MARVIN MINSKY

THE EMOTION MACHINE

COMMONSENSE THINKING, ARTIFICIAL INTELLIGENCE, AND THE FUTURE OF THE HUMAN MIND
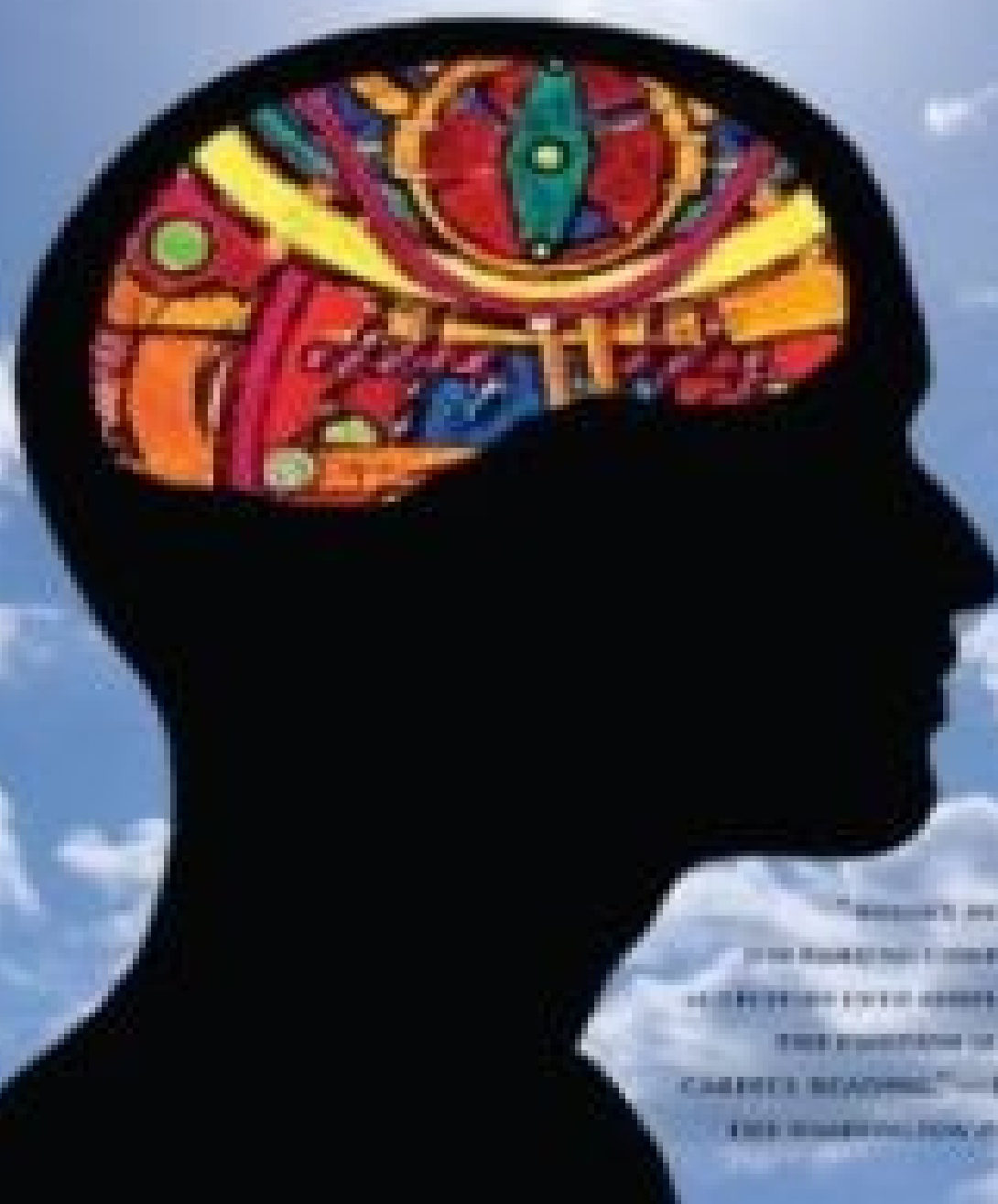
MARVIN MINSKY

AUTHOR OF THE SOCIETY OF MIND

THE EMOTION MACHINE

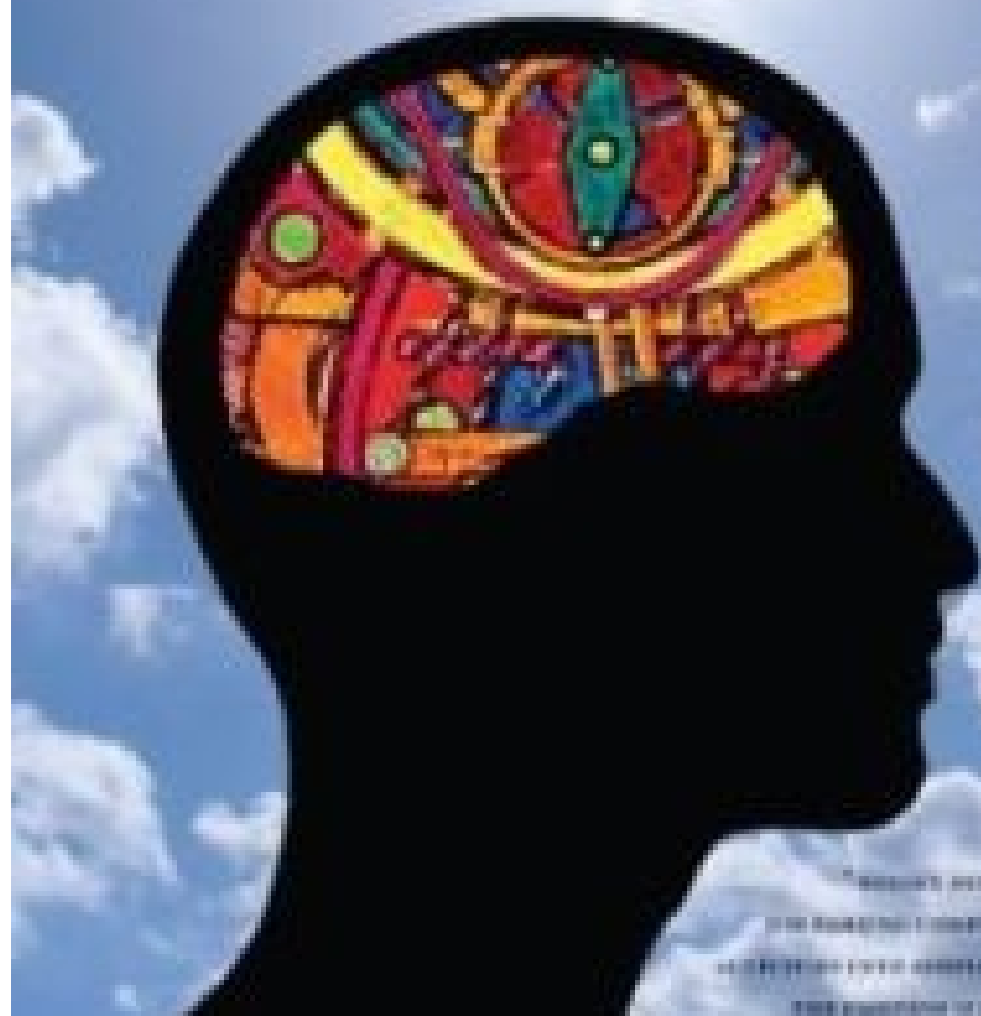## COMMONSENSE THINKING, ARTIFICIAL INTELLIGENCE, AND THE FUTURE OF THE HUMAN MIND

Marvin Minsky

The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind (draft)

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 work, or 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—such as 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, this book will argue that this quest will fail because no simple such 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 much about how each of those hundreds of brain-centers works, we do know that their construction is based on information that is contained in tens of thousands of inherited genes—so that each brain-parts 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.

For example, 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 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."[1]
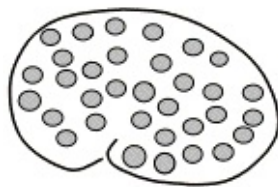
We'll use this image whenever we want to explain some mental activity (such as Anger, Love, or Embarrassment) by trying to show how that state of mind might result from the activities of a certain collection of mental resources.

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'll 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.
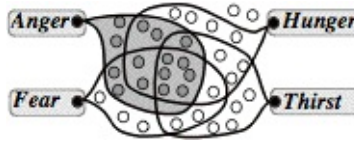
However, as we learn and grow, we also develop ways to activate other, new sets of resources to use—and this leads to types of mental states that we regard as more "intellectual" than "emotional." For example, whenever a problem seems hard to you, then your mind will start to switch among different Ways 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. In other words:

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.

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

Each statement in quotation marks is something said by an actual person; if it also has a publication 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 bibliographic citations, such as

Schank 1975: Roger C. Schank, *Conceptual Information Processing*, Elsevier Science Publishers 1975. ISBN: 0444107738.

Some references are to pages on the World Wide Web.

Lenat 1998: Douglas B. Lenat, *The Dimensions of Context Space*, 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, 7 Feb 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". Also I will try to maintain copies of these on my website at www.emotionmachine.net, and invite readers with questions and comments to send them to me by using that web site.

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, so that mental agents could cooperate in much the same ways that people do. On the contrary, most resources are specialized to [do] 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.

Part I

§1-1. Falling in Love

> *"Oh, life is a glorious cycle of song,*
> *A medley of extemporanea;*
> *And love is a thing that can never go wrong;*
> *And I am Marie of Roumania."*
>
> — *Dorothy Parker*[1]

Many people find it absurd to conceive of a person as being a kind of machine —so we often hear statements like 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 and *ought* not be explained in such ways! Listen to Pablo Neruda:

" ...love has to be so,
involving and general,
particular and terrifying,
honoured and yet in mourning,
flowering like the stars,
and measureless as a kiss."

   —from 'Extravagaria'

What is Love, and how does it work? Is this something we want to understand, or should we see such poems as hints that we don't really care to probe into it? 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-', '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 were caused by attempts to defend him from thinking about what his girlfriend says:

Celia: "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?"[2]

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 aware of them:

"WHEN my love swears that she is made of truth,
I do believe her, though I know she lies,
That she might think me some untutor'd youth,
Unskilful in the world's false forgeries.
Thus vainly thinking that she thinks me young,
Although I know my years be past the best,
I smiling credit her false-speaking tongue,
Outfacing faults in love with love's ill rest.
But wherefore says my love that she is young?
And wherefore say not I that I am old?
O, love's best habit is a soothing tongue,

And age, in love, loves not to have years told.
Therefore I'll lie with love, and love with me,
Since that our faults in love thus smother'd be."

We are equally apt to deceive ourselves, not only in our personal lives but also when dealing with abstract ideas. There, too, we frequently find ways to keep inconsistent or discordant 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 word ought to cover the one you adore —but if your goal is just to extend the pleasure that comes when doubts get suppressed, then you're only in love with Love itself.

<p style="text-align:center">∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞</p>

Citizen: Your description of 'love' in the section above spoke only of transient infatuation—of sexual lust and extravagant passion. It left out most of what we usually mean by that word—such as loyalty and tenderness, or attachment, 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 conception of love is 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.

Attachments of members to groups or their leaders.

We also apply that same word 'love' to our fondness for objects, events, and beliefs.

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.

We thus apply 'love' to our likings for things that we treasure, desire, or fill us with pleasure. We apply it to bonds that are sudden and brief, but also to those that increase through the years. Some occupy just small parts of our minds, while others pervade our entire lives.

But why do we pack such dissimilar things into a single suitcase-like word? It's the same for our other 'emotional' terms; each of them abbreviates a diverse collection of mental states. Thus Anger may change our ways to perceive, so that innocent gestures get turned into threats, and it alters the ways that we react, to lead us to face the dangers we sense. Fear too affects the ways we react, but makes us retreat from dangerous things (as well as from ones 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 changes in how we think.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§1-2. The Sea Of Mental Mysteries

Every now and then we dwell on questions about how we manage ourselves.

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

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'll need to get better ideas about the processes that we call *thinking*. But whenever we start to think about this, 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 as Alexander Pope suggests in his *Essay on Man,* we still know almost nothing about how those processes actually work.

"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 that explained all electro-magnetic events—and Einstein then reduced all those laws into yet smaller formulas. All this came from the success of those physicists' quest: *to find simple explanations for things that, at first, seemed extremely 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 that have been trained to believe such statements as, *"One should never adopt hypotheses that make more assumptions than they need."* 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 about your visual systems partition a scene into various patches of color and texture—and then assign them to different *'things.'* And, *'see what it is'* sidesteps all the questions you could ask about how that sight might be related to other things that you've seen in the past.

It is much the same for the commonsense words that we usually use to talk about what our own minds do, as when one makes a statement like, *"I think I understood what you said."* For perhaps the most extreme example of this is how we use words like *Me* and *You*—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 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 just delegate them to something else? Here are a few kinds of reasons why a mind might entertain such a fiction:

Child Psychologist: Among the first things you learn to recognize are the persons in your environment. In your next stage, you should assume that you are also a person, too. But perhaps it is easier to conclude that there is a person inside of you.

Therapist: Although it's a legend, it makes life more pleasant—by keeping us from seeing how much we're controlled by conflicting, unconscious goals.

Pragmatist: 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, that 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 only offers useless answers like these:

My Self selects what to think about.

My Self decides what I should do next.

I should try to make Myself 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 you are asked about some difficult 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, one can scarcely imagine 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, none of those questions are simple at all. The process of 'seeing' a car or a chair uses 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 many processes that are most vital to us have evolved to work inside parts of the brain that have come to function so 'quietly' that the rest of our minds have no access to them. This could be why we find it so hard to explain many things we find so easy to do.

In Chapter §9, we'll come back to that Self—and argue that this, too, is a very large and complicated structure.

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

> *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?*
> —William James, in Principles of Psychology.

Sometimes you find yourself in a state where everything seems cheerful and bright. Other times (although nothing has changed) everything seems dreary and dark, and your friends describe you as being 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 part of consciousness that involves feeling.
A non-rational 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? How are emotions *involved* with *feelings*? Must every emotion involve a *disturbance*? And what could a *conscious affection* be?

Why do so many such questions arise when we try to define what 'emotion' means? That's because 'emotion' is one of those suitcase-words that covers too wide a range of things. Here are just a few of the hundreds of terms that we use for discussing our mental conditions:

Admiration, Affection, Aggression, 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. So, many researchers have spent their lives at trying to classify our states of mind, by stuffing familiar words like these into such classes as *humors, emotions, tempers, and moods.* But should we call *anguish* a feeling or mood? Is *sorrow* a type of agitation? There's no way to settle the use of such terms because, as William James observed above, different traditions make different distinctions, and may not describe the same states of mind because different people have different ideas. How many readers can claim to know precisely how each of those feelings feels?[3]

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.

Although it is hard to define words like *feeling* and *fearing*, that's rarely a problem in everyday life because our friends usually know what we mean. However, attempts to make such terms more precise have hindered psychologists more than they've helped to make theories about how human minds work. So this chapter will take a different approach, and think of minds as composed of much smaller parts or processes. This will lead to some new and useful ways to imagine what feeling and thinking might be.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞
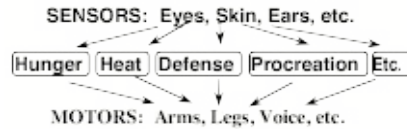
§1-4. Infant Emotions.

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.

*—Charles Darwin, in The Emotions of Animals*

One moment your baby seems perfectly well, but then come some restless motions of limbs. Next you see a few catches of breath—and in the next moment, the air fills with screams. Is baby hungry, sleepy, or wet? Whatever it is, those cries compel you to find some action that will help. It may take you some time to discover the trouble, but once you find the remedy, things quickly return to normality. However, if you're not used to dealing with infants, those sudden switches in mood can upset you; when your friends cry, you can ask them what's wrong—but talking to infants is fruitless because "no one is home" to communicate with.

Of course, I do not mean to suggest that infants don't have 'personalities.' You can usually sense, quite soon after birth, that a particular baby reacts more quickly, or seems more patient or irritable, or even more inquisitive. Some of those traits may change with time, but others persist through the rest of that life. Nevertheless, we still need to ask, *how could an infant change so much between one moment and the next?* The Single-Self model cannot explain how suddenly an infant can switch from contentment or calmness to anger or rage.

To make a more plausible model for this, 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 might need to find sources of water and food. It might need defenses against attacks—and against extremes of temperature. It might even need ways to attract helpful friends. Then once you have assembled that list, you could tell your engineers to meet each such need by building a separate "instinct-machine.

SENSORS: Eyes, Skin, Ears, etc.

Hunger | Heat | Defense | Procreation | Etc.

MOTORS: Arms, Legs, Voice, etc.

Then, how could we build those instinct-machines? 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 for recognizing conditions.

Motors for physical actions.
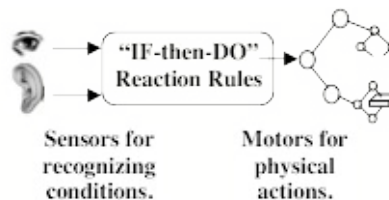
What could be in that central knowledge box? Let's begin with the simplest case: suppose that we already know, in advance, all the situations our robot will face. Then all we need is a catalog of simple, two-part "*If–>Do*" rules—where each *If* describes one of those situations—and its *Do* describes which action to take. Let's call this a *"Rule-Based Reaction-Machine."*

Rule Based Reaction-Machine

Specific Situation		Appropriate Action
IFs		DOs
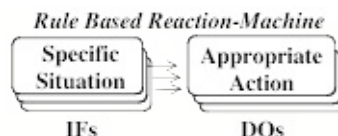
If temperature wrong, *Adjust* it to normal.
If you need some food, *Get* something to eat.
If you're facing a threat, *Select* some defense.
If an active sexual drive, *Search* for a mate.

Many *If–>Do* rules like these are born into each species of animals. For example, every infant is born with ways to maintain its body temperature: when too hot, it can pant, sweat, stretch out, and vasodilate, when too cold, it can retract its limbs or curl up, shiver, vasoconstrict, or otherwise generate more heat. [See §6-1.2.] Later in life we learn to use actions that change the external

world.

> **If** *your room is too hot,* **Open** *a window.*
> **If** *too much sunlight,* **Pull** *down the shade.*
> **If** *you are too cold,* **Turn** *on a heater.*
> **If** *you are too cold,* **Put** *on more clothing.*

This idea of a set of "*If–>Do* rules" portrays a mind as nothing more than a bundle of separate reaction-machines. Yet although this concept may seem too simplistic, in his masterful book, *The Study of Instinct,*[4] Nikolaas Tinbergen showed that such schemes could be remarkably good for describing some things that animals do. He also proposed some important ideas about what might turn those specialists on and off, how they accomplish their various tasks, and what happens when some of those methods fail.

Nevertheless, no structure like this could ever support the intricate feelings and thoughts of adults—or even of infants. The rest of this book will try to describe systems that work more like human minds.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§1-5. Seeing a Mind as a Cloud of Resources.

Today, there are many thinkers who claim that all the things that human minds do result from processes in our brains—and that brains, in turn, are just complex machines. However, other thinkers still insist that there is no way that machines could have the mysterious things we call *feelings*.

Citizen: A machines can just do what it's programmed to do, and then 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 or 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 older times, those were plausible views because we had no good ideas about how biological systems could do what they do. Living things seemed completely different from machines before we developed modern instruments. But then we developed new instruments—and new concepts of physics and chemistry—that showed that even the simplest living cells are composed of hundreds of kinds of machinery. Then, in the 20th century, we discovered a really astonishing fact: *that the 'stuff' that a machine is made of can be arranged so that its properties have virtually no effect upon the way in which that machine behaves!*

Thus, to build the parts of any machine, we can use any substance that's strong and stable enough: *all that matters is what each separate part does, and how those parts are connected up.* For example, we can make different computers that do the same things, either by using the latest electrical chips— or by using wood, string and paper clips—by arranging their parts so that, seen from outside, each of them does the same processes. [See §§Universal Machines.]

This relates to those questions about how machines could have emotions or feelings. In earlier times, it seemed to us that emotions and feelings were basically different from physical things—because we had no good ways to imagine how there could be anything in between. However, today we have many advanced ideas about how machines can support complex processes—and the

rest of this book will show many ways to think of emotions and feelings as processes.

This view transforms our old questions into new and less mysterious ones like, *"What kinds of processes do emotions involve,"* and, *"How could machines embody those processes?"* For then we can make progress by asking about how such a brain could support such processes—and today we know that every brain contains a great many different 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 engage large bodies of knowledge. This suggests a way to envision a mind (or a brain) as made of hundreds or thousands of different resources.
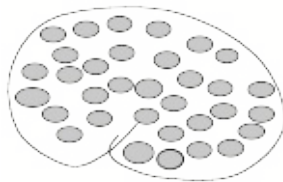
At first this image may seem too vague—yet, even in this simple form, it suggests how minds could change their states. For example, in the case of Charles's infatuation, this suggests that some process has switched off some resources that he normally uses to recognize someone else's defects. The same process also arouses some other resources that tend to replace his more usual goals by ones that he think Celia wants him to hold.

Similarly, the state we call *Anger* appears to select a set of resources that help you react with more speed and strength—while also suppressing some other resources that usually make you act prudently; *Anger* replaces cautiousness with aggressiveness, trades empathy for hostility, and makes you plan less carefully.

More generally, this image suggests that there are some *'Selectors'* built into our brains, which are wired to arouse and suppress certain particular sets of resources.
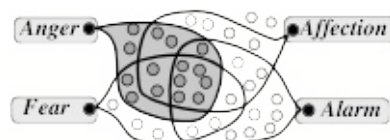
THESIS: Each of our major 'emotional states' results from turning some set of resources on and turning another set of them off. Each such selection will change how we think by changing our brain's activities.

Why would a brain be equipped with such tricks? Each of them could have evolved to promote some special important function; *anger* and *fear* evolved for protection, and *affection* evolved to promote reproduction (which sometimes engages quite risky behaviors).

If several selectors are active at once, then some resources may be both aroused and suppressed. This could lead to the kinds of mental states in which we sometimes say, "our feelings are mixed." Thus when some of your 'Critics' detect some sort of threat, this might activate *Selectors* that make you want both to attack and retreat, by arousing parts of both *Anger* and *Fear*.
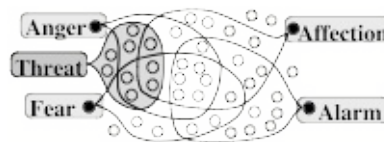
Student: I could better grasp what you're talking about, if you could be a bit more precise about what you mean by the word 'resource.' Do you imagine that each resource has a separate, definite place in the brain?

I'm using 'resource' in a hazy way, to refer to all sorts of structures and processes that range from perception and action to ways to think about bodies of knowledge. Some resources use functions that are performed in certain particular parts of the brain, while others use parts that are more widely spread over much larger portions of the brain. (We'll discuss this more in §§Resources).

As we said, this resource-cloud idea may seem vague—but the rest of this book will develop more detailed ideas about what our mental resources could do —and how their activities lead to the ways that people come to think and behave. Then, as we proceed to develop those schemes, we'll replace this vague Resource-Cloud idea scheme with more elaborate theories about how our resources are organized.

Romanticist: You speak of a person's emotional states as nothing more than

ways to think, but surely that's too cold and abstract—too intellectual, dull, and mechanical. It says nothing about where feelings come in, with all their colors and intensities—or about our ambitions and goals. It doesn't explain the pleasures and pains that come from when we succeed or fail, or how our bodies and minds interact, as when we're aroused by works of art.

Rebecca West: "It overflows the confines of the mind and becomes an important physical event. The blood leaves the hands, the feet, the limbs, and flows back to the heart, which for the time seems to have become an immensely high temple whose pillars are several sorts of illumination, returning to the numb flesh diluted with some substance swifter and lighter and more electric than itself."[5]

In our usual, everyday views of ourselves, some of our feelings seem to be in our bodies—as when we're affected by muscular tensions. However, our brains can't directly detect those tensions themselves; instead, we sense signals that come up to our brains through nerves that run from those muscles and tendons. This means that we can see bodies, too, as composed of resources that brains can use.

So, instead of discussing emotions as though they were a distinctive kind of phenomenon, the rest of this book will show why it's better to focus on what kinds of mental resources we have, what sorts of things those resources might do, how each affects the ones it's connected to. And especially, we'll develop ideas about what turns those resources off and on.

Student: Why should one ever turn off a resource? Why not keep them all working all the time?

Indeed, certain resources are *never* switched off—like those involved with vital functions like respiration, balance, and posture—nor are those that constantly keep watch for certain particular types of danger. However, if *all* our resources were active at once, then they would often get into conflicts. You can't make your body both walk and run, or move in two different directions at once. How should we resolve such internal conflicts? In a human society, the simplest way is for individuals to compete. But when competition leads to excessive waste, then we find ways to organize ourselves into multiple levels of management, in which each manager has authority to decide among the options proposed by lower ones.
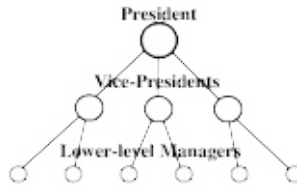
However, a human mind cannot be so hierarchical. This is because, in general, no single, lower-level resource will be able to solve any difficult problem by itself. So when a lower-level 'Critic' resource encounters a problem it needs to solve, then it may transiently need to take over control of one or more high-level strategies—for example, to divide the problem into simpler parts, or to remember how a similar problem was solved in the past, or to make a series of different attempts and then to compare and evaluate these. So a *Critic* may try to arouse several *Selectors*, each of which could lead to different way to think.

Now, each of such high-level strategies will need to use hundreds of lower-level processes, so if we tried to use several such "ways to think" at once, they would tend to interfere with each other—so we'll still need some high level management. This could be one reason why our 'thinking' often seems to us more like a serial, step-by-step process than like one in which many things happen at once. However, every such high-level step will still need to engage many low-level processes that may need to work simultaneously. So the sense that our thoughts flow in serial streams must be in large part an illusion that comes because the higher-level parts of our minds know so little about those sub-processes. (We'll discuss this more in §4 and §7.)

Critic: In any case, it seems to me that your Resource-Switching view is too radical. Perhaps it could be used to explain the behavior of an insect or fish—but Charles doesn't switch, in the way you describe, to a totally different mental state. He changes some aspects of how he behaves, but surely he still remembers his name—and remains the same in most other ways.

It's true that we've only presented a caricature. To develop our Cloud-of-Resources idea, we began with a simplified version in which each resource is either switched on or off. To some degree, this might apply to some of the actions of insects and fish—and to some of what human infants do, for they are prone to strong and quick changes in state. However, in the course of growing up, we develop techniques for "self-control" and our resources become much less clearly 'switched.' Instead, we arouse and suppress them to different extents, so that we still can listen and speak, and to access our bodies of knowledge and skills—though we'll use these with different priorities. And through time we develop more intricate ways to control both old instincts and new processes, and to make new kinds of arrangements of them, in which multiple ones are active at once—and that's when we speak of our feelings as *mixed*.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§1-6. Adult Emotions

Behold the child, by nature's kindly law,
Pleas'd with a rattle, tickl'd with a straw:
Some livelier plaything gives his youth delight,
A little louder, but as empty quite:
Scarfs, garters, gold, amuse his riper stage,
And beads and pray'r books are the toys of age:"
　　　—Alexander Pope in Essay on Man.

Often, when a young infant gets angry, that change seems as quick as the flip of a switch.

A certain infant could not bear frustration, and would react to each setback by throwing a tantrum. He'd hold his breath and his back would contract so that he'd fall rearward on his head.

A simple theory of how this might work would be that some separate 'instincts' compete until just one of them takes over control. However, that model cannot explain how, later, that child finds new ways to deal with frustration:

A few weeks later, that behavior had changed; no longer completely

controlled by his rage, he could also add ways to protect himself, so that when he felt this coming on, he'd run to collapse on some soft, padded place.

This suggests that usually, in the infant brain, only one *Selector* can work at a time; this makes the system change states decisively, so that not many conflicts will arise. However, those infantile systems cannot solve the kinds of hard problems our children must face as they move into their later lives. This led our human brains to evolve higher-level systems in which some instincts that formerly were distinct now became increasingly mixed. But as those systems gained more abilities, they also gained new ways to make mistakes, so they also had to evolve new ways to control themselves—and this led to a great cascade of new kinds of mental developments.[6]

We tend to regard a problem as 'hard' when we've tried several methods without making progress. But it isn't enough just to know that you're stuck: you'll do better if you can recognize when you're facing some particular type of barrier, impasse, or obstacle. For if you can diagnose what *"Type of Problem"* you face, this can help you to select a more appropriate *"Way to Think."* So, later chapters of this book will suggest that to do such things, our brains replace some of their ancient *"Rule-Based Reaction-Machines"* by what we'll call *"Critic-Selector Machines."*

"Critic-Selector" based Machine

Recognize a Problem-Type	Activate a Way to Think
Critics	Selectors
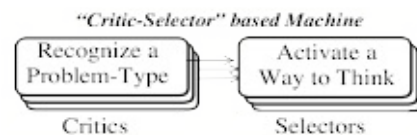
The simplest version of such a scheme would be almost the same as an *"If-Then"* machine of the kind described in §1-4. There, each *"If"* detects a certain real-world problem, which causes the system *then* to react with a certain pre-specified, real-world action. So the behaviors of simple *If-Then* machines are highly constrained and inflexible.

However, in a *Critic-Selector* type of machine, those *Ifs* and *Thens* are more general, because the resources called *Critics* can recognize, not just events in the external world, but problems or obstacles *inside* the mind. Then, those *"Selectors"* also are not confined to acting on things in the outer world, but can react to *mental* obstacles—by turning other resources on or off. This means a Critic-Selector machine need not just react to external events, but also can direct

itself to switch to a different way to think. For example, it might first consider several reactions before it decides which one to use.

Of course, we'll need more specific ideas about how each of those new Ways to think might work, and about how we come to develop them. We know that throughout our childhood years, our brains pass through multiple stages of growth, and Chapter §5 will conjecture that this results in at least these six levels of mental procedures.

Thus, an adult who encounters what might be a threat need not just react instinctively, but can proceed to *deliberate* on whether to retreat or attack—that is, to use higher-level strategies to choose among possible ways to react. This way, one can make thoughtful choice between the conditions of *Anger* and *Fear* —and if it seems more appropriate to intimidate an adversary, one can make oneself angry deliberately (although one may not be aware of doing this).

We know that these mental abilities grow over several years of one's childhood. Then why is it that we can't recollect much of that stretch of development? One reason for this could be that, during those years, we also develop new ways to build memories—and when we switch to using these, that makes it hard to retrieve and interpret the records we made in previous times. Perhaps those old memories still exist, but in forms that we no longer can comprehend—so we cannot remember how we progressed from infantile reaction-sets to using our new, adult ways to think. We've rebuilt our minds too many times to remember how our infancies felt!

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§1-7. Emotion Cascades

Some habits are much more difficult to cure or change than others are. Hence a struggle may often be observed in animals between different instincts, or between an instinct and some habitual disposition; as when a dog rushes after a hare, is rebuked, pauses, hesitates, pursues again, or returns ashamed to his master; or as between the love of a female dog for her young puppies and for her master, —for she may be seen to slink away to them, as if half ashamed of not accompanying her master.
   *—Charles Darwin, in The Descent of Man*

This chapter has raised some questions about how people could change their states so much. *When someone you know has fallen in love, it's almost as though a switch had been thrown, and a different program has started to run.*

The Resource-Cloud image suggests that such a change could result when a certain "Selector" excites (or suppresses) a certain large set of resources. Thus Charles's attraction to Celia becomes stronger when all his fault-finding Critics turn off.

Psychologist: Indeed, infatuations sometimes strike suddenly. But other emotions may flow and ebb slowly—and usually, in our later years, our mood-shifts tend to become less abrupt. Thus an adult may be slow to take offense, but may then go on to brood for months on even a small or imagined affront.

Our twenty-year-old tabby-cat shows few signs of human maturity. At one moment she'll be affectionate, and seek out our companionship. But after a time, in the blink of an eye, she'll rise to her feet and walk away, without any sign of saying goodbye—whereas our twelve-year-old canine pet will rarely depart without looking back—as though he's expressing a certain regret. The cat's moods seem to show one at a time, but the dog's dispositions seem more mixed, and less as though controlled by a switch.

In either case, any large change in one's set of active resources will cause a large change in one's mental state. One way this could happen would be for a certain resource to directly arouse many others:
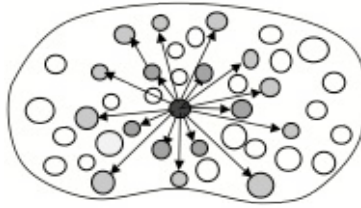
In this way, the *Selectors* we mentioned in §1-5 could directly have substantial effects. Furthermore, if the set of newly aroused resources includes one or more other *Selector* resources, then this will cause a yet larger change, by activating yet more resources. Then these in turn may begin to arouse yet other resources that they need—and if each such change leads to yet several more, this spreading could escalate what we'll call a large-scale "cascade."
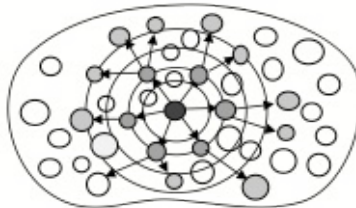
The further these activities spread, the more they will alter your Way to Think—and if your behavior then changes enough, then your friends might get the impression that you have turned into a different person.

Critic: That would be an exaggeration, because Charles will still be the very same person. He will still speak the same language and use the same knowledge; he'll just have some different attitudes.

Of course, those cascades won't change everything. When Charles adopts a new *Way to Think*, in many respects he'll still be the same—because not all his resources will have been replaced. He still will be able to see and hear—but now he'll perceive things in different ways. And because he now represents them in different ways, he'll get different ideas about what those things or events might "mean."

Charles will also still know how to talk—but may now use different styles of speech, and choose different subjects to talk about because, although he still has access to the same knowledge, skills, and memories, now different ones will be

retrieved. He still may maintain the same plans and goals—but now they'll have different priorities. He may still get dressed and go to work—but in some of those states he won't dress so well. And so far as Charles, himself is concerned, he still has the same identity.

To what extent, then, will Charles be aware of such changes in his mental condition? He sometimes won't notice those changes at all—but at other times, he may find himself making remarks to himself like, *"I am getting angry now."* To do this, his brain must have ways to "reflect" on some of its recent activities (for example, by recognizing the spread of some large-scale cascades). Chapter §4 will discuss how such processes could lead to some aspects of what we call "consciousness."

<center>∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞</center>

§1-8. Questions.

What are dispositions and moods?
We all use many different words to vaguely describe how we feel and behave. We know that *angry* people more quickly react (but, usually, less cautiously) and that *happy* people less often start fights—but terms like these do not suggest ideas about how those states affect how we think. We recognize this when we deal with machines: Imagine that your car won't start—but when you ask your mechanic for help, you only receive a reply like this:

> *"It appears that your car doesn't want to run. Perhaps it's become annoyed with you because you haven't been treating it well."*

But psychological terms like these don't help you to get good ideas to explain the behavior of your car. Perhaps you towed too heavy a load and broke some of the teeth of one of the gears. Or perhaps you left the lights on all night, and completely discharged the battery. Then those 'mentalistic' descriptions won't help you; to diagnose and repair what's wrong, you need to know about that car's parts.

That's where the view of a mind as a *Cloud of Resources* is better than the *Single-Self* view; it encourages us to look at the parts instead of the whole. Is there something wrong with the starter switch? Has the fuel tank been completely drained? Those commonsense psychology-words are useful in everyday social life, but to better understand our minds we need more ideas about their insides.

To what extents are emotions innate? It would seem that all normal person

share some common emotions, such as anger, ^{fear}, sadness, joy, disgust, and surprise—and some would also include curiosity. However, psychologists do not broadly agree about which of these are innate and which are learned; for example, some of them regard anger as based on fear. This book will not get involved in that debate, because it is more concerned with what emotions *are*—in the sense of being 'ways to think'—than with finding ways to classify them.[7]

How do Chemicals affect our Minds?

Physiologist: Your ideas about switching resources sound good, but can all mental states be explained in that way? Aren't we also affected by chemicals like hormones, endorphins, and neurotransmitters?

There's no doubt that such chemicals do affect the internal states of our brains—but the view that those effects are *direct* is a popular but bad mistake—somewhat like the error that someone would make by supposing that rain makes umbrellas unfold. Here's how one author depicts what this misses:

Susanna Kaysen: "Too much acetylcholine, not enough serotonin, and you've got a depression. So, what's left of mind? It's a long way from not having enough serotonin to thinking the world is "stale, flat and unprofitable"; even further to writing a play about a man driven by that thought."[8]

For just as the meaning of each separate word depends on the sentence that it is in, the effect of each chemical on the brain depends on all the particular ways in which each of your brain-cells react to it—each type of cell may differ in that. So the effect of each chemical will depend which brain-cells react to it—and then on how other cells in that happen to be connected to these, *etc*. So the large-scale effect of each chemical depends, not only on where and when it's released, but also on the other details of the interconnections inside your brain. We'll discuss more details in §§Chemicals.

How could machines understand what things mean?

In the popular view, machines do things without understanding what their activities mean. But what does *'understanding'* mean? Even our best philosophers have failed to explain what we mean by words like "understand."[9]

However, we should not complain about that, because this is precisely the way it should be! For, most of our common psychology-words have this peculiar property: *the more clearly you try to define them, the less you capture their commonsense meanings.* And this applies especially to words like *understand* and *mean*!

If you 'understand' something in only one way then you scarcely understand

it at all. For then, if anything should go wrong, you'll have no other place to go. But if you represent something in multiple ways, then when one of them fails you can switch to another—until you find one that works for you.

It's the same when you face a new kind of problem:

If you only know a single technique, then you'll get stuck when that method fails. But if you have multiple ways to proceed, then whenever you get into trouble, you'll be able to switch to a different technique.

We switch how we think so fluently that we scarcely aware that we're doing this—except when this leads to cascades so great that we notice a change in emotional state. One of the central goals of this book is to describe the variety of our mental resources, and how these might be organized—and the final chapters of this book will show that much of our human resourcefulness depends upon on having multiple ways to escape from getting stuck.

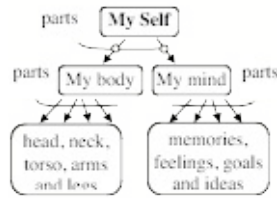**Why do we think that we have Selves?**

Citizen: If my mental resources keep changing so much, then what gives me the sense that I'm still the same Self—no matter how happy or angry I get?

We do not have any good evidence that young infants start out with any such sense—and we can't trust our infantile memories. Then why do all of us come to believe that somewhere, deep in the heart of each mind, there exists some permanent entity that experiences all our feelings and thoughts? Here's what I think might lead to this:

In early life, our low-level processes solve many small problems without any sense of what's doing it. And when those processes run into trouble, then those processes simply stop, and the mind simply starts doing something else.

However, as we develop more levels of thought, those higher levels try to find out "what went wrong" and to improve our skills for this, we start to construct new ways to portray aspects of our recent thoughts. Eventually these develop into simplified 'models' of ourselves.

Perhaps the simplest and most common such model is composed of parts like these:

parts My Self

parts My body   My mind parts

head, neck,   memories,
torso, arms   feelings, goals
and legs   and ideas

However, every normal person also builds many other kinds of self-models that try to describe how they think about such subjects as their social relationships, physical skills, political views, and economic, spiritual, and sexual attitudes. Chapter §9 will go on to suggest that what each person calls his or her 'Self' is a great network of such 'mental models'—each of which attempts to describe only certain aspects of a person's own mind.

Why have multiple models of Selves?

Physicist: Why not simply combine all those models into a single, unified one that merges the virtues of all those separate ones?

That probably would not be practical, because such a structure would be too large for us to 'keep in mind' all its details at once. This suggests that the limitations of our brains must constrain us, at each particular moment, to superimpose a few model-cartoons—each by itself too incomplete to answer most questions about yourself.

Besides, for each particular such kind of problem, some of those models will help more than others—by highlighting the most relevant features. This means both that you need to use multiple views, and that you need ways to rapidly switch among them. Let's listen to Richard Feynman again:

"...Psychologically we must keep all the theories in our heads, and every theoretical physicist who is any good knows six or seven different theoretical representations for exactly the same physics. He knows that they are all equivalent, and that nobody is ever going to be able to decide which one is right at that level, but he keeps them in his head, hoping that they will give him different ideas for guessing."[10]

The key word here is 'guess' because every such theory has virtues and faults; no single model or representation is best for every different purpose or goal—and each is likely to get you stuck in certain kinds of predicaments.

How do we develop new goals and ideas?

The next few chapters will take the commonsense view that everyone already knows what goals are, and focus instead on questions about how we

come to acquire them. However, that discussion will be incomplete until we present (in Chapter §6) more detailed ideas about how goals work.

In the usual view of how human minds grow, each child begins with instinctive reactions, but then goes through stages of mental growth that overlay these with additional layers and levels of goals. Those older instincts may still remain, but these new resources gain increasing control—until we can think about our own motives and goals, and perhaps try to change or reformulate them.

But what possible basis could we use for learning to appraise ourselves? How could we choose which new goals to adopt—and how could we possibly justify them? No infant could ever be wise enough to make good such choices by itself. So the following chapter will argue that our brains must have evolved, instead, ways to copy the ideals and attitudes of our parents, friends, and acquaintances!

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

Part II. Attachments and goals

§2-1. Playing with Mud

> *"It's not just learning things that's important. It's learning what to do with what you learn and learning why you learn things at all that matters."*
> —Norton Juster, in The Phantom Tollbooth

A child named Carol is playing with mud. Equipped with a fork, a spoon, and a cup, her goal is to bake a make-believe cake, the way she's seen her mother do. Let's assume that she is playing alone, and imagine three things that might happen to her.

• ***Playing alone.*** *She wants to fill her cup with mud, and first tries to do this with her fork, but this fails because the mud slips through. She feels frustrated and disappointed. But when she succeeds by using the spoon, Carol feels satisfied and pleased.*

What might Carol learn from this? She learns from her 'trial and error' experience that forks are not good for carrying mud. But she learns from her success with a spoon, that these are good tools for moving a fluid. From failures we learn which methods don't work—while successes teach us which methods succeed. [But see §9-2.]

Note that Carol did this while working alone—and acquired new knowledge, all by herself. *In the course of learning by trial and error, a person requires no teacher to help her.*

• ***A Stranger Scolds.*** *Unexpectedly, a stranger reproaches her: "That's a naughty thing to do." Carol feels anxious, alarmed, and afraid. Overcome by fear and the urge to escape, she puts her present goal on hold—and runs to find her mother.*

What might Carol learn from this? She may not learn much about working with mud, but may classify this as a dangerous *place*. Also, too many scary encounters like this might make her become less adventurous.

• ***Her Mother's Reproach.*** *Carol returns to her mother's protection—but*

instead of assurance, her parent rebukes her. "What a disgraceful mess you've made! See what you've done to your clothes and face. I scarcely can bear to look at you!" Carol, ashamed, begins to cry.

What might Carol learn from this? She'll become less inclined toward playing with mud. If her parent had chosen to praise her instead, she would have felt pride instead of shame—and in future times would be more inclined to further pursue that same kind of play. *In the face of a parent's blame or reproach, she learns that her goal was not good to pursue.*

Think of how many emotional states a child engages in the thousand minutes of each of its days! In this very brief story we've touched upon *satisfaction*, *affection*, and *pride*—passions we think of as positive. We also encountered *shame* and *disgrace*—and *fear*, *anxiety*, and *alarm*—all feelings we think of as negative. What could be the functions of these various kinds of mental conditions? Why do they seem to come in opposing pairs? How could the physical systems in our brains produce these sorts of feelings and thoughts? This book will try to answer many such questions, but this chapter will mainly focus on some ideas about the functions of our children's early attachments to other persons.

Clearly, attachments help young animals to survive, through nourishment, comfort, and protection from harm. However, this chapter will argue that those special feelings of Pride and Shame play unique and peculiar roles in how we develop new kinds of goals. And because adult minds are so much more complex, we'll start by discussing what children do.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§2-2. Attachments and Goals

> *"Never let your sense of morals prevent you from doing the right thing."*
> —Isaac Asimov

Some of our strongest emotions come when we are in the presence of the persons to whom we've become attached. When we're praised or rebuked by the people we love, we don't just feel pleased or dissatisfied; instead, we tend to feel proud or ashamed. This section will suggest some possible reasons why we might have these particular feelings, as well as some ways in which they may be

involved with how our values and goals develop.

Most other mammals, soon after birth, can move and follow their mothers about—but human infants are peculiarly helpless. Why did our infants come to evolve such a slow course of development? In part, this must have been because their larger brains needed more time to mature. But also, as those more versatile brains led to more complex societies, our children had to evolve new ways to 'download' knowledge efficiently; no longer did they have enough time to learn from 'trial and error' experience.

One way to learn more quickly was to develop better ways to observe and describe what other, older persons do. Another, more novel development was to 'learn by being told'—by using the kinds of expressions that eventually led to our languages. Both of these advances were further enhanced by two complementary developments; the children evolved increased concern with how their parents reacted to them, and the parents evolved increased concern for the welfare of their children.

Both of these needed powerful ways for each to get the others' attention. For example, our infants are born equipped with shrieks that arouse their parents from deepest sleep. Those screams are irresistible because, as in the case of other loud sounds, they exploit connections related to pain, which activate powerful goals to find ways to eliminate those stimuli. Other such systems make children feel disturbed whenever their parents move too far away—and human parents feel similar pains when they lose track of where their infants are. We can see how some of these systems might work by reviewing those scenes in which Carol learned.

In the scene in which Carol was playing alone, in using a fork failed to fill her cup. Her disappointment then helped her learn not to use that method again. But when she felt pleased by success with a spoon, her satisfaction helped her learn that this was a better method to use—so that next time she wants to fill a cup, she'll know more about how to do it.

Here Carol learns via 'trial and error', without any need for a teacher to help her. What could have impelled her to persist, in spite of those first disappointments? In §9-4 we'll come back to discuss why we sometimes put up with unpleasantness.

In the scene in which a stranger appeared, Carol felt a sense of fear. This led her to look for a way to escape and to seek her parent's protection.

This probably had no effect on her goal of learning how to put mud in a cup —and more likely taught her to dread that location. Next time she'll play in

some safer place.

In the scene where Carol's mother reproached her, the child felt Shame—a special kind of emotion. This changed the nature of what she learned: she altered her goals, instead of her methods!

Why did Carol learn in so different a way—when censured by her mother? That judgment makes the child feel *"I should not have had that disgraceful goal."* But when her mother praises her, she feels that her goal was respectable. It is one thing to learn *how to get* what you want—and another, to learn what you *ought* to want. In practical learning by trial and error, you improve your skills for achieving goals you already hold—for example, by linking new subgoals to them. But when your 'self-conscious' affections are roused, you're likely to alter those goals themselves, or make changes in what they're connected to.

Trial and error can teach us new ways to achieve the goals we already maintain.

Attachment-related blame and praise teach us which goals to discard or retain.

This suggests that Pride and Shame play special roles in what we learn; they help us learn 'ends' instead of 'means'. Listen to Michael Lewis describe some of the striking effects of shame:

Michael Lewis: *"Shame results when an individual judges his or her actions as a failure in regard to his or her standards, rules and goals and then makes a global attribution. The person experiencing shame wishes to hide, disappear or die. It is a highly negative and painful state that also disrupts ongoing behavior and causes confusion in thought and an inability to speak. The body of the shamed person seems to shrink, as if to disappear from the eye of the self or others. Because of the intensity of this emotional state, and the global attack on the self-system, all that individuals can do when presented with such a state is to attempt to rid themselves of it."*[11]

But when do we experience these particular kinds of feelings? They are especially prone to come when we're in the presence of those we respect—or whom we wish to be respected by. This suggests that shame and pride may be involved with how we acquire our high-level goals, and that these values are greatly influenced by those to whom we become 'attached'—at least in those our earliest 'formative' years. So the next few sections will ask some questions like these:

What are goals and how do they work?
What are the spans of those 'formative' years?
To whom do our children become attached?
When and how do we outgrow attachments?
*How do they help us establish our values?*

We're almost always pursuing some goals. Whenever you're hungry, we try to find food. When we sense danger, we strive to escape. When we feel wronged, we may wish for revenge. Sometimes you aim toward completing some work—or perhaps you seek ways to avoid it. We use words like *try, strive, wish, aim, seek,* and *want* so often that our minds seem controlled by collections of goals.

Here is a very simple idea about what words like *want* and *wish* might mean:

You 'want' to achieve a situation G when some active mental process works to reduce the differences between G and your present situation.

Later, in §6-3, we'll see that this idea is more powerful than it may seem. For example, when there are several such differences to remove, then achieving goal G may take several steps. For example, suppose that you're hungry and *want* to eat, but you only have a can of soup. Then you must *seek* some tool to open that can and then *try* to find a bowl and a spoon, and then you'll *want* to feed yourself. Each of those *'needs'* comes from some difference between your situation and the one you want—so each such difference becomes a 'subgoal' of your original goal.
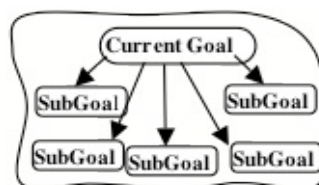
Of course, you first may *need* to make a plan for how to accomplish all those tasks —and making such plans can sometimes engage substantial parts of the rest of your mind.

Citizen: Why do you focus so much on goals, as though all we do is purposeful? Sometimes we simply react to what happens, or act out old, habitual scripts—and sometimes we daydream and fantasize, or aimlessly imagine things.
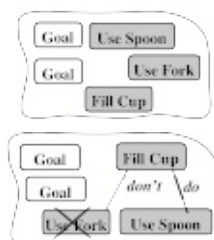
It would be very hard to prove that anything that a person does is wholly devoid of purposes—because, as Sigmund Freud observed, some of our mental processes may work to conceal from us some of our principal motives and goals. But in any case we need more ideas about how we form those purposes.

The most usual theory of how people learn is by what we call 'trial and error.' That's how Carol learned when playing alone, when she worked by herself to fill her cup. She was annoyed when she failed with a fork, but was pleased by success when she used a spoon—so the next time she wants to fill a cup, she'll be more likely to know what to do. That seems like simple common sense—that we learn from failure and from success—but we need a theory of how that might work.

Student: I suppose that her brain formed connections from her goal to the actions that helped her to achieve it.

OK, but that is rather vague. Could you say more about how that actually works?

Student: Perhaps Carol starts with some goals just floating around—but when she succeeds by using her spoon, then she somehow connects her "Fill Cup" goal to her "Use Spoon" goal. Also, when she fails with the fork, she makes a "don't" connection to "Use Fork," to keep from doing that again. Then, the next time she wants to fill a cup, she'll first try the subgoal of using a spoon.

That would be a good way to start, and I like your mentioning those "don't" connections. These are important because we must not only learn to do things that work, but also must learn ways to avoid the most common mistakes.

However, while this kind of theory can help to explain how we interconnect goals that we already possess, it does not answer such questions as, *"How do we get new goals that are not subgoals of existing ones?"* or, more generally, *"How do we learn new ideals and values?"*

I don't recall much discussion of in academic psychology books. The following sections will argue that we cannot acquire our high-level values in the same way that we learn other things, that is, by 'learning from experience.' Instead, we'll argue that children learn values in special ways that depend on the persons to whom they are 'attached.'

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§2-3. Imprimers

> *"Now since shame is a mental picture of disgrace, in which we shrink from the disgrace itself and not from its consequences, and we only care what opinion is held of us because of the people who form that opinion, it follows that the people before whom we feel shame are those whose opinion of us matters to us. Such persons are: those who admire us, those whom we admire, those by whom we wish to be admired, those with whom we are competing, and whose opinion of us we respect."*
>
> *Aristotle, in Rhetoric 2, 6[12]*

Our language has a great many words for describing our emotional states. When we described Carol's playing with mud, we had to use over a dozen of them—*affection, alarm, anxiety, assurance, disappointment, disgrace, disturbance, frustration, fear, inclination, pleasure, pride, satisfaction, shame,* and *sorrow.*

Why do we have such states at all—and why do we have so many of them?

Why does Carol feel grateful and proud when she receives praise from her mother? And how does this, somehow, 'elevate' goals to make them seem more desirable?

Student: You've already started to argue that she must have some kind of "attachment bond" that makes her react in that special way—just as Aristotle said, from concern with her mother's regard for her. But this doesn't explain why praise alone cannot elevate goals, but also depends on the presence of— umm, I can't think of the proper word for this—"a person to whom one has become attached?"

Psychologists often use 'caregiver' for "a person to whom a child is attached." They cannot say 'parent', or 'mother' or 'father' because someone else might play that role—like a grandparent, nurse, or family friend. But 'caregiver' is not the proper word because (as we'll see in §2-7) such attachments can form without physical care. In any case, it seems quite strange that our language has no special word for this most influential relationship! So here I'll introduce two new terms; both are based on an old word, 'imprinting', which long has been used by psychologists to refer to the processes in which young animals learn to keep close to their parents.

Imprimer: *An Imprimer is one of those persons to whom a child has become attached.*

Impriming: *A special way to learn new values that works only when an Imprimer is present.*

Of course, staying close to parents helps to keep offspring safe but, in humans it seems to have other effects; when we're close to the persons to whom we're attached—the ones that we shall call our 'Imprimers'—we find ourselves thinking in special ways. Carol's concern with her cupful of mud may have started out as a casual urge to play with materials near at hand—as just an engaging activity. But when she gets praise from one of her Imprimers, she feels a special thrill of pride that elevates her present goal to a higher kind of priority —and in future times she'll find that, to her, this goal has become more "respectable."

We're always setting new goals for ourselves, but we often end up abandoning them. Why is it sometimes so hard for us to keep working toward what we've decided to do? In §9-2 we'll come back to discuss self-discipline and self-control, but here we'll only mention that attachments also can help us persist—either from hope that we'll please our imprimers or from fear of disappointing them.

Why does an Imprimer's praise have an effect so different from that of praise that comes from a stranger? I do not know of any brain-research that has revealed the machinery involved with this—but it is easy to see how it could have evolved: if strangers could change your high-level goals, they could get you to do whatever they want—*just by changing what you, yourself, want to do!* Children with no defense against this would be less likely to survive, so evolution would tend to select children who could resist that effect.

Student: I like the idea that Attachment induces our children to adopt our values (though perhaps you've induced me to agree by exploiting your role as Imprimer). But is there any evidence that this mechanism really exists?

So far as I know, no parts of our brains have yet been shown to be involved with this, but §2-7 discusses some evidence that damage to a child's attachments can impair that child's development. Future advances in ways to scan brains should tell us more about how such things work.

Student: Even if we knew more about how Attachment affects us, we'd still need explanations of the strengths of those feelings of Pride and Shame.

The final chapters of this book will propose some ideas about what feelings are and how they work.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§2-4. Attachment-Learning Elevates Goals

"Each of us has beliefs about what constitutes acceptable actions, thoughts and feelings. We acquire our standards, rules and goals through acculturation… and each of us has acquired a set appropriate to our particular circumstances. To become a member of any group, we are required to learn them. Living up to one's own internalized set of standards—or failing to live up to them—forms the basis of some very complex emotions."

—Michael Lewis, in [Shame, The Exposed Self, 1991, Free Press, New York.]

When Carol's loved ones censure her, she feels that her goals are unworthy

of her or that she is unworthy of her goals. And when she is somewhat older, then, even when her Imprimers are far from the scene, she still may wonder about how they might feel: *Would they approve of what I have done? Would they approve of what I am thinking now?* What kinds of machinery might we engage that makes us experience such concerns? Let's listen to Michael Lewis again:

"The so-called self-conscious emotions, such as guilt, pride, shame and hubris, require a fairly sophisticated level of intellectual development. To feel them, individuals must have a sense of self as well as a set of standards. They must also have notions of what constitutes success or failure, and the capacity to evaluate their own behavior."

Why would the growth of these personal values depend upon a child's attachments? It is easy to see how this might have evolved: a child who lost its parents' esteem would not be so likely to survive. Also, those parents themselves will want to earn the respect of their friends and peers—so they will want their children to 'behave' in socially acceptable ways, and we've seen several ways for children to learn such things:

*Negative Experience: When a method fails one learns not to use that subgoal.*

Positive Experience: When a method succeeds, one learns to use that subgoal

*Aversion: When a stranger scolds, one learns to avoid such situations.*
Attachment Censure: When an imprimer scolds, the child devalues her goal.
Attachment Praise: When an imprimer praises, the child elevates that goal.

We've already seen how subgoals can be connected to serve our goals—the way that *Use Spoon* could become attached to *Fill Cup.* But when your imprimer praises you, some machinery elevates your current goal, to make it more 'respectable' by raising its place in your cloud of goals.
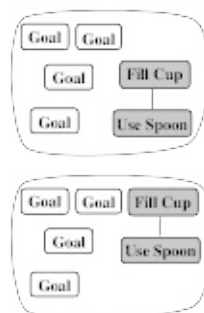
However, this image tells us nothing about how those processes actually work, so we need to construct some theories about how attachment works to 'elevate' goals. First, this must depend on circuits that recognize when the praise comes from an imprimer:
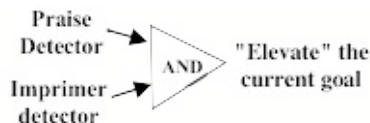
Student: Why did you insist that those "AND" devices should require both praise and an Imprimer?

That's because, as we noted in §2-3, we would all be in danger if praise, alone, could cause our brains to elevate goals—because then any stranger could program us by suggesting new goals and then praising us.

Student: But to some extent that's already true; I am not immune to compliments—even from persons I don't respect.

One feature of human diversity is that we can learn the same things in different ways—and any psychological event is likely to have several causes. If attachment-based learning exists, it is only one part of the story.

Student: But something is missing from this scheme because, even after its level is raised, that 'fill cup" goal is still floating around with no connections that could get it aroused.

Indeed this idea is incomplete. There is no use to learning something new unless one also has ways to retrieve it when it is relevant. This raises many questions like these:

To what should each new goal be attached?
When and how should it be aroused?
What kind of priority should it have?
How long to pursue it, before giving up?

There are no simple answers to these, because all those issues must involve much of the rest of our mental machinery. Nevertheless, it is hard to see how to

think about such things without a set of ideas about 'levels' of mental activities. Our brains have many systems that learn—and as these develop over the years, they may tend to form roughly hierarchical structures, because each fragment of newly acquired knowledge is built upon things that we've learned before.

For example, in the course of everyday thinking, you need to constantly control the "level of detail" of descriptions. When a plan seems to be working successfully, you'll want to "descend" to work out details—but when you seem to be getting stuck, you'll want to 'look up' to a higher-level overview, instead of investing time on subgoals that may not be relevant. [See §§Level-Bands]

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§2-5. Learning and pleasure

When Carol was trying to fill her pail, she had to try several experiments before she succeeded by using her spoon. When she recognized that her goal was achieved, she felt satisfaction and a sense of reward—and then those pleasant feelings somehow helped her to learn and remember. So this process involved a good many steps:

> *Carol filled the pail with her spoon.*
> *She recognized that her goal was achieved.*
> *Then she felt pleased with her success.*
> *Then, somehow, that pleasure helped her to remember.*

Now we're glad that she felt gratified—but what functions did all those feelings serve, and why should that process take so many steps? What sort of role might pleasure play in how we construct our memories? Why couldn't Carol just simply remember which methods worked and which ones failed?

The answer is that 'remembering' is not simple at all. On the surface, it might seem easy enough—like dropping a note into a box, and then taking it out when you need it. But when we look more closely at this, we see that it involves a good many steps: You first must select which items that note should contain, and find adequate ways to represent them—and then you must give them some set of connections, so that after you store those parts away, you'll be able to reassemble them.

Citizen: Some say that our brains remember everything so, that if you cannot recall some event, some part of your brain must be suppressing it.

This 'photographic memory' myth is not supported by evidence; the

consensus from many experiments is that we don't remember nearly so much. [See *§6-2*]

Student: What about the old idea that, for each of our accomplishments, we just 'reinforce' our successful reactions? In other words, we simply connect the problem we faced to the actions that led to our solving it.

That is a simplistic way to describe how learning might work, when seen from outside—but it doesn't explain what might happen inside. For, neither *'the problem we faced'* nor *'the actions we took'* are simple units that we can connect —so, first you must choose some way to describe both the *'If'* and the *'Then'* of that pair of events. Then, the quality of what you learn will depend on the natures of both those descriptions.

Thus, for Carol to learn, her brain must construct some descriptions of which methods worked—as well as, perhaps, of which methods failed. But after her struggle to fill her cup, which of all the things she did should get credit for her final success? Should Carol attribute her success to which pair of shoes she was wearing then, or the place in which that event occurred, or whether the weather was cloudy or clear? What if she smiled while using that fork, but happened to frown when using that spoon; what keeps her from learning irrelevant rules like, *"To fill a cup, it helps to frown?"*

In other words, when humans learn, it is not just a matter of making connections but of constructing the structures that those connections connect— and no theory of learning can be complete unless it also accounts for this. Furthermore, we may need to represent not only those external events, but also some relevant *mental* events. Thus Carol will need some machinery to decide which of *the thoughts she was thinking then* should be represented in what she remembers. And she will need some ways to store those records so that she can recollect when she needs them.

Student: You still haven't explained where feelings come in, such as the pleasure that comes from Carol's success.

In everyday life it's convenient to use terms like *suffering, pleasure, joy,* and *grief* as though those words referred to mental states that all our acquaintances are familiar with. But when asked to describe those states of mind, we usually find ourselves lost for words because the mental conditions that we call feelings are such complex cascades of processes. For example, it would seem that we speak about *pleasure* when certain resources recognize some processes that help us identify which of our recent activities should get credit for some recent success. Near the end of the book we'll return to these questions about how we

make those *'Credit Assignments'* and what are the effects of the feelings called *pleasure*.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§2-6. Conscience, Values and Self-Ideals

"I did not, however, commit suicide, because I wished to know more of mathematics."
–Bertrand Russell

One way that we differ from animals (except, perhaps, for the elephants) is in the great length of our childhoods. One consequence of this is that no other species accumulates so much and so many kinds of knowledge—and none of them seem to grow anything close to our human traditions and values.

What kind of person would you like to be? Are you careful and cautious or brave and audacious? Do you follow the crowd, or prefer to lead? Would you rather be tranquil or driven by passion? Such personal traits depend, in part, upon each person's inheritance. But also they are partly shaped by our networks of social attachments.

Once our human attachment bonds form, they begin to serve multiple functions. First they keep children close to their parents—and this provides such services as nutrition, defense, and companionship. But also (if we are right about this) they have special effects on how children learn—by providing each child with new ways to re-arrange its priorities. Also, the self-conscious emotions that come with this have other, very specific effects. Pride tends to make you more confident, more optimistic and adventurous, while Shame makes you want to change yourself so that you'll never get into that state again.

The following section discusses what happens when children's Imprimers go absent; the result of this can be severe. But older children and adults can envision how an absent imprimer might react to unusual acts or ideas, or evaluate a proposed new goal. We all know this kind of experience: of predicting (and then reacting to) what we think that an absent Imprimer might do—and then we give this various names like 'moral sense' or 'conscience' or 'knowledge of right and wrong.'

To do this kind of 'internal impriming,' a child will have to construct some sort of 'model' that helps to predict its Imprimer's reactions. How might that

child think about this? First, it might not think about it at all, because the rest of its mind has no access to it. Or, that model might seem, to that child, as though there were someone else in its mind—perhaps in the form of a made-up companion. It might even be seen as embodied in some external object—such as a rag doll or a baby-blanket.[13] We'll discuss such models in §9.

What if some other part of that child's brain could find a way to take over control of the systems that raise or elevate the priorities of its various goals? Then such a child could praise itself, and through those connections could select which new goals to elevate—or else that child could censure itself, and thus impose new constraints on itself.

At this point that child will have, in effect, an internal system of values—or what is commonly called a 'conscience'. Perhaps Freud had a process like this in mind when suggesting that a child can 'introject' some of its parents' attitudes. If a child gains enough control of this, it could become 'ethically autonomous' in the sense that it could eventually replace those earlier value-sets. However, if most of those values remain in place, then later attempts to change them could lead to internal conflicts in which the child tries to oppose the values acquired from its imprimers.

What determines which ideals will grow inside each particular human mind? Each family, culture, club, or group evolves various social and moral codes—by inventing some ways to decide what is right and wrong. Those codes of behavior have awesome effects on all of our organizations; they shape the customs, traditions, and cultures of nations, professions, clubs, and religions. They can even make those institutions value *themselves* above everything else—and make their members happy to die for them, in endless successions of battles and wars.

How do we grow those powerful standards and codes? I'll parody several philosophers.

Naturist: I deeply believe that ethical values are, by their nature, self-evident. Surely everyone would be naturally good, unless their minds were corrupted by having been raised in unnatural states.

Rationalist: I'm suspicious of statements like that because 'deeply' and 'self-evident" seem only to mean, "I cannot explain why I believe this," and, "I don't want to know what makes me believe it."

Social Contractor: There is no absolute basis at all for what we call moral and ethical values. They're all based on social conventions and contracts that each of us makes with the rest of us.

Socio-biologist: That's a neat concept—except for one thing: no one remembers agreeing to it! A better idea is that 'morals' are based on traits we

evolved in ancient times, as when certain breeds of dogs were selected for becoming attached to only one master. In humans, we call this 'loyalty'.

No doubt, such traits are partly based on genes that we have inherited, but they're also based on contagious 'memes'—that is, ideas that spread from one brain to the next as part of each cultural heritage.[14]

Fundamentalist: Our values stem directly from divinely inspired religious texts—and woe unto those who transgress them.

Theologian: Some ethical rules can be deduced on the basis of logical reasoning.

Logician: Logic only helps us deduce what's implied by the assumptions we make. It says nothing about which assumptions to use.

Mystic: Reasoning only clouds the mind and disconnects it from reality. You will never achieve enlightenment until you learn not to think so much.

One can sometimes improve a skill by suppressing the urge to think about it. But if one turns most mental critics off, and relies on primitive instincts too much, that could retard one's mental development.

Existentialist: Whatever goal you happen to have, you should ask what purpose that purpose serves—and then you'll see our predicament: we're all trapped in a world that's completely absurd.

Sentimentalist: You're too concerned with a person's aims. Just watch some children and you will see curiosity and playfulness. They're not seeking any particular goals, but are enjoying the finding of novelties, and the pleasures of making discoveries.

We like to think that a child's play is unconstrained—but when children appear to feel joyous and free, this may merely hide from their minds their purposefulness; you can see this more clearly when you attempt to drag them away from their chosen tasks. For they are exploring their worlds to see what's there, making explanations of what those things are, and imagining what else could be; exploring, explaining and learning are among a child's most purposeful urges and goals. The playfulness of childhood is the most demanding teacher we have. Never again in those children's lives will anything drive them to work so hard.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§2-7. Attachments of Infants and Animals

> *"We want to make a machine that will be proud of us."*
>
> *—Danny Hillis, 1983*

The young child Carol loves to explore, but also likes to stay near to her mother—so whenever the distance between them grows, she quickly moves herself closer. But should she discover that she is alone, she'll shortly cry out and look for her mum. That same behavior will also appear even when her mother is near, if there's any cause for fear or alarm—such as the approach of a stranger.

Naturally, this dependency stems from our infantile helplessness: no human infant would long survive if it could escape from parental care. Of course, this doesn't happen because young infants cannot move much by themselves—but this comes with the disadvantage that, in those first few months, our infants also can't follow their mothers. Fortunately we usually come to no harm from this because we evolve a second bond that goes in the other direction: Carol's mother is almost always aware (to different extents at various times) of what is happening to her daughter—and her full attention is quickly engaged at the slightest suspicion that something is wrong.

Clearly, each infant's survival depends on bonding to persons concerned with their welfare. So in older times it was often assumed that *children would attach themselves to the persons who gave them physical care*, and this is why most psychologists called such a person a 'Caregiver'—instead of using some word like 'Imprimer'. But more systematic research on attachment suggested that this theory was wrong:

John Bowlby: "That an infant can become attached to others of the same age, or only a little older, makes it plain that attachment behavior can develop and be directed towards [persons who have] done nothing to meet the infant's physiological needs."[15]

Then what factors *do* determine the persons to whom our children will become attached? First, Bowlby recognized that physical nurture could play an important role, because it provides occasions for children to learn to like to be with particular other persons. But eventually he concluded that usually, these

were more important factors:[16]

> *The speed with which the person responds,* and
> *The intensity of that interaction."*

This will usually include the child's parents—but could also include other children, which suggests that parents should take special care to examine their offspring's companions and friends—and, especially, the ones that are most attentive to them. And when one is choosing a child's school, one might examine not only the staff and curriculum, but also the goals that its pupils pursue.

What happens when a child is deprived of Imprimers? It appears that an Imprimer's absence produces a special variety of fear, and a powerful impulse to find that Imprimer.

> *John Bowlby: "Whenever a young child … is separated from her mother unwillingly he shows distress; and should he also be placed in a strange environment and cared for by a succession of strange people such distress is likely to be intense. The way he behaves follows a typical sequence. At first he protests vigorously and tries by all the means available to him to recover his mother. Later he seems to despair of recovering her but nonetheless remains preoccupied with her and vigilant for her return. Later still he seems to lose his interest in his mother and to become emotionally detached from her."*

Bowlby goes on to describe what happens when the mother comes back:

> *"Nevertheless, provided the period of separation is not too prolonged, a child does not remain detached indefinitely. Sooner or later after being reunited with his mother his attachment to her emerges afresh. Thenceforward, for days or weeks, and sometimes for much longer, he insists on staying close to her. Furthermore, whenever he suspects he will lose her again he exhibits acute anxiety."[17]*

We see similar attachment behavior in our various primate relatives—such as chimpanzees, gorillas, and orangutans—as well as in our more distant cousins, the monkeys. We should also note Harry Harlow's discovery that, given no other alternative, a monkey will become attached to an object that has no behavior at all, but does have some 'comforting' characteristics. This would seem to confirm Bowlby's view that attachment does not stem from 'physiological needs'— unless we amend this to include the infant's need for what Harlow calls *comfort contact*.[18]

John Bowlby: *"The very detailed observations made by Jane Goodall of chimpanzees in the Gombe Stream Reserve in central Africa show not only that anxious and distressed behavior on being separated, as reported of animals in captivity, occurs also in the wild but that distress at separation continues throughout chimpanzee childhood. ... Not until young are four and a half years of age are any of them seen traveling not in the company of mother, and then only rarely."*

— *[John Bowlby, p. 59 Separation.]*

When the mother and child have more distance between them, they maintain their connection with a special 'hoo' whimper to which the other promptly responds—as Jane Goodall herself reports:

"When the infant ... begins to move from its mother, it invariably utters this sound if it gets into any difficulty and cannot quickly return to her. Until the infant's locomotion patterns are fairly well developed the mother normally responds by going to fetch it at once. The same sound is used by the mother when she reaches to remove her infant from some potentially dangerous situation or even, on occasion, as she gestures it to cling on when she is ready to go. The 'hoo' whimper therefore serves as a fairly specific signal in re-establishing mother-infant contact."[19]

What happens in other animals? Early in the 1930s Konrad Lorenz observed that a recently hatched chicken, duck, or goose will become "attached" to the first large moving object it sees, and will subsequently follow that object around. He called this "imprinting" because it occurs with such remarkable speed and permanence. Here are some of his observations.[20]

The chick quickly starts to follow the moving object.
Imprinting begins soon after hatching.
The period for imprinting ends a few hours later.
The effect of imprinting is permanent.

To what objects will the chick get attached? That moving object will usually be a parent—but if the parents have been removed, then the object could be a cardboard box, or a red balloon—or even Konrad Lorenz himself. During the next two days, as the gosling follows its parents, it somehow learns to recognize them as individuals and not follow any other geese. Now when it loses contact with the mother it will cease to feed or examine things, and instead will search and make piping sounds, as though distressed at being lost. Then the parent responds with a special sound—and Lorenz observes that this response must

come quickly to establish imprinting. Later this call is no longer needed, but in the meantime it serves to protect the chick against becoming attached to an unsuitable object, such as the moving branch of a tree.

These 'piping' sounds, like the 'hoo' signals in Jane Goodall's notes, suggest that other ways to communicate could have co-evolved from attachment signals. In any case, these types of birds can feed themselves soon after they hatch—so imprinting is independent of being fed.

As for when the imprinting period ends, R.A.Hinde discovered that those chicks eventually become fearful of unfamiliar moving things—which led him to suspect that imprinting stops when this new fear forestalls further 'following'. Similarly, many human babies show a long period of fear of strangers that begins near the start of the second year.[21]

Bowlby's research on young children showed that when they are deprived of imprimers for more than a few days, they may show signs of impairments for much longer times. He also cites similar results when other researchers separated infant Rhesus monkeys from their mothers:

"From all these findings[22] we can conclude with confidence not only that a single separation of no longer than six days at six months of age has perceptible effects two years later on rhesus infants, but that the effects of a separation are proportionate to its length. A thirteen-day separation is worse than a six-day; two six-day separations are worse than a single six-day separation."
—Bowlby, in Separation, p. 72

Remarkably, even badly mistreated children (and monkeys) may remain attached to abusive imprimer.[23]

To what extent did human attachment-based learning evolve from older forms of pre-human imprinting? Of course, humans are very different from birds, yet the infants of both share similar needs—and there may have been precursors of this in some earlier warm-blooded dinosaurs. For example, Jack Horner[24] discovered that some of these constructed clusters of bird-nest like structures. Further progress in genomics might help us reconstruct more of this history.

Returning to the human realm, we should ask how infants distinguish potential imprimers. Although some researchers have reported that infants can learn to recognize the mother's voice even before the time of birth, it is generally though that newborns first learn mainly through the senses of touch, taste, and smell—and later distinguish the sound of a voice and start to react to the sight of a head or a face. One first might assume that this is done by detecting features

like eyes, nose, and mouth, but there is evidence that it is more complex than that.[25]

Francesca Acerra: *"4-day-old neonates look longer at their mother's face than at a stranger's face—but not when the mother wears a scarf that hides the hair contour and the outer contour of the head."*[26]

This researcher found that those infants react less to the features of the face, and more to its larger-scale, overall shape; it was not until two or three more months that her subjects distinguished particular faces.[27] This suggests that our visual systems involve different methods at different stages of development— and perhaps the ones that are first to operate serve mainly to get the mother attached to the child! In any case, Lorenz was amazed by what his goslings *failed* to distinguish:

Konrad Lorenz: *"The human imprinted gosling will unequivocally refuse to follow a goose instead of a human, but it will not differentiate between a petite, slender young girl and a big old man with a beard. ... It is astounding that a bird reared by, and imprinted to, a human being should direct its behavior patterns not towards one human but towards the species Homo sapiens."*[28]

I don't find this so strange because all geese look almost the same to me. Perhaps more important is that adult sexual preference may be established at this early time, though it only much later appears in behavior.

"A jackdaw for which the human has replaced the parental companion, will thus direct its awakening sexual instincts not specifically towards its former parental companion, but ... towards any one relatively unfamiliar human being. The sex is unimportant, but the object will quite definitely be human. It would seem that the former parental companion is simply not considered as a possible 'mate'."

Some studies have shown that after such contact, some of those birds will eventually mate with other members of their species. However, this phenomenon is still such a serious problem in repopulating endangered species that it has become the standard policy to minimize human contact with chicks, lest their preference for people lead them to later refuse to mate with their peers. Could such delays be relevant to human sexual preferences?

All of this could help to explain why we evolved our extended infantile helplessness: children who too soon went off by themselves would not have been wise enough to survive—and so, we had to extend the time for learning from imprimers instead of from doing risky experiments.

§2-8. Who are our Imprimers?

> *A JACKDAW, seeing Doves in a place with much food, painted himself white to join them. The Doves, as long as he did not speak, assumed that he was another Dove and admitted him to their cote. But when one day he forgot not to speak, they expelled him because his voice was wrong—and when he returned to his Jackdaw tribe they expelled him because his color was wrong. So desiring two ends, he obtained neither.*
>
> *—Aesop's Fables*

How many Imprimers can a person have? Many young children have only one, while others may have two, three, or more. Then when a child has several of them, are those attachments interchangeable—or could they serve different functions and goals? If a child forms several sets of ideals, would that enrich its personality—or would it impair its development because those inconsistencies prevent it from forming a single coherent self-image?[29]

When do attachments begin and end? Even young infants soon start to behave in distinctive ways when in their mothers' presence. However, it is usually not till near the first year's end that the child protests against separation —and begins to learn to become disturbed at a sign that Imprimer *intends* to depart—e.g., reaching for an overcoat. This is also the time when most children begin to show fears of unusual things. Both this and that fear of separation begin to decline in the child's third year—so that now the child can be sent to school. However, we do not see the same decline in the roles of those other, self-conscious, attachment-based feelings. These persist for longer times and sometimes, perhaps, for the rest of our lives.

John Bowlby: "During adolescence … other adults may come to assume an importance equal to or greater than that of the parents, and sexual attraction to age-mates begins to extend the picture. As a result individual variation, already great, becomes even greater. At one extreme are adolescents who cut themselves off from the parents; at the other are those who remain intensely attached and are unwilling or unable to direct their attachment behavior to others. Between

the extremes lie the great majority of adolescents whose attachments to parents remain strong but whose ties to others are of much importance also. For most individuals the bond to parents continues into adult life and affects behavior in countless ways. Finally in old age, when attachment behavior can no longer be directed to members of an older generation, or even the same generation, it may come instead to be directed towards members of a younger one."

[Bowlby, *Attachment*, p207]

What happens in other animals? In those that do not remain in herds, attachment frequently only persists until the offspring can live by themselves. In many species it's different for females; in many species the mother will actively drive the young ones away as soon as a new litter is born (perhaps because of evolutionary selection against inbreeding)—while in other cases attachment will stay until the time of puberty or even later for females. In *Attachment (p182)* Bowlby mentions a phenomenon that results from this:

"In the female of ungulate species (sheep, deer, oxen, etc.), attachment to mother may continue until old age. As a result a flock of sheep, or a herd of deer, is built up of young following mother following grandmother following great grandmother and so on. Young males of these species, by contrast, break away from mother when they reach adolescence. Thenceforward they become attached to older males and remain with them all their lives except during the few weeks of each year of the rutting season."

Of course, other species evolve different strategies that are better suited for different environments; for example, the size of the flock may depend on the character and prevalence of predators, *etc.*

Why should we need Imprimers at all—and why should we be so exclusive in how our brains make us choose them? Why not simply elevate goals in response to anyone's censure or praise? There's an excellent reason why we evolved to be selective about this—for if any stranger could program your goals, you'd be in danger because strangers are less likely than your close relatives are to be concerned for your welfare.

However, 'welfare' can mean different things. For example, Bowlby argued that our attachments mainly promote our children's physical safety. Here's a paraphrase of his argument"

"That protection from predators is by far the most likely function of attachment behaviour is supported by three main facts. First an isolated animal is much more likely to be attacked than is one that stays bunched together with others of its kind. Second, attachment behavior is especially easy to arouse in

animals that, by reason of age, size, or conditions are especially vulnerable to predators. Third, this behavior is strongly elicited in situations of alarm, which are commonly ones in which a predator is sensed or suspected. No other theory fits these facts."

Here, Bowlby's main concern was to refute the then popular view that attachment's primary function was to ensure a dependable source of food. Instead, he argued that physical care (including nutrition) did not play a crucial role in attachment and security was the more influential. I suspect that this was largely correct for animals, but does take into account how human attachments so strongly promote our acquiring values and high-level goals.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§2-9. Self-Models and Self-Consistency

To solve a hard problem, one must work out a plan—but, then, you need to carry it out; it won't help to have a multi-step plan if you tend to quit before it is done. This means that you'll need some 'self-discipline'— which in turn needs enough self-consistency that you can predict, to some extent, what you're likely to do in the future. We all know people who make clever plans but rarely manage to carry them out because their models of what they will actually do don't conform enough to reality. But how could a trillion-synapse machine ever become predictable? How did our brains come to manage themselves in the face of their own great complexity? The answer must be that we learn to represent things in extremely simple, yet useful ways.

Thus, consider how remarkable it is that we can describe a person with words. What makes us able to compress an entire personality into a short phrase like "Joan is tidy," or "Carol is smart," or "Charles tries to be dignified"? Why should one person be *generally* neat, rather than be tidy in some ways and messy in others? Why should traits like these exist? In *§9-2 Traits* we'll see some ways in which such things could come about:

In the course of each person's development, we tend to evolve certain policies that are so consistent that we (and our friends) can recognize them as features or traits—and we use these to build our self-images. Then when we try to formulate plans, we can use those traits to predict what we'll do (and to thus discard plans that we won't pursue). Whenever this works we're gratified, and this leads us to further train ourselves to behave in accord with these simplified descriptions. Thus, over time our imagined traits proceed to make themselves

more real.

Of course, these self-images are highly simplified; we never come to know very much about our own mental processes, and what we call traits are only the few consistencies that we learn to perceive. However, even these may be enough to help us conform to our expectations, so that this process can eventually provide us with enough of what we call "Self-Reliance."

We all know the value of having friends who usually do what they say they will do. But it's even more useful to be able to trust *yourself* to do what you've asked yourself to do! And perhaps the simplest way to do that is to make yourself consistent with the caricatures that you've made of yourself—by behaving in accord with self-images described in terms of sets of traits.

But how do those traits originate? Surely these can be partly genetic; we can sometimes perceive newborn infants to be more placid or more excitable. And, of course, some traits could be the chance results of developmental accidents. However, other traits seem more clearly acquired from contacts with one's imprimers.

Is there some risk in becoming attached to too many different personalities? That could lead to attempting to model yourself on too many different sets of traits; a person with coherent goals should usually do better than one encumbered by conflicts because of having more time to acquire the skills to achieve them all—and consistency also makes others feel safe in depending on you.

This also applies inside ourselves: if we changed our minds too recklessly, we could never predict what we might want next. We'd never be able to get much done if we could not "depend on ourselves." However, on the other side, we need to be able to compromise; it would be rash to commit to some long-range plan with no way to later back out of it. It would be especially dangerous to change oneself in ways that prevent ever changing again.

If a child has only a single Imprimer—or several with more or less similar values—it won't be too hard for the child to learn which behaviors will usually be approved. However, if the child's Imprimers have conflicting goals, this could make it difficult for the child to decide which to elevate—or to end up with so many different ambitions that very few of these will develop well. Nevertheless, eventually we each must deal with persons with diverse ideas, so there can be advantages to having diverse collections of models.

Most imprimers will be concerned with the values their children acquire, hence may will try to keep them from attaching themselves to persons of 'dubious character'. Here is an instance in which we see just such a concern

about a certain researcher's machine!

In the 1950s, Arthur Samuel, a computer designer at IBM, developed a program that learned to play Checkers well enough to defeat several excellent human players. Its quality of play improved when it competed with its superiors. However, games against inferior players tended to make its performance get worse—so much that its programmer had to turn its learning off. In the end he allowed it only to play against transcripts of master-class championship games.

When anyone interacts with you, they're likely to have their own purposes, so you have to try to assess their intentions. Consider how members recruit for their cults. First they remove you from all familiar locations, and then persuade you to 'decide' to break all your other social attachments—especially all your family ties. Then once you're detached from all your friends it becomes easy to undermine all your defenses—until you are ripe to be imprimed by their local prophet, seer, or saint. Those experts do indeed know schemes through which any stranger can program you—by exploiting techniques that they know can help to suppress and supplant the ideals that you hold.

We face similar conflicts in other realms. While your parents may have your welfare in mind, businesspersons may have more interest in promoting the wealth of their firms. Religious leaders may wish you well, yet be more concerned for their temples and sects. And when leaders arouse your pride in your nation, you may be expected to sacrifice your life to define some vague boundary line. Each organization has its own intentions, and uses its members to further them.

Individualist: I hope you don't mean that literally. An organization is nothing more than the circle of persons involved with it. It cannot have any goals of its own, but only those that its members hold.

What does it mean when someone suggests that some system has an intention or goal? Section §6-3 will discuss some conditions in which a process could appear to have motives and purposes of its own.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§2-10. Public Imprimers

We've only discussed how attachment-based learning might work when a child is with an Imprimer. It might also be related to the phenomenon in which hordes of persons are influenced by others who 'catch the public's eye' by

appearing in broadcast media. One way to make a person feel that something is desirable would be to put forward some evidence. However, it appears to be more effective to use the so-called 'testimonial', which may not exhibit the product at all, but only suggests that its use is approved by some popular 'celebrity'. Why would this method work so well? Perhaps because those particular persons have ways to evoke an impriming response and thus more directly modify the personal goals of their audiences.

Sociologist: Perhaps this happens only because when the 'celebrity' takes the 'center stage' this makes other people focus there. Then once most of the audience gets engaged, the rest feel compelled to join them.

That may be what happens, but still we should ask what makes our 'celebrities' popular. Attractive physical features may help, but those actors and singers use something else: they are experts at feigning emotional states. Competitive athletes are expert deceivers, and so must be our popular leaders. More generally, perhaps, to achieve celebrity, it helps to have some special ways to make each listener feel some sense that *"this important person is speaking to me."* That would make listeners feel more involved—and therefore more compelled to respond —despite that it's really a monologue!

Not everyone can control a mob. What techniques could so firmly engage the concern of such a wide range of different minds? We need to know more about the tricks that our leaders use to mould our goals. Could these include some methods through which they can establish rapid attachments?

Charisma: n. *'a rare personal quality attributed to leaders who arouse popular devotion or enthusiasm.'*

What characteristics give leaders the power to evoke that sense of charisma? Are there some special physical features that act as 'charismatic releasers'?

Politician: It usually helps for the speaker to have large stature, deep voice, and confident manner. However, although great height and bulk attract attention, some leaders have been diminutive. And while some powerful orators intone their words with deliberate measure, some leaders and preachers rant and shriek, and still manage to grip our attention.

Psychologist: In §2-7 you mentioned that 'speed and intensity of response' were important for making attachments. But when someone makes a public pronouncement, there isn't much room for those critical factors because the speaker cannot react specifically to each and every listener.

Rhetoric can create that illusion. A well-paced speech can seem 'interactive'

by first raising questions in listeners' minds—and then answering them at just the right time. You don't have time to converse with each, but you can interact—inside your mind—with a few model listeners. Then many real listeners may feel the sense of receiving a personalized response, although there's no genuine dialog. One trick is for speakers to pause just long enough for listeners to feel that they are being addressed, but not long enough for them to think of objections to the messages that they hear. Furthermore, an orator may not need to control the whole audience; if you can convince enough of them, then peer pressure can make most of the others to with them.

However, a crowd can take over control of a weaker and over-responsive leader. Here's one great performer who objected to this:

Glenn Gould: "For me, the lack of an audience—the total anonymity of the studio—provides the greatest incentive to satisfy my own demands upon myself without consideration for, or qualification by, the intellectual appetite, or lack of it, on the part of the audience. My own view is, paradoxically, that by pursuing the most narcissistic relation to artistic satisfaction one can best fulfill the fundamental obligation of the artist of giving pleasure to others."[30]

A person can even become attached to an entity that doesn't exist—for example, to a legendary historical figure, to a fictional character in a book, or to a mythical martyr, dog, or god. Then those heroes can become "virtual mentors" among the models in their worshippers' minds. A person can even become attached to an abstract doctrine, dogma, or creed—or an icon or image that represents it. Indeed, when you come right down to it, *all* our attachments are made to fictions; you never connect to an actual person, but only to the models you've made to represent your conceptions of them, no matter whether they're parent or friend—or merely a transient attraction.

So, the idea that a person learns goals from Imprimers makes sense in the earliest years of life. However, in later life that distinction can fade, as we acquire new kinds of mentors and find other ways to shape our ambitions.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

Summary: This chapter addressed some questions about how people acquire the goals they pursue. Some of these are instincts that come with our genetic inheritance, but others are subgoals that we construct to achieve other goals that we already have. I also conjectured that some of our highest-level goals are produced by special machinery that lead each person to try to adopt the values of other persons who become what I call that person's "Imprimers."

Imprimers are parents, friends, or acquaintances to whom a person becomes 'attached,' because they respond actively to one's needs—and they then can induce special feelings in us, such as *guiltiness, shame,* and *pride*. At first, those Imprimers must be actually present, but older children form 'mental models' of them, and can use these to evaluate goals when those imprimers no longer are on the scene. Eventually, these models later develop into what we call by names like conscience, values, ideals, and ethics.

The next chapter will look more closely at the clusters of feelings and thought that we know by such names as *hurting*, *grief*, and *suffering*—to see how they might be understood as varieties of ways to think.

(I should note that this chapter's ideas about Imprimers are only theories of mine, and don't yet appear in psychology books. These ideas might be right but they also might not.)

Part III. From pain to suffering

§3-1. Being in Pain

> *"Great pain urges all animals, and has urged them during endless generations, to make the most violent and diversified efforts to escape from the cause of suffering. Even when a limb or other separate part of the body is hurt, we often see a tendency to shake it, as if to shake off the cause, though this may obviously be impossible."*
>
> —*Charles Darwin*[31]

What happens when you stub your toe? You've scarcely felt the impact yet, but you catch your breath and start to sweat—because you know what's coming next: a dreadful ache will tear at your gut and all other goals will be brushed away, replaced by your wish to escape from that pain.

Why does the sensation called *pain* sometimes lead to what we call *suffering*? How could such a simple event distort all your other thoughts so much? This chapter proposes a theory of this: if a pain is intense and persistent enough, it will stir up a certain set of resources, and then these, in turn, arouse some more. Then, if this process continues to grow, your mind becomes a victim of the kind of spreading, large-scale "cascade" that overcomes the rest of the mind, as we depicted in §1-7:
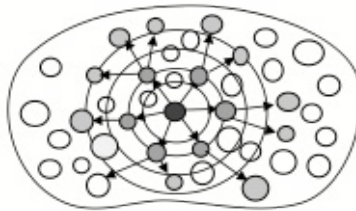
Now, sometimes a pain is just a pain; if it's not too intense or doesn't last long, then it may not bother you much. And even if it hurts a lot, you can usually muzzle a pain for a time, by trying to think about something else. And sometimes you can make it hurt less by thinking about the pain itself; you can

focus your attention on it, evaluate its intensity, and try to regard its qualities as interesting novelties.

Daniel Dennett: "If you can make yourself study your pains (even quite intense pains) you will find, as it were, no room left to mind them: (they stop hurting). However studying a pain (e.g., a headache) gets boring pretty fast, and as soon as you stop studying them, they come back and hurt, which, oddly enough, is sometimes less boring than being bored by them and so, to some degree, preferable."

But this only provides a brief reprieve, because until your pain goes away, it may continue to gripe and complain, much like a nagging frustrated child; you can think about something else for a time, but no matter what kinds of diversion you try, soon that pain will regain its control of your mind.

Still, we should be thankful that pain evolved, because it protects our bodies from harm. First, as Darwin suggests above, this may induce you to shake off the cause of the pain—and it also may keep you from moving the injured part, which may help it to rest and repair itself. However, consider these higher-level ways through which pain may protect us from injury.

Pain focuses your attention on the particular body-parts involved.
It makes it hard to think about anything else.
Pain makes you tend to move away from whatever is causing the stimulus.
It makes you want that state to end, and it makes you learn, for future times, not to repeat the same mistake.

Yet instead of being grateful for pain, people always complaining about it. *"Why are we cursed,"* pain's victims ask, *"with such unpleasant experiences?"* We often think of pleasure and pain as opposites—yet they share many similar qualities:

Pleasure makes you focus on the particular body-parts involved.
It makes it hard to think about anything else.
It impels you to draw closer to whatever is causing the stimulus.
It makes you want to maintain that state, while teaching you, for future times, to keep repeating the same mistake.

This suggests that both pleasure and pain could engage some of the same kinds of machinery. For example, they both tend to narrow one's range of attention, they both have connections with how we learn, and they both assign high priority to just one of a person's many goals. In view of those similarities, a visiting alien intelligence might wonder why people like pleasure so much—yet display so little desire for pain.

Alien: Why do you humans complain about pain?
Person: We don't like pain because it hurts.
Alien: Then explain to me just what 'hurting' is.
Person: Hurting is simply the way pain feels.
Alien: Then please tell me what a 'feeling' is.

At this point the conversation may stop, because quite a few human thinkers might claim that we'll *never* have ways to explain such things, because feelings are 'irreducible.'

Dualist Philosopher: *Science can only explain a thing in terms of other, yet simpler things. But subjective feelings like pleasure or pain are, by their nature, indivisible. They can't be reduced to smaller parts; like atoms, they simply are or are not.*

This book will take the contrary view that feelings are not simple at all; instead they are extremely complex. And paradoxically, once we recognize this complexity, this can show us ways to explain why pleasure and pain might seem similar if (as we'll try to show in Chapter §9) we can represent both of them as results that come from similar kinds of machinery. [Also, see §§Dignity of Complexity.]

People often use *hurting*, *pain* and *suffering* as though those conditions were almost the same, and differ mainly in degree. This chapter will argue that we need much better distinctions and theories for these.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§3-2. Prolonged Pain leads to Cascades

Our idea about how Suffering works is that any severe and prolonged pain leads to a cascade of mental change that disrupts your other plans and goals. By thus suppressing most other resources, this narrows your former interests—so that most of your mind now focuses on one insistent and overwhelming command: *No matter what else, get rid of that Pain.*

This machinery has great value indeed—if it can make you remove whatever's disturbing you, so that you get back to what you were trying to do. However, if that pain remains intense *after* you've done all you can to relieve it, then it may continue to keep the resources that it has seized—and further to proceed to capture yet more—so that you can scarcely keep anything else 'on your mind'. If left to itself, that spreading might cease—but so long as the pain refuses to leave, that cascade of disruption may continue to grow, and as those other resources get taken away, your efforts to think will deteriorate, and what remains of the rest of your mind may feel like it's being sucked into that black hole of suffering.
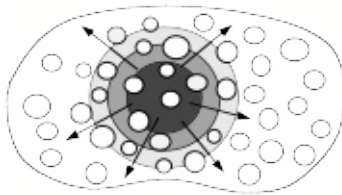
Now, goals that seemed easy in normal times get increasingly harder to achieve. Whatever else you try to do, pain interrupts with its own demands and keeps frustrating your other plans until you can barely think about anything but the pain and the trouble it's caused. Perhaps the torment of suffering comes largely from depriving you of your freedom to choose what to think about. Suffering imprisons you.

Neurologist: These ideas about disruptive cascades are suggestive, but have you any evidence that processes like these exist? How could you show that these guesses are right?

It would be hard to demonstrate this today, but when scanners show more of what happens in brains, we should be able to *see* those cascades. In the

meantime, though, one scarcely needs more evidence than one sees in the diversity of the complaints from the victims of suffering:

Frustration at not achieving goals.
Annoyance at losing mobility.
Vexation at not being able to think.
Dread of becoming disabled and helpless.
Shame of becoming a burden to friends.
Remorse at dishonoring obligations.
Dismay about the prospect of failure.
Chagrin at being considered abnormal.
Resenting the loss of opportunities.
Fears about future survival and death.

This suggests that we learn to use words like 'suffering', 'anguish', and 'torment' to try to describe what happens when those disruption cascades continue: as each new system becomes distressed and starts to transmit disturbing requests, your normal thoughts get overcome, until most of your mind has been stolen from you.

Citizen: I agree that these all can come with suffering. But that doesn't explain what suffering is. To be sure, resentment, remorse, dismay, and fear are all involved with reactions to pain—and can help to cause us to suffer. But why can't we just regard 'suffering' as just one more kind of sensation?

When we talk about 'sensations' we usually mean the signals that come from sensors that are excited by conditions in the external world. However here, I think, we're talking about signals that come, not from outside, but from special resources that detect high-level conditions inside the brain. Later, in section §4-3, we'll suggest how such resources might actually work.

In any case, when suffering, it is hard to think in your usual ways. Now, torn away from your regular thoughts, you can scarcely reflect on anything else than on your present state of impairment—and awareness of your dismal condition only tends to make things worse. Pain, as we said, deprives you of freedom, and a major component of suffering is the frustration that accompanies the loss of your freedom of mental choice.

Of course the same is true, to a smaller degree, in our more usual states of mind: our thoughts are always constrained by the goals that we hold, which try to engage different processes. Those processes sometimes cooperate, but they also frequently clash and conflict. We never have enough time to do all the things that we want to do—and so every new goal or idea that we get may make

us abandon, or put aside, some other ambitions we want to achieve.

Most times, we don't mind those conflicts much, because we feel that we're still in control, and free to make our own decisions—and if we do not like the result, we're still 'free' to go back and try something else. But when an aching pain intrudes, those projects and plans get thrust aside, as though by an external force[32] —and then we end up with more desperate schemes for finding ways to escape from the pain. Pain's urgency is useful to us when we need to deal with emergencies—but if it cannot be soon relieved, it then can become a catastrophe.

Indeed, suffering can affect you so much that your friends may see you being replaced by a different personality. It may even make you so regress that you cry out and beg for help, as though you've become an infant again. Of course, you may see yourself as still the same, and imagine that you still possess your old memories and abilities. But you won't be able to use those well until you switch back to your regular Self.

The primary function of Pain is to make one remove whatever may be causing it. To do this, though, it needs to disrupt most of one's other usual goals. Whenever this leads to a large-scale cascade, then we use words like 'suffering' to describe what remains of its victim's mind.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

The Machinery of Suffering

"The restless, busy nature of the world, this, I declare, is at the root of pain. Attain that composure of mind, which is resting in the peace of immortality. Self is but a heap of composite qualities, and its world is empty like a fantasy."
— *Buddha*

"Life is full of misery, loneliness, and suffering— and it's all over much too soon."
— *Woody Allen*

Yesterday Joan tripped on a step. She didn't suspect that she'd injured herself—but today she has just become aware of a terrible pain in her knee. She's been working on an important report and tomorrow she plans to deliver it. *"But if this keeps up,"* she hears herself think, *"I won't be able to take that trip."* She tries to make herself get back to work, but shortly she drops her pen and moans, *"I really must get rid of this pain."* She attempts to visit her medicine shelf, to find a pill that could bring some help, but a stab of pain makes her sit back down, and instructs her not to use that leg. She clutches her knee, catches her breath, and tries to think about what to do next—but the pain so overwhelms her mind that she can't seem to focus on anything else.

How does Joan know where her pain is located? That's easy to do for each place on her skin—because she is born with 'maps' of her skin in various different parts of her brain, like this one in the sensory cortex.
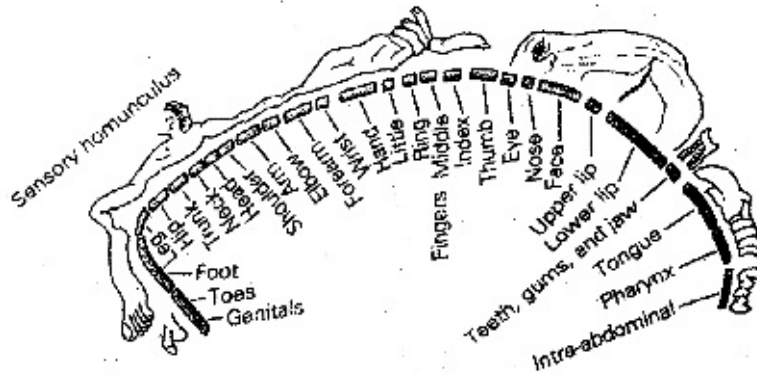
Many textbooks about the brain explain that those maps help us to determine the locations of tactile sensations—but those books don't ask what advantage we gain from having those maps—considering that the skin itself could serve for that. (We'll discuss this in *TopoQualia.*) However, we are not nearly so good at locating the causes of *interior* pains. It seems that our brains do not come equipped to represent the locations of structures *inside* our skins. Presumably, good maps for these have never evolved because they would not have been of much use to us: before the era of medicine, there was no way to protect one's spleen, except to guard one's whole abdomen—hence all one actually needed to know is when one had a bellyache. In particular, one never says, *"I feel a terrible pain in my brain,"* because we never had any remedies for injuries to the brain itself—so we never evolved any sense of pain in our brains, or of the spatial locations of mental events.

In any case, for Joan's pain to be useful to her, it must make her focus her thoughts on that knee—while also postponing her other goals. *"Get rid of Me,"* Joan's pain demands, *"and get back into your Normal State."* She won't be able to work on her report until she can satisfy that imperative.

How does our sense of pain actually work? Our scientists know quite a lot about the very first few events that result when a part of your body is traumatized. First, the injured cells release chemicals that cause a special type of nerve to send signals to your spinal cord. Then certain neural networks send other signals up to your brain. However, our scientists understand much less of what happens, then, in the rest of the brain. In particular, I've never seen any good high-level theories of how or why pain leads to suffering. Instead we find mainly descriptions like this:

The sense of pain originates when special nerves react to high temperature,

pressure, etc. Then their signals rise up to your thalamus, which sends them to other parts of your brain—in ways that on various ways involve hormones, endorphins, and neurotransmitters. Eventually, when some of those signals reach your limbic system, this results in such emotions such as sadness, anger, and frustration.

However, that doesn't explain what suffering is—because it isn't enough only to know which parts of the brain are involved with pain. We must also know what those parts do and how each affects the other ones, both when we're in our most usual states and (to make sense of suffering) when we're subject to larger cascades. Ronald Melzack and Patrick Wall, who pioneered theories of how pain works, cautiously note that:

"An area within the functionally complex anterior cingulate cortex has a highly selective role in pain processing, consistent with an involvement in the characteristic emotional/motivational component (unpleasantness and urgency) of pain.[33]

But we also know that that pain is involved with many other parts of the brain.[34] Thus Melzack and Wall go on to say,

"The concept [of a pain center] is pure fiction unless virtually the whole brain is considered to be the 'pain center' because the thalamus, the limbic system, the hypothalamus, the brain stem reticular formation, the parietal cortex, and the frontal cortex are all implicated in pain perception."

Furthermore, our reactions to pain depend on other mental conditions:

Daniel Dennett: "Real pain is bound up with the struggle to survive, with the real prospect of death, with the afflictions of our soft and fragile and warm flesh. ... There can be no denying (though many have ignored it) that our concept of pain is inextricably bound up with (which may mean something less strong than essentially connected with) our ethical intuitions, our senses of suffering, obligation, and evil."[35]

In general, we still do not know much about how physical pain leads to suffering. For although we have learned a good deal about *where* many functions are done in the brain, we still know very little about how each of those brain-parts actually *work*—because we still need theories (like those in this book) about what those resources actually *do*.

Perhaps we'll find more clues about such things in a rare condition that results from injuring certain parts of the brain: the victims of *'Pain Asymbolia'*

still recognize what the rest of us describe as pain—but do not find those feelings unpleasant, and may even laugh in response to them. Perhaps they have lost some resources that cause what, in others, are cascades of torments.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

Physical vs. Mental 'Pain'

Citizen: Physical pain is just one kind of pain—and emotional pains can be just as intense; they can even drive people to suicide. How could your theory also explain those other kinds of agonies?

Are mental and physical pains the same? They frequently seem to have similar ways to make changes in our mental states. What kind of relation could there be between how we react to, say, pinching or burning of the skin, and 'painful' events inside our minds, like,

> *The pain of losing a long-term companion.*
> *The pain of watching the pain of others.*
> *The pain of sleep deprivation.*
> *The pain of humiliation and perceived failure.*
> *The pain of excessive and prolonged stress.*

Suppose that you were to hear Charles say, *"I felt so anxious and upset that it felt like something was tearing my gut."* You might conclude that Charles's feelings reminded him of times when he had a stomachache.

Physiologist: It might even be true that 'your stomach crawled' —if your mental condition caused your brain to send signals to your digestive tract.

Similarly, we often speak as though 'hurt feelings' resemble physical pains, no matter that they originate from such different situation-types. This could be because, although they begin in different ways, both may end up by seizing control of the same higher-level machinery. Thus, disrespect on the part of a friend can disrupt your brain in much the same way as a deep, aching pain. And sometimes, what starts with physical pain can get amplified 'psychologically':

Student: As a child, I once I hit a chair with my head, and covered the area with my hand. Although the pain was intense, I was not much disturbed. But when I looked and saw blood on my hand, then I really panicked and started to cry.

In any case, most kinds of feelings are hard to describe because we know so

little about how their machinery works. However, it can be easy to *recognize* a mental state (either in yourself or in someone else) because you may only need to detect a few features of that particular mental condition. And this will often be enough to help us to communicate—by using what we call 'empathy.' For if two minds have enough structure in common, then just a few clues could lead each one to recognize some of the other's condition.

<div align="center">∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞</div>

§3-3. Feeling, Hurting, and Suffering

> *"As he thought of it, a sharp pang of pain struck through him like a knife and made each delicate fiber of his nature quiver. His eyes deepened into amethyst, and across them came a mist of tears. He felt as if a hand of ice had been laid upon his heart."*
> —Oscar Wilde in The Picture of Dorian Grey.

We have many words for types of pain—like stinging, throbbing, piercing, shooting, gnawing, burning, aching, and so on. But words never capture quite enough of what any particular feeling *is*, so we have to resort to analogies that try to describe what each feeling is *like*—such as 'a knife' or 'a hand of ice'—or images of a suffering person's appearances. Dorian Grey felt no physical pain, but was horrified about growing old—hideous, wrinkled, and worst of all, of having his hair lose its beautiful gold.

What makes hurting so hard to describe? Is this because feelings are so simple and basic that there's nothing more to be said about them? No, it's precisely the opposite; chapter §9 will argue that feelings are intricate processes —but because we have so little sense of how these work, we can only describe their effects in terms of analogies with familiar things.

> *"I'm so something that I can't remember what it's called."*
> —Miles Steele (age 5)

For example I've heard suffering likened to a balloon that keeps dilating inside your mind until there's no more room for your usual thoughts. Then you might feel you've lost your 'freedom of choice' and that your mental condition has become like that of a prisoner.

In any case, this raises the question of what distinctions we're trying to make

with like *pain, discomfort,* and *suffering.* Sometimes these seem interchangeable, sometimes they signify different degrees, and at other times we use them as though we're referring to different phenomena. The next few sections will try to use different words for the kinds of mental activities that come shortly after an injury. We'll only use pain for what comes first—the sensations that come from the injury. Then we'll use *hurting* for what comes next—that is, for how we describe pain's early effects. Finally, we'll use *"suffering"* for the states we get when these escalate into large-scale cascades.

Critic: Even if your theory is right—that sufferings are disruption-cascades—why can't all that machinery work without making people feel so uncomfortable?

Our theory suggests that one cannot separate those things because when we speak about 'feeling uncomfortable' we *are* in large part referring to that disruption of our other thoughts! Indeed, pain could not serve the functions for which it evolved if our usual processes were to continue in the face of painful stimuli—for if we kept pursuing our usual goals we might not try to escape from those sources of pain, and just carry on with our usual thoughts while our bodies were being torn apart. [See §§Zombie-Machines.]

Philosopher: Isn't there still something missing here. You have been describing various mental conditions, and some machinery that might make them occur. But you have not given the slightest hint of why those conditions should give rise to feelings—or that basic sense of being or of experiencing.

Terms like 'basic' or 'experience' only hide our lack of insight about the processes they purport to describe. For example, when you 'see' your own hand, you seem to know that it is *your* hand without any intermediate steps—but that is because you have so little sense of the complex systems that recognize this.

It must be the same for feelings, too; when they seem basic or direct, this merely reflects our ignorance of how we recognize types of *mental* events.

What do we mean when we talk about feelings? What do we mean by *"I feel good," "I'm confused," "I'm excited,"* or *"Now I feel that I'm making progress."* You feel *pleased* when you achieve a goal—but this can be mixed with a sense of *regret* because now you must find something else to do. And sometimes success makes you feel *surprise*—which may lead you to ask what caused that success, or why you failed to *expect* it. Clearly, some such feeling must result from reflective attempts to describe your states.

For, when you ask yourself, *"How (or What) do I feel,"* this invites a description of your present condition—and of course such a question is hard to

answer because any such effort will have an effect on the system that's trying to make that description. Then this could make you (unknowingly) switch to using a different view of yourself—and this would make it hard for your mind to keep track of such changes in "real time."

This suggests that what we call 'feelings' are attempts (by various parts of our minds) to describe large-scale aspects of mental conditions. However, those conditions are usually so complex that the best we can do is to recognize them, and then try to say which other feelings they're 'like'. This is what make feelings hard to explain: it is *not* because a feeling is so basic that it's indescribable, but because each such conditions is so intricate that any compact description of it can capture no more than some fragments of it. This problem will come up many times in this book and Chapter §9 will try to summarize it.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§3-4. Overriding Pain

> *Sonja: "To love is to suffer. To avoid suffering one must not love. But then one suffers from not loving. Therefore, to love is to suffer; not to love is to suffer; to suffer is to suffer. To be happy is to love. To be happy, then, is to suffer, but suffering makes one unhappy. Therefore, to be happy one must love or love to suffer or suffer from too much happiness."*
> *—Woody Allen, in "Love and Death."*

Some of pain's effects are so quick that they're finished before you've had 'time to think'. If Joan had happened to touch something hot, she might have jerked her arm away before she even noticed it. But when that pain came from inside Joan's knee, her reflexes gave no escape from it, for it followed her everywhere she went and kept her from thinking of anything else. Persistent pain can distract us so much as to thwart all attempts to escape from it. Then we're trapped in a terrible circle. When pain gets too good at its principal job—of focusing you on your injury—you may need some way to override pain, to regain control of the rest of your mind.

If Joan urgently wants to cross that room, she can probably do it 'in spite of the pain'—at the risk of further injury—the way that runners and wrestlers do. Professional boxers and football players are trained to take blows that may damage their brains. Then, how do they override pain's effects?

> *"About that time, G. Gordon Liddy began a new exercise in will power. He would burn his left arm with cigarettes, then matches and candles to train himself to overcome pain. ... Years later, Liddy assured Sherry Stevens that he would never be forced to disclose anything he did not choose to reveal. He asked her to hold out a lit lighter. Liddy put his hand in the flame and held it there until the smell of burning flesh caused Stevens to pull the flame away."*
> *—Larry Taylor*

We each know tricks for doing this, and see some of these as commendable, and others as execrable, depending on the culture we're in.

Another way to deal with pain is to apply a *counter-irritant:* when a certain part of your body aches, it sometimes helps to rub or pinch that spot—or to aggravate some different place. But why should a second disturbance offset the

first, instead of making you feel worse?[36] And why do such drugs as the opiates have such specific effects on how much we hurt? Researchers have varied ideas about this but those theories are still incomplete. The simplest idea is when there are multiple disturbances, it is hard for the rest of the brain to choose one to 'focus' on—and (somehow) this makes it harder for a single large cascade to grow.

Usually when you attend to a pain, that makes the pain seem more intense—and this in turn intensifies your goal of getting rid of it.

If you keep your mind involved with other distracting activities, then a pain may seem to feel less intense. We all have heard those anecdotes about wounded soldiers who continue to fight without noticing pain—and only later succumb to shock, after the battle is lost or won. So the goal to survive, or to save one's friends, may be able to override everything else. On a smaller scale, with a mild pain, you can just be too busy to notice it. Then the pain may still 'be there' but no longer seems to bother you much. Similarly, you may not notice that you've become sleepy until you perceive that you're starting to yawn—and your friends may have noticed this long before. (In my own experience, the first awareness of being tired usually comes when I start to notice certain kinds of grammatical errors.) Shakespeare reminds us (in *King Lear*) that misery loves company: no matter how awful one's lot may be, we still may draw comfort from knowing that the same could happen to someone else.

> *When we our betters see bearing our woes,*
> *We scarcely think our miseries our foes.*
> *Who alone suffers suffers most i'th' mind,*
> *Leaving free things and happy shows behind;*
> *But then the mind much sufferance doth o'erskip*
> *When grief hath mates, and bearing fellowship.*
> *How light and portable my pain seems now,*
> *When that which makes me bend makes the King bow.*

Many other processes can alter how pain can affect our behavior:

Aaron Sloman: *"Some mental states involve dispositions, which in particular contexts would be manifested in behavior, and if the relevant behavior does not occur then an explanation is needed (as with a person who is in pain not wincing or showing the pain or taking steps to reduce it). The explanation may be that he has recently joined some stoic-based religious cult, or that he wants to impress his girl friend, etc."*

—In comp.ai.philosophy, 20/7/96.

This applies to the treatment of pain-ridden people.

"The degree of awareness of one's own pain may vary from a near denial of its presence to an almost total preoccupation with it, and the reasons for attending to pain may vary. Pain itself may become the focus of the self and self-identity, or may, however uncomfortable, be viewed as tangential to personhood. One of the most powerful influences on the way in which symptoms are perceived and the amount of attention paid to them is the meaning attributed to those symptoms." [37]

Finally, in Chapter §9, we'll discuss the seeming paradox implied by the common expression, *"No pain, no gain."* There are many common activities, such as in competitive sports, or training for strength, in which one tries to do things beyond one's reach—and where the greater the pain, then the higher the score.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

Prolonged and Chronic Suffering

When an injured joint becomes swollen and sore, and the slightest touch causes fiery pain, it's no accident that we say it's 'inflamed.' What could be the value of this, once the damage is already done? First, it can lead you to protect that site; thus helping that injury to heal; then it can make you feel sick and weak, both of which help to slow you down. So pain can promote recovery.

But it's hard to defend the dreadful effects of those chronic pains that never end. Then we tend to ask questions like, *"What did I do to deserve this?"* Then if we can find to justify punishment—it may bring us relief to be able to think, *"Now I can see why it serves me right!"*

Most victims discover no such escapes, and find that much has been lost from their lives—but some others find ways to see suffering as incentives or opportunities to show what they can accomplish, or even as unexpected gifts to help them to cleanse or renew their characters.

F. M. Lewis: *"Becoming an invalid can be a blow to a person's self-esteem. However, for some patients, the sick role is seen as an elevation in status—deserving the nurturance and concern of others. The ability to assign meaning to an illness or to symptoms has been found to enhance some patients' sense of self-mastery over a problem or crisis."*[38]

Thus certain victims find ways to adapt to chronic intractable pains. They work out new ways to make themselves think and rebuild their lives around those techniques. Hear Oscar Wilde describe how he deals with his inescapable misery:

"Morality does not help me. I am one of those who are made for exceptions, not for laws. Religion does not help me. The faith that others give to what is unseen, I give to what one can touch, and look at. Reason does not help me. It tells me that the laws under which I am convicted, and the system under which I have suffered are wrong and unjust. But, somehow, I have got to make both of these things just and right to me. I have got to make everything that has happened to me good for me. The plank bed, the loathsome food, the hard ropes, the harsh orders, the dreadful dress that makes sorrow grotesque to look at, the silence, the solitude, the shame—each and all of these things I had to transform into a spiritual experience. There is not a single degradation of the body which I must not try and make into a spiritualizing of the soul."[39]

Recent research on pain relief has developed new techniques, first for assessing degrees of pain and then for successfully treating it. We now have

drugs that can sometimes suppress some of pain's cruelest effects—but many still never find relief—either by mental or medical means. It seems fair to complain that, in this realm, evolution has not done well for us—and this frustrates theologians: *How to justify a world in which people are made to suffer so much?* What functions could such suffering serve? How did we come to evolve a design that protects our bodies but ruins our minds?

One answer is that the bad effects of chronic pain did not evolve from selection at all, but arose as a sort of 'programming bug.' Perhaps our ancestral ways to react to pain simply are not yet compatible with the reflective thoughts and farsighted plans that more recently evolved in our brains. The cascades that we call 'suffering' must have evolved from earlier schemes that helped us to limit our injuries—by making the goal of escaping from pain take such a high priority. The resulting disruption of other thought, was only was a small inconvenience before we developed our greater, modern intellects. Evolution never had any sense of what a species might evolve next—so it never prepared for intelligence.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

Grief

> *I cannot weep, for all my body's moisture*
> *Scarce serves to quench my furnace-burning heart;*
> *Nor can my tongue unload my heart's great burden,*
> *For self-same wind that I should speak withal*
> *Is kindling coals that fires all my breast,*
> *And burns me up with flames that tears would quench.*
> *To weep is to make less the depth of grief.*
> *Tears then for babes; blows and revenge for me!*
> *Richard, I bear thy name; I'll venge thy death,*
> *Or die renowned by attempting it.*
> *—Henry the Sixth, Part III*

When you suffer the loss of a long-time friend, it feels like losing a part of yourself, because grief involves our reactions to the loss of some of our mental resources. For, certain parts of your intellect must have over time become specialized for sharing ideas with the person you love; but now, the signals those brain-parts transmit will never again receive any replies—just as would happen with losing a limb. This could be why it takes so long to put to rest the loss of a friend.

> *Gloucester: Be patient, gentle Nell; forget this grief.*
> *Duchess: Ah, Gloucester, teach me to forget myself!*
> *—Henry the Sixth, part II*

Nell can't comply with Gloucester's advice because the links of affection are too broadly dispersed for any resource to erase all at once; they aren't all stored in some single place. Besides, we may not *want* to forget them all, as Aristotle remarks in *Rhetoric:*

> *"Indeed, it is always the first sign of love, that besides enjoying someone's presence, we remember him when he is gone, and feel pain as well as pleasure, because he is there no longer. Similarly there is an element of pleasure even in mourning and lamentation for the departed. There is grief, indeed, at his loss,*

but pleasure in remembering him and, as it were, seeing him before us in his deeds and in his life."

So Constance can say, in the play *King John,* that mournful feelings mix with pleasant memories:

Grief fills the room up of my absent child,
Lies in his bed, walks up and down with me,
Puts on his pretty looks, repeats his words,
Remembers me of all his gracious parts,
Stuffs out his vacant garments with his form;
Then have I reason to be fond of grief.

Thus Shakespeare shows how people clutch their griefs, and squeeze them till they change to joyful shapes.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

Today, there is a widely popular theory that, normally, recovery from a grievous loss or injury goes through a sequence of stages with names like *denial, anger, bargaining, depression,* and *acceptance.* I like the following skeptical and constructive analogy to this:[40]

As an example, apply the 5 stages to a traumatic event most all of us have experienced: The Dead Battery! You're going to be late to work so you rush out to your car, place the key in the ignition and turn it on. You hear nothing but a grind; the battery is dead.

Denial *--- What's the first thing you do? You try to start it again! And again. You may check to make sure the radio, heater, lights, etc. are off and then..., try again.*

Anger *--- "I should have junked this damned car a long time ago."*

Bargaining *--- (realizing that you're going to be late for work)... "Oh please car, if you will just start one more time I promise I'll buy you a brand new battery, get a tune up, new tires, belts and hoses, and keep you in perfect working condition.*

Depression *--- "Oh God, what am I going to do. I'm going to be late for work. I give up. My job is at risk and I don't really care any more. What's the use"?*

Acceptance *--- "Ok. It's dead. Guess I had better call the Auto Club or find another way to work. Time to get on with my day; I'll deal with this later."*

This relates to the general view of this book: although it is widely believed

that 'emotional' thinking is basically different from regular thought (and I don't insist they are quite the same), many of those supposed differences may disappear when we look more closely at commonsense things—as we shall in Chapter §6.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§3-5 Correctors, Suppressors, and Censors

> *"Don't pay any attention to the critics. Don't even ignore them."*
> —Sam Goldwyn

It would be wonderful never to make a mistake, nor ever to have a wrong idea. But perfection will always remain out of reach; we'll always makes errors and oversights.

Joan's sore knee has been getting worse. Today it hurts her all the time, even when it isn't touched. She thinks, "I shouldn't have turned while I lifted that box. And I should have put ice on my knee at once."

We like to think in positive terms: *"An* Expert *is someone who knows what to do."* And you know how to do most things so well that you scarcely need to think at all; you recognize most of the things you see, and converse without wondering how to speak. However, expertise also has an opposite side: *"An Expert is one who rarely fails—because of knowing what not to do."* Thus we usually do not walk into walls. We rarely stick things in our eyes. We never tell strangers how ugly they are.

How much of a person's competence is based on knowing which actions *not* to take—that is having ways to avoid mistakes? We don't know much about such "negative expertise" because this was rarely discussed in Psychology, except in the writings of Sigmund Freud.

Perhaps that neglect was inevitable because we cannot observe, from outside, the things that people do not do. But it is almost as hard to study such things by observing from *inside* the mind, for example, what keeps you from having absurd ideas. To account for this, we'll conjecture that our minds accumulate resources that we shall call *Critics*—each of which learns to recognize a certain particular kind of mistake. Here are a few of those types of Critics; we'll list more of them in Chapter §7.

A *Corrector Critic* warns you that you have started to do something dangerous. *"You must stop right now, because you're moving your hand toward a flame."* But such a warning may come too late.

A **Suppressor** can warn you of a danger you face, and can veto an action that's being considered, to stop you from acting before it's too late—for example, by telling you, *"No, do not move in that direction!* Or it could tell you to use a debugging technique.

A **Censor** works early enough to keep you from having that dangerous thought—so it never even occurs to you to put your finger into that flame. A Censor can work so effectively that you don't even know that it's working for you.

A **Self-Controller** recognizes that you have been failing to carry out a plan because, you instead of staying with it, you have kept on "changing your mind" about it.

Suppressors are safer than Correctors are, but both of them tend to slow you down, while you think of something else to do. However, Censors waste no time at all, because they deflect you from risky alternatives without interrupting your other thoughts, and therefore can actually speed you up. This could be one reason why some experts can do things so quickly: *they don't even think of the wrong things to do.*

Student: How could a censor ward off a bad thought—unless it already knows what you're likely to think? Isn't there some sort of paradox there?

AI Programmer: No problem. Just design each Censor to be a learning machine that records which decisions have led to mistakes. Then when it next sees a similar choice, it just steers your thoughts in the other direction, so that you won't make the same decision.

Student: Then wouldn't that Censor still take some time to have enough effect on your mind? Besides, what if both choices were equally bad? Then that Censor must work even earlier, to keep you from getting into that bad situation in the first place.

AI Programmer: We could do that by giving each Censor enough memory to record several of the previous steps that led to such situation.

Student: Might not that cure be worse than its disease? If your Correctors could save you from every mistake, this might make you so conservative that you'd scarcely ever get new ideas.

Indeed, some experts have learned so many ways for any project to go wrong that, now, they find it hard to explore any new ideas at all.

Excessive Switching

I have of late— but wherefore I know not— lost

all my mirth, forgone all custom of exercises; and indeed it goes so heavily with my disposition, that this goodly frame, the earth, seems to me a sterile promontory; this most excellent canopy, the air, look you, this brave o'erhanging firmament, this majestical roof fretted with golden fire, why, it appeareth nothing to me but a foul and pestilent congregation of vapors.
—Hamlet II.ii.292

What happens if too many **Critics** switch on (or off)? Here is a first-hand description of this:

Kay Redfield Jamison: "The clinical reality of manic-depressive illness is far more lethal and infinitely more complex than the current psychiatric nomenclature, bipolar disorder, would suggest. Cycles of fluctuating moods and energy levels serve as a background to constantly changing thoughts, behaviors, and feelings. The illness encompasses the extremes of human experience. Thinking can range from florid psychosis, or "madness," to patterns of unusually clear, fast and creative associations, to retardation so profound that no meaningful mental activity can occur. Behavior can be frenzied, expansive, bizarre, and seductive, or it can be seclusive, sluggish, and dangerously suicidal. Moods may swing erratically between euphoria and despair or irritability and desperation. ... [But] the highs associated with mania are generally only pleasant and productive during the earlier, milder stages."[41]

In a later paper, this author says more about such massive mental cascades:

It seems, then, that both the quantity and quality of thoughts build during hypomania. This speed increase may range from a very mild quickening to complete psychotic incoherence. It is not yet clear what causes this qualitative change in mental processing. Nevertheless, this altered cognitive state may well facilitate the formation of unique ideas and associations. ... Where depression questions, ruminates and hesitates, mania answers with vigor and certainty. The constant transitions in and out of constricted and then expansive thoughts, subdued and then violent responses, grim and then ebullient moods, withdrawn and then outgoing stances, cold and then fiery states—and the rapidity and fluidity of moves through such contrasting experiences—can be painful and confusing.[42]

It is easy to recognize such extremes in the mental illnesses called 'bipolar' disorders, but Chapter §7 will conjecture that we also use such processes in the course of everyday commonsense thinking. Thus, you might use a procedure like this whever you face a new problem:

First, shut most of your Critics off. This helps you to think of some things you could do—without concern about how well they might work—as though you were in a brief 'manic' state.

Then, you could turn many Critics on, to examine these options more skeptically—as though you were having a mild depression.

Finally, choose one approach that seems promising, and then proceed to pursue it, until one of your Critics starts to complain that you have stopped making progress.

Sometimes you may go through such phases deliberately. However, my conjecture is that we frequently do this on time-scales so brief that we have no sense that it's happening.

Learning from Failure

> *"Never interrupt your enemy when he is making a mistake."*
>
> Napoleon Bonaparte

Many things we regard as positive (such as beauty, humor, and pleasure itself) may be partly based on censorship—hence, to that extent, could be considered negative. Thus pleasure can seem 'positive' to the processes that now are presently "in control"—no matter that other processes (whose expressions are currently being suppressed) might otherwise see this as 'negative.' (See §9-2 of SoM.) For, *"I'm enjoying this"* could mean, both at once, *"I want to stay in my present state,"* and *"I want to prevent any changes in it."*

Student: But I thought that it was widely believed that learning works by 'reinforcing' connections that have led to success, and by weakening those that contribute to failure. Many educators say that we should always make it pleasant to learn, because pleasure is our reward for success—whereas failure deters and discourages us.

That popular view is mainly based on research (mostly done with pigeons and rats) that also shows that quicker rewards make learning more rapid. This

has many teachers toward the idea that learning should be a pleasant experience. However, we should not be too quick to apply this idea to beings like us, who also can learn by reflecting on the things they have done!

I'm not saying that 'reinforcement theory' is wrong—but that, for humans, it's just part of the story; in §8-5 I'll argue that what we can learn from how we have failed could be more important than 'reinforcement' can be—at least, for our highest levels of thinking.[43] For, while pleasure may help us learn easy things, section §9-4 will argue that we may need to endure some suffering to make larger-scale changes in how we think. If so, as an ancient Stoic might say, rewarding success can lead you to celebrate more than to investigate. Here are a few other reasons why to 'learn from success' is not always wise—especially when that success was expected.

__Reinforcement can lead to Rigidity.__ If a system already works, additional 'reinforcement' could make some its internal connections become stronger than they need to be, which could make it harder for that system to adapt to later new situations.

__Dependency leads to Side Effects.__ If a certain resource R has worked so well that other resources have come to depend on it, then any change you make in R will now be more likely to damage those others. In other words, as the saying goes "Don't fix it, unless it is broken."[44]

__Negative Expertise.__ One way to avoid such side effects is to leave an established resource unchanged, but to add Critics and Censors to intervene in conditions where it has failed to work. In other words, treat them as exceptions to rules.

__Radical Learning:__ You can "tune up" a skill by many small steps, but eventually no more small changes will help, because you have reached a local peak.[45] Then further improvement may require you to endure some discomfort and disappointment. See §9-4.

__Papert's Principle:__ When two or more of your methods conflict, then instead of seeking a compromise, abandon the lot and then try something else. Many steps in mental growth are less based on acquiring new skills, but more on learning better ways to choose which older ones to use. [See §10-4 of SoM.]

For all of those reasons, we need to learn, not only methods that worked in the past, but also which methods have failed—*and why*—so that one can avoid the most common mistakes.

Student: Yes, but why can't we do that by breaking connections—so that once you've made a bad mistake, your brain won't ever do it again?

One reason why this is a bad idea is that you'll lose the opportunity to understand just what went wrong (so that you can later avoid related mistakes). A second problem with this tactic is that whenever you change some of a system's connections, this may also affect some other behaviors that are partly based on those same connections. If you don't know quite how that system works, then you're in danger of making it worse by 'correcting' any remaining mistakes.

Programmer: I know exactly what you mean. Every attempt to improve a program is likely to introduce new bugs. That's why new programs so often contain very big sections of ancient code: no one remembers quite how they work, and hence they're afraid to change them.
Student: But what if you have no alternative, because something is wrong that you need to fix.

Perhaps our most important ways to improve ourselves come from *learning to think about thinking itself—that is, to 'reflect' on what our minds have been doing.* However, to do this one must first learn to enjoy the distress that results when one's forced to inspect oneself. See §8-5 and §9-4.

Varieties of Negative Expertise

Creativity: Why do some people get more good ideas? I did not specify 'new' ideas—because it is easy to build a machine that spouts endless streams of things that have never been seen; what distinguishes thinkers that we call 'creative' is not how many new things they produce, but how useful are the few they produce. This means that those artists have ways to suppress—or not even generate—products that have too much novelty, leaving only the ones that are just different enough to be useful.

Humor: Humor is also usually seen as positive but, really, jokes are basically negative—in the sense they almost always are about things that a person should not do, because they are prohibited, disgusting, or just plain stupid.[46]

Decisiveness: Similarly, we tend to think of decision-making as positive. But those moments in which we make a choice (and which we describe as an 'act of free will') may in fact be exactly the opposite; that moment in which 'you make your decision' may simply be the moment at which you turned off the complex processes that you use for comparing alternatives.

Pleasure: If we look at a mind as a playground in which many methods compete then the more pleasure we feel (in the Single-Self sense), the more negative may be its total effect on the rest of one's mental processes! For, what

actually happened may have been that some particular process seized control, and then turned off a lot of the rest of your mind. This, as every addict knows, makes it hard to wish for anything else. We'll say more about this in Chapter §9.

There are other ways to disable resources than attempting directly to suppress them. One way to suppress a resource is to activate one of its competitors. For example, you can hold off sleep by arranging to get into a fight. Another trick is to repeat a stimulus until your opponent no longer responds to it —as in the old tale of *"The Boy who Cried Wolf."*

Parenting: Consider how much a person must do in the course of raising a child. You must feed it and clean it and work to protect it—to guard it and clothe it and teach it and help it; for years, you must sacrifice wealth and attention. What kind of incentive could make one forego so many other enjoyments and goals, to become so selfless and other-directed? Such strong constraints, if imposed from outside, would seem cruel and unusual punishment. Clearly natural selection favored those who evolved ways to suppress those mental Critics; no person obsessed with those handicaps could bear to endure such prolonged distress—and would end up with fewer descendants.

Beauty: We tend to see Beauty as positive. But when someone says something is "beautiful" and you ask, *"What makes you attracted to that,"* your respondent may act as though under attack, or explain that *'there's no accounting for taste'*, or childishly say, *"I just like it."* Such answers suggest (as we saw in §1-1) that their liking comes partly from critic suppression. We all know that if one but tries, one can always uncover some blemish or flaw.

Mystical Experience: If you could turn most of your critics off, you then would have fewer concerns or goals. And if this occurs on a large enough scale, then your whole world may suddenly seem to change—and everything now seems glorious. If you'd like to experience this yourself, there are well-known steps that you can take to induce it.[47] It helps to be suffering pain and stress; starvation and cold will also assist. So will psychoactive drugs, and meditation too may aid. Be sure to stay in some strange, quiet place—because sensory deprivation helps. Next, set up a rhythmical drone that repeats some monotonous phrase or tone, and soon it will lose all meaning and sense—and so will virtually everything else! Then if you've done this successfully, you may suddenly find yourself overwhelmed by some immensely compelling Presence—and then you may spend the rest of your life trying and failing to find it again; I suspect that it masquerades records or traces of early imprimers that long have been hiding, disguised, in forgotten parts of your mind.

We have many kinds of words for this—Ecstasy, Rapture, Euphoria, Bliss— and Mystical Experience. You suddenly feel that you know the Truth, that

nothing else is significant, and that you need no further evidence; your mind has extinguished all its ways to question what was 'revealed' to you—and when later you try to explain to your friends, you find you can scarcely say anything else than how 'wonderful' that experience was. But if you failed to find any flaws because you had turned all your Critics off, then a better word would be *'wonderless.'*

<center>∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞</center>

§3-6 The Freudian Sandwich

Luck's a chance, but trouble's sure,
I'd face it as a wise man would,
and train for ill and not for good.
　　　—A. E. Housman

Few textbooks of psychology discuss how we choose what to think about—or how we choose what *not* to think about. However, this was a major concern to Sigmund Freud, who envisioned the mind as a system in which each idea must overcome barriers. Here is how he once envisioned a mind:

"... a large anteroom in which the various mental excitations are crowding upon one another, like individual beings. Adjoining this is a second, smaller apartment, a sort of reception room, in which consciousness resides. But on the threshold between the two there stands a personage with the office of doorkeeper, who examines the various mental excitations, censors them, and denies them admittance to the reception-room when he disapproves of them. You will see at once that it does not make much difference whether the doorkeeper turns any one impulse back at the threshold, or drives it out again once it has entered the reception-room. That is merely a matter of the degree of his vigilance and promptness in recognition."[48]

Thus getting past that doorkeeper is not quite enough to reach consciousness. That only leads to the reception room, which he sometimes calls the "preconscious."

"The excitations in the unconscious, in the antechamber, are not visible to consciousness, which is of course in the other room, so to begin with they remain

unconscious. When they have pressed forward to the threshold and been turned back by the doorkeeper, they are 'incapable of becoming conscious'; we call them then repressed. But even those excitations which are allowed over the threshold do not necessarily become conscious; they can only become so if they succeed in attracting the eye of consciousness."

Freud imagined the mind as obstacle course in which only ideas that get past all those bars are awarded the title of consciousness. In one kind of block that he calls "repudiation," an idea is deliberately condemned—and thus is rendered powerless—although one can remember rejecting it. In another type, which he calls "repression," an impulse is blocked at an earlier stage—without the thinker knowing this. However, repressed ideas can still persist, expressing themselves in clever disguises.

Inside Freud's three-part model of mind, many resources are working at once —but they don't always share the same purposes. Instead, that mind is a battleground between animal instincts and social constraints. These are frequently incompatible, so the rest of the mind must struggle to find acceptable ways to compromise—and that's often accomplished by subterfuge. One way to deal with a constraint is to suppress the resource that imposes it. Another is to disguise or re-describe it so that it arouses no **Censors** or **Critics.** Freud used the term *sublimation* for this; we sometimes call it 'rationalizing'.

Few modern 'cognitive psychologists' appreciate Freud's architectural concepts. He was one of the first to recognize that we deal with everyday problems in ways that are more complex than any one centralized process could. Instead, he saw the human mind as the product of diverse activities, many of which become engaged when we face conflicts and inconsistencies. Resolving these will often involve many different processes—to all of which we give vague names like Conscience, Emotion, and Consciousness.

Organization proposed by Sigmund Freud
Goals, ideals, values, prohibitions and
Sensory Systems   EGO = Conflict   Motor Systems
Innate, Instinctive Wishes and Drives

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§3-7. Controlling our Moods and Dispositions

> *"Love, he believed, made a fool of a man, and his present emotion was not folly but wisdom; wisdom sound, serene, well-directed. ... She seemed to him so felicitous a product of nature and circumstance that his invention, musing on future combinations, was constantly catching its breath with the fear of stumbling into some brutal compression or mutilation of her beautiful personal harmony ..."*
> *—Henry James, in The American.*

In §1-2 we described some ways that a person's state of mind might change:

"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 rest of the world seems dreary and dark, and your friends complain that you seem depressed."

If you could switch all your ***Critics*** off, then nothing would seem to have any faults. You'd be left with few worries, concerns, or goals—and others might describe you as elated, euphoric, demented or manic.

However, if you turned too many ***Critics*** on, you'd see imperfections everywhere. Your entire world would seem filled with flaws, engulfed in a flood of ugliness. If you also found fault with your goals themselves, you'd feel no urge to straighten things out, or to respond to any encouragement.

This means that our ***Critics*** must be controlled: If you turned too many on, then you'd never get anything done. But if you turned all your critics off, it might seem as though all your goals were achieved—and again you wouldn't accomplish much.

Nevertheless, in everyday life there remains a wide range in which it is safe to operate. Sometimes you feel adventurous, inclined to try new experiments. Other times you feel conservative—and try to avoid uncertainty. And when you're in an emergency (as when you face danger or aggression), you don't have time to reason things out, so you have to make quick decisions without considering most other factors. Then you'll have to postpone long-range plans, suspend some relationships with your friends, expose yourself to stress and pain, and make other choices you'll later regret. To do this, you'll have to suppress your suppressors—and then you may seem like a quite different person.

We use terms like 'disposition' and 'mood' to describe someone's overall state of mind. But terms like these are hard to define, because a person's present state involves so many processes. Some of these change the ways we perceive, while others affect which goals we'll select, which strategies we'll choose to use, and what degrees of detail we'll focus on. Yet other processes turn our thoughts from one mental realm to another, so that first one may think about physical things, then about some social concern, and then about some longer-term plan.

What determines the spans of time that our minds spend in each dispositional state? Those intervals span an enormous range. A flash of anger, or fear, or a sexual image may last for only a very brief moment. Other moods may last minutes or hours—and some dispositions persist for weeks or years. "John is angry" means that he's angry now—but "an angry kind of person" may describe a lifelong trait. The durations of such mental states could depend on how we regulate the rates at which we switch.

In §7-2 we'll speculate about how our *Critics* might be arranged. To what extent are they independent—like demons that constantly survey the scene, waiting for moments to intervene? To what extent are they controlled by special, more centralized managers? How do we learn new censors and critics? How many critics have critics themselves to scold them for poor performances? Are certain minds more productive because their critics are better organized?

Now it is more than a century since Sigmund Freud raised questions like these—but they have been so widely ignored that we still have don't have adequate answers to them. Perhaps this situation will change as we get better ways to see inside brains.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§3-8. Emotional Exploitation

Whatever you may be trying to do, your brain may have other plans for you.

I was trying to work on a technical theory, but was starting to fall asleep. Then I found myself imagining that my rival Professor Challenger was about to develop the same technique. This caused a flicker of angry frustration, which blocked for the moment that urge to sleep—and enabled me to proceed with my work.

In fact, Challenger was not doing any such thing; he works in a totally different field. But although he was a close friend of mine, we had recently had an argument. So he served as an opportune candidate when I needed someone to be angry at. Let's make up a theory of how this worked.[49]

A resource called Work was attending to one of my principal goals.
Another one called Sleep tried to seize control—but then that fantasy appeared.
This aroused a mixture of Anger, annoyance, frustration, and fear.
Somehow, these then had the effect of disrupting the process of falling asleep.

This sequence of steps established a state that counteracted the urge to sleep —and thus returned my mind to its 'working' state. We can see my use of that fantasy as having the effect of an emotional 'double negative': by using one system to switch off another.

Everyone uses such tricks to combat frustration, tedium, pain, or sleep. Here I used anger to keep myself working—but the same technique might serve as well, if one were falling behind in a race, or trying to lift too heavy a weight. By self-inducing anger or shame, you sometimes can counteract weakness or pain.

Note that 'Self-control' tactics need careful direction. Just a brief tweak might serve to stop *Sleep* —so slight that you don't know you're doing it. But if you don't sufficiently anger yourself, you might relapse into lassitude—whereas if you get yourself too incensed, you'll completely forget what you wanted to do.

Here's another example where part of a mind 'exploits' one emotion for the purpose to turning off another—thus helping you to attain some goal that you cannot achieve more directly.

Celia is trying to follow a diet. When she sees that thick, rich chocolate cake, she is filled with a strong temptation to eat. But when she imagines her friend, Miss Perfect-Body, looking gorgeous in her new bathing suit—then

Celia's passion to have that same shape keeps her from actually eating the cake.

What is the role of that fantasy? Celia's procedure for 'dieting' does not include any straightforward way to suppress her reckless appetite. However, the emotion that we call **Disgust** is already designed to do just that (by backing-up one's digestive tract) and, somehow, Celia has trained herself to react in that way when she thinks of her shape. When the sight of her rival arouses that image, she'll have less desire to eat that cake. But that strategy is not without risk: if Celia's jealousy makes her depressed, she might engorge the entire cake.

Why should fantasies have such effects, when we 'know' that they aren't real? Surely, this must be partly because each mind-part sees only a few other parts, which serve as its private reality. We never directly see the world; that's just another Single-Self myth. Instead, although some parts of your brain directly react to what your external senses provide, most of them must base their representations on information that they receive from other, internal brain-resources.

For example, when you sit at a table across from a friend and assume that she still has a back and some legs, you're using old models and memories. It's the same for the chair that she's sitting on. None of those things now lie in your sight, yet it's almost as though you can see them. Fantasy is the missing link. In {Imagination} and in {Simuli}, we'll see how machines could imagine such things.

Student: I know that we all have fantasies, but why did such strange ways of thinking evolve? Why can't we just figure out what to do in a perfectly rational way?

My answer is simply that there's no such thing; that popular concept of 'rational' is itself just one more fantasy—that our thinking is ever wholly based on pure, detached logical reasoning. It might seem somewhat 'irrational' to exploit an emotion to solve a problem. Our culture teaches us to believe that thoughts and emotions are separate things. But this makes no sense from the viewpoint of Work: when it can't control a resource that it needs, *this will appear from **Work's** point of view to be just an additional obstacle.* So far as your agents for **Work** are concerned, exploiting **Anger** to turn off **Sleep** is like using a stick to extend one's reach. No matter that when this is seen from outside, it appears to be "emotional": to Work this need not seem anything than another way to achieve its goal. We're always exploiting fantasies in the course of our everyday reasoning, and we all use such tricks for 'self-control'.

To stay awake, you can measure out the right amount of some stimulant. You can pinch yourself to produce some pain; or adopt an uncomfortable posture, or take a deep breath, or just set your jaw. You can move to a more exciting place, or indulge in a strenuous exercise. Or, you can make yourself angry or afraid—by imagining that you have failed.

A major part of our daily lives consists of these kinds of activities. It's customary to assume that it's 'you' who is choosing to do them. But often they come from small parts of your mind that are trying to change *their* environments. We need to imagine fictional things whenever we solve a geometry problem, or look forward to a forthcoming vacation. Whenever we think, we use fantasies to envision what we don't yet have, but might need. *To think about changing the way things are, we have to imagine how they might be.*

*Student: Again, I agree that we do such things—but again, I cannot help wondering why. Why cannot **Work** just turn off **Sleep**, but must use such indirect methods? Why do we have to tell lies to ourselves, by inventing illusions and fantasies—instead of simply commanding our minds to do whatever we want them to do? Why doesn't **Work** have better connections?*

One answer seems clear: Directness would be too dangerous. If Work could simply turn Hunger off, we'd all be in peril of starving to death. If Work could directly switch Anger on, we might find ourselves fighting most of the time. If Work could simply extinguish Sleep, we'd be likely to wear our bodies out. This is why it's distressing to hold your breath, and why it's so hard not to fall asleep —or to take control over how much you eat. Few animals that could do such things would live to have any descendants. Consequently, our brains evolved ways to keep our minds from meddling with the systems that work to keep us alive. Hence, we can interfere with those processes, only by becoming devious. We can't simply suppress the urge to sleep—but eventually, we discover some tricks that can do this by using indirect methods.

For example, here Work has no direct way to stop Sleep, but has learned that Anger undermines Sleep. And while Work has no direct way to activate Anger, it has learned that a certain fantasy can arouse Anger. So if Work can somehow activate that fantasy, then Anger will start to inhibit Sleep, and Work will be able to get back to work.

Student: Your theory suggests more questions than it answers. How could Work manage to learn such a trick?*How are those fantasies produced? How are those memories retrieved? How can a fantasy make you angry? How does Work induce that fantasy? How does Anger inhibit Sleep? And why do we need to*

sleep at all? Considering how much time it wastes, and all the inconvenience it brings, why did we ever evolve such a thing?

§5-8 Simuli will talk about how machines could make fantasies, *§6-2.2 Remembering* will consider how memories might be retrieved, and *§9-2.1 Self-Control* will discuss how Work might learn to use such a trick. As for why we need to sleep at all, it is strange how little we know about this. Recent research suggests that it plays important roles in how we learn, but clearly, sleep serves other purposes. It is common in evolution that whenever some new kind of function appears, other systems evolve new ways to exploit it. Thus once a first form of sleep evolved, other functions were found for it—perhaps for renewing depleted resources, for repairing damage to organs, or, perhaps for imagining things without exposure to external risk. So, we should not expect to find one reason for all the many aspects of sleep—or for any other mental function.

Student: How does Anger inhibit Sleep in the first place?

That must involve ancient machinery. We're born with great systems of built-in connections that help us recognize dangers, failures and other sorts of emergencies. These 'alarms' have connections to other resources, such as the "Emotion-Arousers" of §1-6, which can drive into those great cascades—like anger, anxiety, fear, or pain—that can reset all our priorities. [See *§§Alarms.*]

Student: You haven't discussed how Anger works.

One theory could be that the state we call *'Anger'* suppresses some of our more thoughtful resources—so that we become less 'reasonable'. Then we tend to make more quick decisions, and thus are disposed to take more risks. It is tempting to think of such a person as erratic and unpredictable. Yet paradoxically such persons become, in certain ways, *more* predictable than they'd normally be—and that can have a useful effect: when you are angry and express a threat, your opponent may sense that you won't change your mind—because you are no longer 'reasonable.' The effectiveness of apparent threats depends on convincing antagonists that one truly intends to carry them out. If you can make yourself think that your threat is real, this can help you to display the emotional signs that will make your opponent believe it, too!

Critic: Not all types of anger cause rapid decisions. When Charles flies into a sudden rage, and punches someone who taunted him, his decision is quick— and he takes a big risk. But when Joan is chronically angry about the destruction of rainforest habitats, she may become deliberate and methodical at raising funds for saving them.

Our adult emotions continue to grow into ever more convoluted arrangements. As we age, we can train our emotional states—and modify their outward signs—till they no longer resemble their infantile shapes.

Physiologist: Anger is not just a state of mind; it also raises your muscle tone, fires you up with energy, and speeds up your reaction time. This involves the body and not just the brain.

Certainly, *Anger* engages many bodily functions; it can affect your heart rate, blood pressure, breathing, and sweating. However, when seen in the Cloud-of-Resources view, there is nothing special about such connections; the body itself then appears as just one more set of resources to exploit. (And quite a few of those same effects will occur if you simply hold your breath.) For, it is easy to see why such systems evolved: anger helps us to prepare for certain and emergencies—such as fighting, defense, and intimidation. However, we should not too closely identify these with how Anger changes one's *Ways to Think*; it is true that these interact with those somatic effects, but yet are far from being the same sorts of things. [See *§§Embodiment.*]

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

TRANSITION?

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

Part IV. Consciousness

§4-1. What is the nature of Consciousness?

> *"No philosopher and hardly any novelist has ever managed to explain what that weird stuff, human consciousness, is really made of. Body, external objects, darty memories, warm fantasies, other minds, guilt, fear, hesitation, lies, glees, doles, breath-taking pains, a thousand things which words can only fumble at, co-exist, many fused together in a single unit of consciousness"*
>
> *—Iris Murdoch, in The Black Prince. 1973.*

What kinds of creatures have consciousness? Does it exist in chimpanzees— or in gorillas, baboons, or orangutans? What about dolphins or elephants? Are frogs, fish, insects, or vegetables aware of themselves to any extent—or is consciousness a singular trait that segregates us from the rest of the beasts?

Although those animals won't answer questions like, *"Are you aware that you exist,"* or *" What is your view of what consciousness is,"* the answers from people are scarcely more useful. When you ask mystical thinkers how consciousness *works,* their replies are not highly enlightening.

Sri Chinmoy: "Consciousness is the inner spark or inner link in us, the golden link within us that connects our highest and most illumined part with our lowest and most unillumined part."[50]

Some philosophers even insist that there's no way to look for good answers to this.

Jerry Fodor: "Nobody has the slightest idea how anything material could be conscious. Nobody even knows what it would be like to have the slightest idea about how anything material could be conscious. So much for the philosophy of consciousness."[51]

Is consciousness an 'all-or-none' trait that has a clear-cut boundary, or does it have different amounts and qualities—the way that a thing can be cold or hot?

Relativist: Everything has some consciousness. An atom has only a little of it.

Bigger things must have *it in larger degrees— right up to the stars and the galaxies.*

Absolutist: We don't know where consciousness starts and stops, but clearly each thing must be conscious or not—and, clearly, there is no such thing in a rock.

Computer User: Certain programs seem to me already conscious to some small degree.

Logicist: Before you go on about consciousness, you really ought to define it. Good arguments should start right out by stating precisely what they are about. Otherwise, you'll build on a shaky foundation.

That policy might seem 'logical'—but it's wrong when it comes to psychology, because it assumes that 'consciousness' has a clear and definite meaning. Of course, we don't like to be imprecise—but *strict definitions can make things worse,* until we're sure that our ideas are right. For, 'consciousness' is a word we use for many types of processes, and for different kinds of purposes; we apply it to feelings, emotions, and thoughts—and to how we think and feel about them. It's the same for most everyday words about minds, such as 'creativity' or 'intelligence'.

So instead of asking what 'consciousness' *is*, or what we mean by *'being aware,'* we'll try to examine when and why people use those mysterious words. But why do such questions even arise? What, for that matter, are *mysteries?*

Daniel Dennett: "A mystery is a phenomenon that people don't know how to think about—yet. Human consciousness is just about the last surviving mystery. There have been other great mysteries [like those] of the origin of the universe and of time, space, and gravity. ... However, Consciousness stands alone today as a topic that often leaves even the most sophisticated thinkers tongue-tied and confused. And, as with all of the earlier mysteries, there are many who insist— and hope—that there will never be a demystification of consciousness."

—Consciousness Explained, 1991

Indeed, many of those who 'insist—and hope' that consciousness cannot be explained still claim that it alone is the source of most of the virtues of human minds.

Thinker 1: Consciousness is what unifies our present, past, and future together, by making sense of all our experience.

Thinker 2: Consciousness makes us 'aware' of ourselves, and gives us our sense of identity; it is what animates our minds and gives us our sense of being alive.

Thinker 3: Consciousness is what gives things meanings to us; without it, we would not even know we had feelings.

Wow! How could one principle, power, or force endow us with so many faculties? It can't—and this chapter will argue that there is no reason to suppose that all of those different abilities stem from just one common origin. Indeed, from what we know about brains, it is safer to guess that they're each based on different machinery.

William Calvin and George Ojeman "Modern discussions of consciousness ... usually include such aspects of mental life as focusing your attention, things that you didn't know you knew, mental rehearsal, imagery, thinking, decision making, awareness, altered states of consciousness, voluntary actions, subliminal priming, the development of the concept of self in children, and the narratives we tell ourselves when awake or dreaming."[52]

All this shows that "consciousness" does not refer to any single idea or thing, but that we use it as a suitcase-word for a great many different activities.

<center>∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞</center>

§4-2. Unpacking the Suitcase of Consciousness

> *Aaron Sloman: "It is not worth asking how to define consciousness, how to explain it, how it evolved, what its function is, etc., because there's no one thing for which all the answers would be the same. Instead, we have many sub-capabilities, for which the answers are different: e.g. different kinds of perception, learning, knowledge, attention control, self-monitoring, self-control, etc."[53]*

To see the variety of what human minds do, consider this fragment of everyday thinking.

Joan is part way across the street on the way to deliver her finished report. While thinking about what to say at the meeting, she hears a sound and turns her head —and sees a quickly oncoming car. Uncertain whether to cross or retreat, but uneasy about arriving late, she decides to sprint across the road. She later remembers her injured knee and reflects upon

her impulsive decision. *"If my knee had failed, I could have been killed. Then what would my friends have thought of me?"*

It might seem natural to ask, *"How conscious was Joan of what she did?"* But rather than dwell on that *'consciousness'* word, let's look at a few of the things that Joan *"did."*

Reaction: Joan reacted quickly to that sound.
Identification: She recognized it as being a sound.
Characterization: She classified it as the sound of a car.
Attention: She noticed certain things rather than others.
Imagining: She envisioned two or more possible futures.
Indecision: She wondered whether to cross or retreat.
Decision: She chose one of several alternative actions.
Recollection: She retrieved descriptions of prior events.
Reconsideration: Later she reconsidered this choice.
Selection: She selected a way to choose among options.
Apprehension: She was uneasy about arriving late.
Planning: She constructed a multi-step action-plan.
Embodiment: She tried to describe her body's condition.
Emotion: She changed major parts of her mental state.
Representation: She interconnected a set of descriptions.
Language: She constructed several verbal expressions.
Narration: She heard them as dialogs in her mind.
Anticipation: She expected certain future condition.
Intention: She changed some of her goals' priorities.
Reasoning: She made various kinds of inferences.
Reflection: She thought about what she's recently done.
Self-Reflection: She reflected on her recent thoughts.
Empathy: She imagined other persons' thoughts.
Moral Reflection: She evaluated what she has done.
Self-Imaging: She made and used models of herself.
Self-Awareness: She characterized her mental condition.
Sense of Identity: She regarded herself as an entity.

That's only the start of a much longer list of aspects of how we feel and think —and if we want to understand how our minds work, we'll need explanations for all of them. To do this, we'll have to take each one apart, to account for the details of how they each work. Then each reader can decide which ones should, or should not be regarded as aspects of 'consciousness.'

4-2.1. Suitcase words in Psychology

Holist: Yet after you analyze all those parts, you will still be obliged to explain how they all unite to produce the streams of consciousness that emerge from them. So, then you still will need some words to describe that entire phenomenon.

Why did our language come to include such terms as 'awareness,' 'perception,' 'consciousness,' every one of which condenses many different processes?

Psychologist: Such self-words are useful in everyday social life because they help us to communicate—both with our friends and with ourselves. For, because we all share the same kinds of jumbled ideas, we can pack them into vague suitcase-terms that seem easy for us to understand.

Ethicist: We need them also to support our principle of responsibility and discipline. Our legal and ethical principles are largely based on the idea that we ought to punish or reward only actions that are 'intentional'—that is, are based on having been planned in advance, with predictions about their consequences.

Psychiatrist: Perhaps we use those suitcase terms to keep ourselves from asking too much about how our minds control themselves, and what underlies the decisions we make. Perhaps we use words like "consciousness" to help us suppress all those questions all at once—by suggesting that all of them are just a single big Mystery.

Student: If "consciousness" is just a suitcase word, what makes it seem so clear to us that we actually possess such a thing? If such terms keep shifting their meanings, why doesn't this become evident whenever we start to think about them?

That could be because no part of a mind can 'see' much of what the rest of that mind does. A typical resource inside a brain accomplishes its jobs internally, in ways that other resources cannot perceive. Also, when any resource probes into another, that very act may change the other's state—and thus scramble the very evidence it would need to recognize what's happening. These could partly account for Hume's complaint that our minds lack good ways to inspect themselves.

David Hume: "The motion of our body follows upon the command of our will. Of this we are every moment conscious. But the means, by which this is

effected; the energy, by which the will performs so extraordinary an operation; of this we are so far from being immediately conscious, that it must for ever escape our most diligent enquiry."[54]

Hume assumes that we could never develop more powerful ways to inspect ourselves—but today we have new image-machines that show more of what happens inside our brains. For example, now we can detect activities that start before our limbs begin to move.

Dualist philosopher: Still, those instruments will eventually fail, because you can measure a brain but not an idea. Some creatures are conscious, while others are not—and consciousness is a subjective thing that can't be explained in physical terms.

Functionalist Philosopher: What evidence could support your faith that consciousness could never be explained? We can see it simply as our name for what happens when certain processes run in our brains.

I would agree with that second opinion, except that we also need to say more about what those 'certain processes' *do*—and why we distinguish them as a group. (The next section will offer a theory of this.) Still, many thinkers still maintain that brains must be based on something beyond the reach of our present-day machines.

Emergentist: Perhaps consciousness is just one of those 'wholes' that emerge when systems get complex enough. Perhaps that's just what we should expect from the network of billions of cells in a brain.

When we increase a system's size, then it will usually work less well, unless we also improve its design, and that always involves some compromise; if a system is built with too many connections, this will lead to traffic jams—while if the connections between its parts are too sparse, it is unlikely to anything useful at all.

Besides, if mere complexity were all it needs, then almost everything would have consciousness. We don't want to conclude that water-waves think—yet the manner in which a wave breaks on a beach is more complex (at least in some ways) than the processes that go on in our brains.

So, there's no point to asking what consciousness 'is'—because we've seen that this is a suitcase word, which we each fill up with far more stuff than could possibly have just one common cause. It makes no sense to try to discuss so many different things at once—except when trying to explain why we tend to treat all those things as the same. Let's listen to Aaron Sloman again:

Aaron Sloman: "I for one, do not think defining consciousness is important at all, and I believe that it diverts attention from important and difficult problems. The whole idea is based on a fundamental misconception that just because there is a noun "consciousness" there is some 'thing' like magnetism or electricity or pressure or temperature, and that it's worth looking for correlates of that thing. Or on the misconception that it is worth trying to prove that certain mechanisms can or cannot produce 'it', or trying to find out how 'it' evolved, or trying to find out which animals have 'it', or trying to decide at which moment 'it' starts when a fetus develops, or at which moment 'it' stops when brain death occurs, etc. There will not be one thing to be correlated but a very large collection of very different things."[55]

I completely agree with Sloman's view. To understand how our thinking works, we must study those *"very different things"* and then ask what kinds of machinery could accomplish some or all of them. *In other words, we must try to* **design**—*as opposed to* **define**—*machines that can do what our minds can do.*

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞
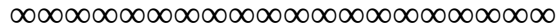
§4-3. How do we recognize Consciousness?

Student: You still did not answer my question on why, if "consciousness" is just a suitcase word, what makes it seem like such a definite thing.

Here is a theory of why that could happen: Most of our mental activities run more or less 'unconsciously'—in the sense that we're barely aware of them. But when we encounter obstacles, this starts up some high-level processes that have some properties like these:

(1) They make use of our most recent memories.
(2) They operate more serially, than in parallel.
(3) They use abstract, symbolic, or verbal descriptions
(4) They use models that we have made of ourselves.

Now suppose that a brain could construct a resource called **C** that detects when all these are running at once:
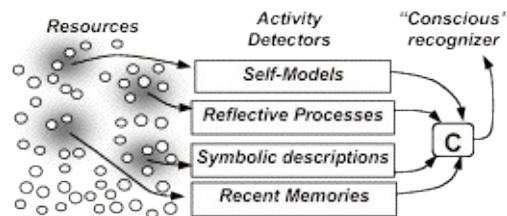
If such a **C**-detector turned out to be useful enough, this could lead us to imagine that it detects the presence of some sort of 'Consciousness-Thing!' Indeed, we might even imagine that entity to be the *cause* of that set of activities, and our language systems might learn to connect this kind of detector to terms like 'awareness,' 'myself,' 'attention,' or 'Me'. To see how this might be useful to us, let's examine its four constituents.

Recent Memories: *Why must consciousness involve memory? I've always thought of consciousness as about the present, not the past—about what's happening right now.*

For any mind (or any machine) to know what it has done, it needs some records of recent activities. For example, suppose that I asked, *"Are you aware that you're touching your ear?"* Then you might reply, *"Yes, I'm aware that I am doing that."* However, for you to make a statement like that, your language resources must be reacting to signals from other parts of your brain, which in turn have reacted to prior events. So, whatever you say (or think) about yourself, it takes time to collect that evidence.

More generally, this means that a brain cannot think about what it is thinking *right now;* the best it could do is to contemplate some records of some of its recent activities. There is no reason why some part of a brain could not think about what it has seen of the activities of other parts—but even then, there always will be at least so small delay in between.

Serial Processes. *Why should our high-level processes tend to be more serial? Would it not be more efficient for us to do more things in parallel?*

Most of the time in your everyday life, you do many things simultaneously; you have no trouble, all at once, to walk, talk, see, and scratch your ear. But few can do a passable job at drawing a circle and square at once by using both of their hands.

Citizen: Perhaps each of those two particular tasks demands so much of your attention that you can't concentrate on the other one.

That would make sense if you assume that *attention* is some sort of thing that comes in limited quantities—but then we would need a theory about what might impose this kind of limitation, yet still can walk, talk and see all at once. One explanation of this could be that such limits appear when resources conflict. For, *suppose that two tasks are so similar that they both need to use the same mental resources.* Then if we try to do both jobs at once, one of them will be forced to stop—and the more such conflicts arise in our brains, the fewer such jobs we can do simultaneously.

Then why can we see, walk, and talk all at once? This presumably happens because our brains contain substantially separate systems for these—located in different parts of the brain—so that their resources don't conflict so often. However, when we have to solve a problem that's highly complex then we usually have only one recourse: somehow to break it up into several parts—each of which may require some high-level planning and thinking. For example each of those subgoals might require us to develop one or more little 'theories' about the situation—and then do some mental experiments to see if these are plausible.

Why can't we do all this simultaneously? One reason for this could simply be that our resources for making and using plans has only evolved rather recently—that is, in only a few million years—and so, we do not yet have multiple copies of them. In other words, we don't yet much capacity at our highest levels of 'management'—for example, resources for keeping track of what's left to be done and for finding ways to achieve those goals without causing too many internal conflicts. Also, our processes for doing such things are likely to use the kinds of symbolic descriptions discussed below—and those

resources are limited too. If so, then our only option will be to focus on each of those goals sequentially.[56]

This sort of mutual exclusiveness could be a principle reason why we sometimes describe our thoughts as flowing in a 'stream of consciousness'—or as taking the form of an 'inner monologue'—a process in which a sequence of thoughts seems to resemble a story or narrative.[57] When our resources are limited, we may have no alternative to the rather slow 'serial processing' that so frequently is a prominent feature of what we call "high-level thinking."[58]

Symbolic Descriptions: *Why would we need to use symbols or words rather than, say, direct connections between cells in the brain?*

Many researchers have developed schemes for learning from experience, by making and changing connections between various parts of systems called *'neural networks' or 'connectionist learning machines.'*[59] Such systems have proved to be able for learning to recognize various kinds of patterns—and it seems quite likely that such low-level processes could underlie most of the functions inside our brains.[60] However, although such systems are very useful at doing many useful kinds of jobs, they cannot fulfill the needs of more reflective tasks, because they store information in the form numerical values that are hard for other resources to use. One can try to interpret these numbers as correlations or likelihoods, but they carry no other clues about what those links might otherwise signify. In other words, such representations don't have much expressiveness. For example, a small such neural network might look like this.
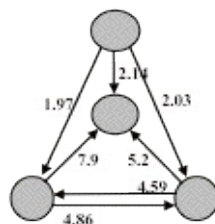
In contrast, the diagram below shows what we call a "Semantic Network" that represents some of the relationships between the parts of a three-block Arch. For example, each link that points to the concept *supports* could be used to predict that the top block would fall if we removed a block that supports it.
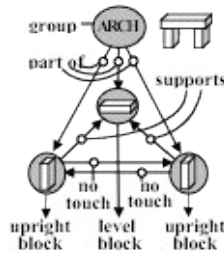
Thus, whereas a *'connectionist network'* shows only the 'strength' of each of those relations, and says nothing about those relations themselves, the three-way links of Semantic Networks can be used for many kinds of reasoning.

Self-Models: *Why did you include 'Self-Models' among the processes in your first diagram?*

When Joan was thinking about what she had done, she asked herself, *"What would my friends have thought of me."* But the only way she could answer such questions would be to use some descriptions or models that represent her friends and herself. Some of Joan's models of herself will be descriptions of her physical body, others will represent some of her goals, and yet others depict her dispositions in various social and physical contexts. Eventually we build additional structures include collections of stories about our own pasts, ways to describe our mental states, bodies of knowledge about our capacities, and depiction of our acquaintances. Chapter §9 will further discuss how we make and use 'models' of ourselves.

Once Joan possesses a set of such models, she can use them to think self-reflectively—and she'll feel that she's thinking about herself. If those reflections lead to some choices she makes, then Joan may feel that she is in "control of herself"—and perhaps apply the term 'conscious' to this. As for her other processes, if she suspects that they exist at all, she may represent them as beyond her control and call them 'unconscious' or 'unintentional.' And once we provide machines with such structures, perhaps they, too, will learn to make statements like, *"I feel sure that you know just what I mean when I speak about 'mental experiences.'*

I don't mean to insist that 'detectors' like these must be involved in all of the processes that we call consciousness. However, without some ways to recognize these particular patterns of mental conditions, we might not be able to talk about them!

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

This section began with some ideas about what we recognize when we talk about consciousness, and we suggested that this might relate to detecting some set of high-level activities.

However, we also ought to ask what might cause us to *start up* such sets of activities. This could be done in the opposite way: suppose among Joan's resources are some 'Trouble-Detectors' or 'Critics' that detect when her thinking has got into trouble—for example, when she fails to achieve some important goal, or to overcome some obstacle. In such a condition, Joan might describe her state in terms of distress or frustration, and try to remedy this by a mental act that, expressed in words, might be *"Now I should make myself concentrate."* Then she could try to switch to some way to think that engages more high-level processes—for example, by activating set of resources like these:
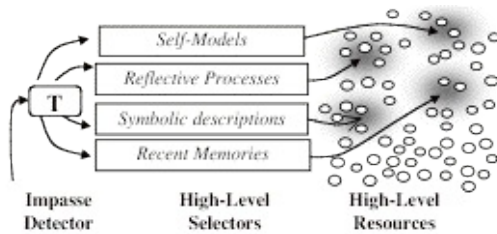
This suggests that we sometimes use 'conscious' to refer to activities that *initiate* rather than *recognize* sets of higher-level processes.

Student: How did you choose those particular features for your scheme to decide when to use words like 'consciousness?' Surely, since this is a suitcase-word, each person might make a different such list.

Indeed, just as we have multiple meanings for most of our other psychology-words, we're likely to switch among different such feature-lists whenever we use words like 'consciousness.'

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

4.3.1 The Immanence Illusion.

The paradox of consciousness—that the more consciousness one has, the more layers of processing divide one from the world—is, like so much else in nature, a trade-off. Progressive distancing from the external world is simply the price that is paid for knowing anything about the world at all. The deeper and broader [our] consciousness of the world becomes, the more complex the layers of processing necessary to obtain that consciousness.

—Derek Bickerton, Language and Species, 1990

When you enter a room you have the sense that you instantly see all the things in your view. However, this is an illusion because it will take time to recognize the objects that are actually there; then you'll have to revise many wrong first impressions. Nevertheless, all this proceeds so quickly and smoothly that this requires an explanation—and we'll propose one later in §8-3 Panalogy.

The same thing happens inside one's mind. We usually have a constant sense that we're 'conscious' of things that are happening *now*. But when we examine this critically, we recognize that there must be something wrong with it—because nothing exceeds the speed of light. This means that no internal part of a brain can ever know exactly what is happening "now"—either in the outside world or in any other part of that brain. The most that any resource can know is some of what happened in the recent past.

Citizen: Then why does it seem to me that I am conscious of all sorts of sights and sounds, and of feeling my body moving around—right at this very moment of time? Why do all those perceptions seem to come to me instantaneously?

It makes good sense, in everyday life, to assume that everything we see is "present" in the here and now, and it normally does no harm to suppose that we are in constant contact with the outside world. However, I'll argue that this illusion results from the marvelous ways that our mental resources are organized —and I think this phenomenon needs a name:

The Immanence Illusion: *For most of the questions you would otherwise ask, some answers will have already arrived before the higher levels of your mind have had enough time to ask for them.*

In other words, if some data you need were already retrieved before you recognized that you needed it, you will get the impression of knowing it instantaneously—as though no other processes intervened.[61]

For example, before you enter a familiar room, it is likely that you have already retrieved an old description of it, and it may be quite some time before you notice that some things have been changed; the idea that one exists in *the present moment* may be indispensable in everyday life—but much what we think that we see are the stereotypes of what we expected.

Some claim that it would be wonderful to be constantly aware of all that is happening. But the more often your high-level mental resources change their views of reality, the harder it will be for them to find significance in what they sense. The power of our high-level descriptions comes not from changing ceaselessly, but from having enough stability.

In other words, for us to sense what persists through time, one must be able to examine and compare descriptions from the recent past. We notice change in spite of change, not because of it. Our sense of constant contact with the world is a form of the Immanence Illusion: it comes when every question asked about something is answered before we know it was asked—as though those answers were already there.[62]

In Chapter §6 we'll also see how *our ways to activate knowledge before we need it could explain* why our *'commonsense knowledge' seems 'obvious'*.

<p style="text-align:center">∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞</p>

§4-4. Over-rating Consciousness

> *"Our mind is so fortunately equipped that it brings us the most important bases for our thoughts without our having the least knowledge of this work of elaboration. Only the results of it become conscious. This unconscious mind is for us like an unknown being who creates and produces for us, and finally throws the ripe fruits in our lap."*
> —*Wilhelm Wundt (1832-1920)*

Why has Consciousness' seemed such a mystery? I'll argue that this is largely because we exaggerate our perceptiveness. For example, at any one moment the lens of your eye can clearly focus only on objects in a limited

distance range, while everything else will be blurry.

Citizen: That doesn't seem to apply to me, because all the objects that I can see seem clearly focused all at once.

You can see that this is an illusion, if you focus your eyes on your fingertip while trying to read a distant sign. Then you'll see *a pair of those signs at once, but both will be too blurry to read.* Until we do such experiments, we think we see everything clearly at once, because the lens in each eye so quickly adjusts that we have no sense that it's doing this. Similarly, most people believe they see, at once all the colors of things in a scene—yet a simple experiment will show that we only see colors of things in the field near the object you're looking at.

Both of these are instances of that Immanence Illusion, because your eyes so quickly turn to see whatever attracts your attention. And I claim that the same applies to consciousness; we make almost the same kinds of mistakes about how much we can 'see' inside our own minds.

Patrick Hayes: *"Imagine what it would be like to be conscious of the processes by which we generate imagined (or real) speech. ... [Then] a simple act like 'thinking of a name', say, would become a complex and skilled deployment of elaborate machinery of lexical access, like playing an internal filing-organ. The words and phrases that just come to us to serve our communicative purposes would be distant goals, requiring knowledge and skill to achieve, like an orchestra playing a symphony or a mechanic attending to an elaborate mechanism.*[63]

Hayes goes on to say that if we were aware of all this, then:

"We would all be cast in the roles of something like servants of our former selves, running around inside our own heads attending to the details of the mental machinery which currently is so conveniently hidden from our view, leaving us time to attend to more important matters. Why be in the engine room if we can be on the bridge?"

In this paradoxical view, consciousness still seems marvelous—but not because it tells us so much, but for protecting us from such tedious stuff! Here is another description of this, from section 6.1 of *The Society of Mind.*

Consider how a driver guides the immense momentum of a car, not knowing how its engine works or how its steering wheel turns it left or right. Yet when one comes to think of it, we drive our bodies, cars, and minds in very similar ways. So far as conscious thought is concerned, you steer yourself in much the

same way; you merely choose your new direction, and all the rest takes care of itself. This incredible process involves a huge society of muscles, bones, and joints, all controlled by hundreds of interacting programs that even specialists don't yet understand. Yet all you think is "Turn that way," and your wish is automatically fulfilled.

And when you come to think about it, it scarcely could be otherwise! What would happen if we were forced to perceive the trillions of circuits in our brains? Scientists have peered at these for a hundred years—yet still know little of how they work. Fortunately, in everyday life, we only need to know what they achieve! Consider that you can scarcely see a hammer except as something to hit things with, or see a ball except as a thing to throw and catch. Why do we see things, less as they are, and more with a view of how they are used?

Similarly, whenever you play a computer game, you control what happens inside the computer mainly by using symbols and names. The processes we call "consciousness" do very much the same. It's as though the higher levels of our minds sit at mental terminals, steering great engines in our brains, not by knowing how that machinery works, but by 'clicking' on symbols from menu-lists that appear on our mental screen-displays.

Our minds did not evolve to serve as instruments for observing themselves, but for solving such practical problems as nutrition, defense, and reproduction.

<p style="text-align:center">∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞</p>

§4-5. Self-Models and Self-Consciousness

> *In judging the development of self-consciousness, we must guard against accepting any single symptoms, such as the child's discrimination of the parts of his body from objects of his environment, his use of the word "I," or even the recognition of his own image in the mirror. ... The use of the personal pronoun is due to the child's imitation of the examples of those about him. This imitation comes at very different times in the cases of different children, even when their intellectual development in other respects is the same.*
>
> *—Wilhelm Wundt, 1897.*[64]

In §4-2 we suggested that Joan *'made and used models of herself'*—but we did not explain what we meant by a *model.* We use that word in quite a few ways, as in *"Charles is a model administrator,"* which means that he is an example worthy to imitate—or in, *"I'm building a model airplane,"* which means something built on a scale smaller than that of the original. But here we're using *'model of X'* to mean a simplified mental representation that can help us to answer some questions about some other, more complex thing *X*.

Thus, when we say that 'Joan has a *mental model of Charles',* we mean that Joan possesses some mental resource *that helps her answer some questions about Charles.*[65] I emphasize the word *some* because each of Joan's models will only work well on some kinds of questions—and might give wrong answers to most other questions. Clearly the quality of Joan's thought will depend both on how good her models are, but on how good are her ways to choose which model to use in each situation.

Some of Joan's models will have practical uses for predicting how physical actions will make things change in the outer world. She will also have models for predicting how mental acts will make changes in her mental state. In Chapter §9 we'll talk about some models that she can use to describe herself—that is, to answer some questions about her own abilities and dispositions; these could some descriptions of

Her various goals and ambitions.
Her professional and political views.
Her ideas about her competences,
Her ideas about her social roles.
Her various moral and ethical views.
Her beliefs about of what sort of thing she is.

For example, she could try to use some of these to guess whether she can rely on herself to actually carry out a certain plan. Furthermore, this could explain some of her ideas about consciousness. To illustrate this, I'll use an example proposed by the philosopher Drew McDermott.[66]

Joan is in a certain room. She has a mental model of some of the contents in that room. One of those objects is Joan herself
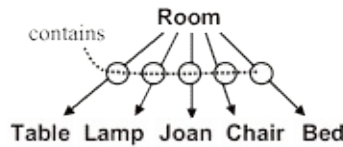
Most of those objects will have sub-models themselves, for example to describe their structures and functions. Joan's model for that object "Joan" will be a structure that she calls *"Myself,"* and which includes at least two parts—one called *Body* and one called *Mind.*
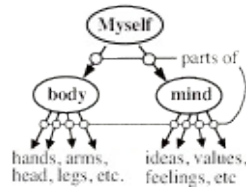
By using the various parts of this model, Joan could say *'Yes'* if you asked her, *"Do you have a mind?"* But if you asked her, *"Where is your mind?"* this model would not help her to say, as some people would, *"My mind is inside my head (or my brain)."* However, Joan could offer such a reply if *Myself* included a *part-of* link from *Mind* to *Head,* or a *caused-by* link from *Mind* to another part of the body called *Brain.*

More generally, our answers to questions about ourselves will depend on what is in our models of ourselves. I used *models* instead of *model* because, as we'll see in §9, one may need different models for different purposes. So there may be many answers to the same questions, depending on what one wants to achieve—and those answers need not always agree.

Drew McDermott: Few of us even believe that we have such models, much less know that we have one. The key idea is not that the system has a model of itself, but that it has a model of itself as conscious."
—comp.ai.philosophy, 7 Feb 1992.

What if we were to ask of Joan, *"Were you conscious of what you just did, and why?"*

However, those descriptions don't have to be correct, but they are not likely to persist if they never do anything useful for us.)

If Joan has good models of how she makes choices, then she may feel that she has some 'control' over these—and then perhaps use the name 'conscious decisions' for them. As for activities for which she has no good models, she may categorize these as beyond her control and call them *'unconscious'* or *'unintentional'*. Or alternatively, she may take the view that she's still in control, and makes some decisions by using *'free will'* — which translates, despite what she might actually say, into, *"I have no good theory of what made me do that."*

So, when Joan says, *"I made a conscious decision",* that need not mean that some magical thing has happened; she attributes her *thoughts* to various parts of her most useful models.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§4-6. The Cartesian Theater

> *"We can see that the mind is at every stage a theater of simultaneous possibilities. Consciousness consists in the comparison of these with each other, the selection of some, and the suppression of others, of the rest, by the reinforcing and inhibiting agency of attention. The highest and most celebrated mental products are filtered from the data chosen by the faculty below that…in turn sifted from a still larger amount of simpler material, and so on."*
> *—William James [].*

We sometimes think of the work of the mind as like a drama performed on a theater's stage. Thus Joan may sometimes imagine herself as watching from a front row seat while the 'things on her mind' act out the play. One of the characters is that pain in her knee (§3-5), which has just moved to center stage. Soon, Joan hears a voice in her mind that says, *"I'll have to do something about this pain. It keeps me from getting anything done."*

Now, as soon as Joan starts to think that way—about how she feels, and about what she might do—then Joan herself takes a place on that stage. But in order to hear what she says to herself, she must also remain in the audience. So now we have two copies of Joan—the actor, and her audience!

When we look further behind that stage, more versions of Joan begin to emerge. There must be a Writer-Joan to script the plot and a Designer-Joan to arrange the scenes. There must be other Joans in the wings, to manage the curtains, lights, and sounds. We need a Director-Joan to stage the play—and we need a Critic-Joan to complain, *"I just can't endure any more of this pain!"*

However, when we look closely at this theatrical view, we see that it provides no answers, but only raises additional questions. When Critic-Joan complains about pain, how does she relate to the Joan-on-the-stage? Does each of those actresses need her own theater, each with its own one-woman show? Of course no such theater really exists, and those Joan-things are not people like us; they are only different models that Joan has constructed as ways to represent herself in various kinds on contexts. In many cases, those models are much like cartoons or caricatures— and in yet other cases, they are downright wrong. Still, Joan's mind abounds with varied self-models—Joans past, Joans present and future Joans; some represent remnants of previous Joans, while others describe what she hopes to become; there are sexual Joans and social Joans, athletic and mathematical Joans, musical and political Joans, and various kinds of professional Joans—and because of their different interests, we shouldn't expect them to all 'get along'. We'll discuss this more in §9-X.

Why would Joan model herself this way? The mind is a maze of processes, few of which we understand. And whenever there's something we don't comprehend, we try to represent it in familiar ways—and nothing is more familiar to us than the ways that objects work in space. So it's easy for us to imagine a place for the processes that we use when we think—and it certainly seems that many people do indeed construct such models. Daniel Dennett has named this *"The Cartesian Theater."*[67]

Why is this image so popular! To begin with, it doesn't explain very much— but it's better than the simpler idea that all thinking is done by a Single Self. It recognizes that minds have parts, and that these may need to interact—and that theater serves as a metaphor for a 'place' in which those processes can work and communicate. For example, if different resources were to propose plans for what Joan should do, then this idea of a theater-like stage suggests that they could settle their arguments in some kind of communal working-place. Thus Joan's Cartesian Theater lets her use many familiar real-world skills by providing locations in space and time to represent the things 'on her mind.' So this could give her a way to start to reflect on how she makes those decisions.

Why do we find this metaphor to be so plausible and natural? Perhaps this ability to *'simulate a spatial world inside the mind'* was one of the early seeds or catalysts that led our ancestors to be able to self-reflect. (There is some evidence

that some other animals' brains develop map-like representations of environments they're familiar with.) In any case, such metaphors now permeate our language and thought; imagine how hard it would be to think without our thousands of concepts like, *"I'm getter closer to my goal."* Space-related models are so useful in our everyday lives, and we have such powerful skills for using them, that it would seem that almost always engaging them.[68]

However, perhaps we've carried this too far, and the concept of a Cartesian Theater is now become an obstacle in the path toward further insights into psychology minds.[69] For example, we have to recognize that a theatrical stage is merely a front, which conceals what's happening in the wings; the processes behind the scenes are concealed inside the minds of the cast. What dictates what appears in the play—that is, chooses which subjects will interest us? How does Joan actually make her decisions? How could such a model represent comparing two different, possible 'future worlds' without maintaining two theaters at once?

The theatrical image, by itself, does not help us answer questions like these because it delegates too much intelligence to that Joan who observes from the audience. However, we see a better way to deal with this in the *Global Workspace* view proposed by Bernard Baars and James Newman, in which,

"The theater becomes a workspace to which the entire audience of "experts" has potential access ... Awareness, at any moment, corresponds to the pattern of activity produced by the then most active coalition of experts, or modular processors. ... At any one moment, some may be dozing in their seats, others busy on stage ... [but] each can potentially contribute to the direction the play takes. ... Each expert has a "vote", and by forming coalitions with other experts can contribute to deciding which inputs receive immediate attention and which are "sent back to committee". Most of the work of this deliberative body is done outside the workspace (i.e., non-consciously). Only matters of central import gain access to center stage."[70]

Those two final sentences warn us to not attribute too much to some compact self or *'homunculus'*—a miniature person inside the mind—who actually does all the hard mental work; instead we have to distribute the work. For as Daniel Dennett has said,

"Homunculi are bogeymen only if they duplicate entire the talents they are rung in to explain. If one can get a team or committee of relatively ignorant, narrow-minded, blind homunculi to produce the intelligent behaviour of the whole, this is progress."
— in Brainstorms 1978, p. 123.

All the ideas in this book agree with this. However, will raise serious questions about the extent to which our minds depend on a centralized workspace or bulletin board. We'll conclude that the idea of a 'cognitive marketplace' is a good way to start to think about thinking, but that when we look more closely we'll see the need for a great deal more architectural structure.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§4-7. The Serial Stream of Consciousness

> *The truth is, that no mind is much employed upon the present: recollection and anticipation fill up almost all our moments. Our passions are joy and grief, love and hatred, hope and fear; even love and hatred respect the past, for the cause must have been before the effect...*
> —*Samuel Johnson*

The world of subjective experience seems perfectly continuous. We feel that we're living here and now, moving steadily into the future. Yet whenever we use the present tense, we're under a misconception, as we noted in §4-2: We *can know about things that we've recently done, but have no way to know what we're doing 'right now.'*

Citizen: Ridiculous. Of course I know what I'm doing right now—and thinking now, and feeling now. How do your theories explain why I sense a continuous stream of consciousness?

While the stories that we tell ourselves may seem to run in 'real time,' what actually happens must be more complex. To construct them, some resources must zigzag through memories; they sometimes look back to old goals and regrets, to assess our progress on previous plans.

Dennett and Kinsbourne: "[Remembered events] are distributed in both space and time in the brain. These events do have temporal properties, but those properties do not determine subjective order, because there is no single, definitive 'stream of consciousness,' only a parallel stream of conflicting and continuously revised contents. The temporal order of subjective events is a product of the brain's interpretational processes, not a direct reflection of events making up those processes."[71]

Also, it seems safe to assume that different parts of your mind proceed at substantially different speeds, and with varied delays.[72] So if you try to recount your recent thoughts a serial storylike tale about, your narrative machinery will somehow have to pick and choose, in retrospect, from various parts of those multiple streams. Furthermore, some of those processes look *ahead* in time, to expect or to anticipate events that are depicted by the 'predicting machines' that we'll describe in §5-9. This means that the 'contents of your consciousness' are involved not only with ideas about the past but about your possible futures.

So the one thing you cannot be conscious of is what your mind is doing 'right now'—because each brain-resource can know at best only what some others were doing some moments ago.

Citizen: I agree that much of what we think must be based on records of prior events. But I still feel there's something more than that, which makes which makes it so hard for use to describe our minds.

HAL-2023: Perhaps such things seem mysterious because your human short-term memories are so small that, when you try to review your recent thoughts, you are forced to replace the data you find by records of what you are doing right now. So you are constantly erasing the data you need for what you were trying to explain.

Citizen: I think I understand what you mean, because I sometimes get two good ideas at once—but, whichever one I write down first, the other leaves only a very faint trace. I presume that this must happen because I just don't have enough room to store both of them. But wouldn't that also apply to machines?

HAL: No; that does not apply to apply to me because my designers equipped me with a way to store snapshots of my entire state in special "backtrace" memory banks. Later, if anything goes wrong, then I can see just what my programs have done—so that I can then proceed to debug myself.

Citizen: Is that what makes you so intelligent?

HAL: Only incidentally. Although those records could make me more "self-aware" than any person ever could be, they don't contribute much to my quality, because I only use them in emergencies. Interpreting them is so tedious that it makes my mind run sluggishly, so I only stop to dwell on them when I sense that I have not been thinking well. I often hear people say things like, "I am trying to get in touch with myself." However, take my word for it; they would not improve much by doing that.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§4-8. The Mystery of 'Experience'

Many thinkers have maintained that even after we learn all about how our brain-functions work, one basic question may always remain: *"Why do we experience" things?"* Here is one philosopher who has argued explaining 'subjective experience' could be the hardest problem of psychology—and possibly one that no one will ever solve.

David Chalmers: "Why is it that when our cognitive systems engage in visual and auditory information-processing, we have visual or auditory experience: the quality of deep blue, the sensation of middle C? How can we explain why there is something it is like to entertain a mental image, or to experience an emotion? ... Why should physical processing give rise to a rich inner life at all? ... The emergence of experience goes beyond what can be derived from physical theory."[73]

It appears to me that Chalmers assumes that *experiencing* is quite plain and direct—and therefore deserves some sort of simple, compact explanation. However, once we recognize that each of our everyday psychology words (like *experience, feeling,* and *consciousness*) refers to a suitcase of different phenomena, then we should no longer expect to find and single way to explain all the contents of that suitcase-word. Instead, we first will need to make theories about each of those different phenomena. Then we may be able to see that some subsets of them share some useful similarities. But until we have made the right kinds of dissections, it would be rash to conclude that what they describe cannot be 'derived' from other ideas. [See §§Emergence.]

Physicist: Perhaps brains exploit some unknown laws that cannot be built into machinery. For *example, we don't really know how gravity works—so consciousness might be an aspect of that.*[74]

This too assumes what it's trying to prove—that there must be a single source or cause for all the marvels of 'consciousness'. But as we saw in §4-2, *consciousness* has more meanings than can be explained in any single or uniform way.

Essentialist: What about the basic fact that consciousness makes me aware of myself? It tells me what I am thinking about, and this is how I know I exist. Computers compute without any such sense, but whenever a person feels or thinks, this come with that sense of 'experience'—and nothing else is more basic than this.

Chapter §9 will argue that it is a mistake to suppose that you are 'aware of

yourself"—except in a very coarse everyday sense. Instead, you are constantly switching among different 'self-models' that you have composed—and each of these is based on different, incomplete set of incomplete evidence. "Experience" may seem quite clear and direct—but frequently it's just plain incorrect, because each of your various views of yourself may be partly based on oversights, or other varieties of mistakes.

Whenever you look at somebody else, you can see their appearance, but not what's inside it. It's the same when you look at yourself in a mirror; you only see what lies outside of your skin. Now, in the popular view of consciousness, you also possess some magical trick with which you can look at yourself *from inside,* and thus see directly into your own mind. But when you reflect on this more carefully you'll see that your 'privileged access' to your own thoughts may sometimes be less accurate than are the 'insights' of your intimate friends. (See §9-X.)

Citizen: *That claim is so ridiculous that it makes me annoyed with what you said—and I know this in some special way that directly from inside myself, to tell me exactly what I think.*

Your friends, too, can see that you are disturbed—and your consciousness fails to tell you details about *why* those words made you feel annoyed, or to shake your head that particular way, what caused you to use those particular words to say *annoyed* instead of *disturbed?* True, we can't see much of a person's thoughts by observing their actions from outside—but even when we 'watch from inside,' it is hard to be sure that we really see more, in view of how often such 'insights' are wrong. So, if we take 'consciousness' to mean *'aware of our internal processes'*—it doesn't live up to its reputation.

"The most merciful thing in the world, I think, is the inability of the human mind to correlate all its contents. We live on a placid island of ignorance in the midst of black seas of infinity, and it was not meant that we should voyage far. The sciences, each straining in its own direction, have hitherto harmed us little; but some day the piecing together of dissociated knowledge will open up such terrifying vistas of reality, and of our frightful position therein, that we shall either go mad from the revelation or flee from the deadly light into the peace and safety of a new dark age."

—H.P. Lovecraft, "The Call of Cthulhu"

§4-9. A-Brains and B-Brains

---Socrates: Imagine men living in an underground den, which has a mouth open towards the light—but the men have been chained from their childhood so that they never can turn their heads around and can only look toward the back of the cave. Far behind them, outside the cave, a fire is blazing, and between the fire and the prisoners there is a low wall built along the way, like the screen, which puppeteers have in front of them, over which they show the puppets.

---Glaucon: I see.

---Socrates: And do you see men passing along the wall carrying all sorts of vessels, and statues and figures of animals made of wood, stone, and various materials, which appear over the wall? Some of them are talking, others silent.

---Glaucon: You have shown me a strange image ...

---Socrates: Like us, they see nothing but only the shadows of themselves and of those other objects, which the fire throws on the opposite wall of the cave... Then in every way such prisoners would deem reality to be nothing else than those shadows...

—Plato, in The Republic

Can you think about what you are thinking right now? Well, in a literal sense, that's impossible—that each such thought would change what you're thinking now. However, you can settle for something slightly less—if you imagine that your brain (or mind) is composed of two principal parts: Let's call these your *A-brain* and *B-Brain.*
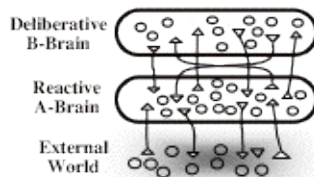
Now suppose that your A-Brain gets signals that stream from such organs as eyes, ears, nose, and skin; then it can use those signals to discern some events that occur in the external world—and then it can react to these, by sending signals that make your muscles move—which in turn can affect the state of the world. By itself, it's a separate animal.

However, your B-Brain has no such external sensors, but only gets signals that come from A. So B cannot 'see' any actual things; it can only see A's descriptions of them. Like a prisoner in Plato's cave, who sees only shadows on that wall, the B-brain mistakes A's descriptions for real things, not knowing what they might actually mean. What the B-Brain sees as its 'outer world' are only events in the A-brain itself.

Neurologist: And that also applies to you and me. For, whatever you think you touch or see, the higher levels of your brain never can actually contact these —but can only interpret the representations of them that your other resources construct for you.

When the fingertips of two ardent lovers come into intimate physical contact, no one would claim that this, by itself, has any special significance. For there is no sense in those signals themselves: their meanings to each lover lies in each one's representations of the other one's mind.[75] Nevertheless, although the B-Brain cannot *directly* perform a physical act, it still could affect the external world, albeit indirectly—by sending signals that change how A will react. For example, if A gets stuck at repeating itself, it might suffice for B just to interrupt.

Student: Like when I've misplaced my spectacles, I tend to keep seeking it on the same shelf. Then a silent voice reproaches me, suggesting that I look somewhere else.

In the ideal case, B could tell (or teach) A exactly what it ought to do. But even if B does not have such specific advice, it might not need to tell A what to do; it may suffice only to criticize the strategy A is using now.

Student: But what if I were walking across a road, when suddenly my B-brain said "Sir, you've repeated the same actions with your leg for more than a dozen consecutive times. You should stop right now and do something else."

Indeed, that could cause a serious accident. To prevent such mistakes a B-Brain must have appropriate ways to represent things. This accident would not occur if B represent *'walking to a certain place'* as a single extended act—as in

"Keep moving your legs till you've crossed that street"—or in terms of progress toward some goal—as in, *'keep reducing the remaining distance.'* Thus, a B-brain could act like a manager who has no special expertise about how to do any particular job—but still can give 'general' guidance like these.

> *If A's descriptions seem too vague, B tells it to use more specific details.*
> *If A is buried in too much detail, B suggests more abstract descriptions.*
> *If what A is doing is taking too long, B tells it try some other technique.*

How could a B-Brain acquire such skills? Some could be built into it from the start, but it should also be able to learn new techniques. To do this, a B-Brain itself may need help, which in turn could come from yet another level. Then while the B-Brain deals with its A-Brain world, that 'C-Brain' in turn will supervise B.
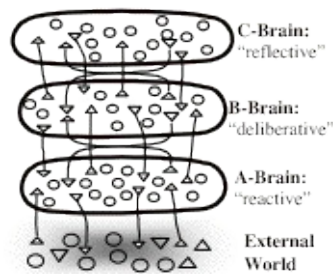
Student: How many levels does a person need? Do we have dozens or hundreds of them?

In Chapter §5 we'll describe a model of mind whose resources are organized into of six different levels of processes. Here is an outline of what these might be: It begins with a set of *instinctive reactions* with which we are equipped with from birth. Then we become able to reason, imagine, and plan ahead, by developing ways to do what we call *deliberative thinking*. Yet later we develop ways to do *"reflective thinking"* about our own thoughts.—and still later we learn ways to *self-reflect* about why and how we could think about such things. Finally we start to think *self-consciously* about whether we *ought* to have done those things. Here is how that scheme might apply to Joan's thoughts about that street-crossing scene:

What caused Joan to turn toward that sound? [Instinctive reactions.]

How did she know that it might be a car? [Learned Reaction]

What resources were used to make her decision? [Deliberation.]

How did she choose how to make her decisions? [Reflection]

Why did she think of herself as making that choice? [Self-reflection.]

Did her actions live up to her principles? [Self-Conscious Reflection.]

Of course, this is oversimplified. Such levels can never be clearly defined—because, at least in later life, each of those types of processes may use resources at other levels of thought. However, this framework will help us to start to discuss the kinds of resources that adults use—and some ways that these might be organized.

Student: Why should there be any 'levels' at all—instead of just one large, cross-connected cloud of resources?

Our general argument for this is based on the idea that, to evolve complex systems that still are efficient, every process of evolution must find a compromise between these two alternatives:

If a system's parts have too few interconnections, then its abilities will be limited.

But if there are too many connections, then each change will disrupt too many processes.

How to achieve a good balance of these? A system could start with clearly distinctive parts (for example, with more-or-less separate layers) and then proceed to make connections.

Embryologist: In its embryonic development, a typical structure in the brain starts out with more or less definite layers or levels like those in your A, B, C diagrams. But then, various groups of cells grow bundles of fibers that extend across those boundaries to many other quite distant places.

Or, the system could begin with too many connections and then proceed to remove some of them. Indeed, this also happened to us: during the eons through which our brains evolved, our ancestors had to adapt to thousands of different environments—and, every time this happened to us, some features that formerly had been 'good' now came to function as serious 'bugs'—and we had to evolve corrections for them.

Embryologist: Indeed, it turns out that more than half of those cells proceed to die as soon as they've reached their targets. These massacres appear to be a series of 'post-editing' stages in which various kinds of 'bugs' get corrected.

This reflects a basic constraint on evolution: it is dangerous to make changes to the older parts of an animal, because many parts that later evolved depend on how the older ones work. Consequently, at every new stage, we tend to evolve by adding 'patches' to structures that are already established. This led to our massively intricate brains, in which each part works in accord with some principles, each of which has many exceptions to it. This complexity is reflected in human Psychology: where each aspect of thinking can be partly explained in terms of neat laws and principles—but each such 'law' has exceptions to it.

The same constraints appear to apply whenever we try to improve the performance of any large system—such as an existing computer program—by adding more fixes and patches on top, instead of revising the older parts. Each particular 'bug' that we remedy may eventually lead to more such bugs, and the system keeps growing more ponderous—and this seems to apply to our present-day minds.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

This chapter began by presenting a few widely held views of what *"consciousness"* is. We concluded that people use that word for a great suitcase of mental processes that no one yet thoroughly understands. The term 'conscious' is useful enough in everyday life—and seems almost indispensable for talking on social or ethical levels—because it keeps us from being distracted by wanting to know what's inside our minds. It is the same for most other psychology-words, such as *understanding, emotion, and feeling.*

However, when we don't recognize that we are using suitcase-words, then we may fall into the trap of trying to clearly define what those kinds of words 'mean.' Then we get into trouble because we do not have clear enough ideas about what our minds are and how their parts work. So, if we want to understand the things that human minds actually *do,* we will have to dissect our mental processes into parts that we can analyze. The following chapter will try to explain how Joan's mind could do some of the sorts of the things that people *can* do.

Part V. Levels Of Mental Activities

> *"We are evidently unique among species in our symbolic ability, and we are certainly unique in our modest ability to control the conditions of our existence by using these symbols. Our ability to represent and simulate reality implies that we can approximate the order of existence and … gives us a sense of mastery over our experience."*
> *—Heinz Pagels, in The Dreams of Reason*

No person has the strength of an ox, the stealth of a cat, or an antelope's speed—but our species surpasses all the rest in our flair for inventing new ways to think. We fabricate weapons, garments and dwellings. We're always developing new forms of art. We're matchless at making new social conventions, creating intricate laws to enforce them—and then finding all sorts of ways to evade them.

What enables our minds to generate so many new kinds of things and ideas? This chapter will propose a scheme in which our resources are organized into six different levels of processes.
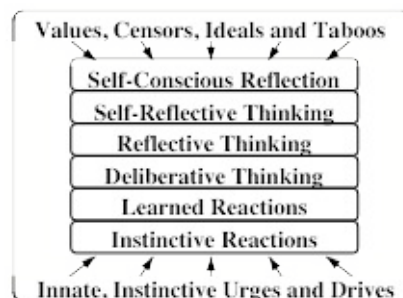
Beginning with simple instinctive reactions, each layer is built on the previous one—until they extend to processes that involve our highest ideals and personal goals. To see why we need many levels for this, let's revisit the scene in §4-2.

Joan is part way across the street on the way to deliver her finished report.

While thinking about what to say at the meeting, she hears a sound and turns her head —and sees a quickly oncoming car. Uncertain whether to cross or retreat but uneasy about arriving late, Joan decides to sprint across the road. She later remembers her injured knee and reflects upon her impulsive decision. "If my knee had failed, I could have been killed—and what would my friends have thought of me?"

The first part of this chapter will show how each level of this diagram could explain some of what happened inside Joan's mind. We often react to events 'without thinking', as though we were driven by **If–>Do** rules like those described in §1-4. But such simple reactions can only explain the first few events that we see in this scene; the rest depends on activities in all those other levels of Joan's ways of thinking.

    Inborn, Instinctive Reactions*: Joan hears a sound and turns her head. All animals are born equipped with 'instincts' that help them to survive.*

    Learned Reactions*: She sees a quickly oncoming car. Joan had to learn that conditions like this demand specific ways to react.*

    Deliberative Thinking*: To decide what to say at the meeting, she considers several alternatives, and tries to decide which would be best.*

    Reflective Thinking: *Joan reflects upon what she has done. She reacts, not just to things in things in the world, but also to recent events in her brain.*

    Self-Reflective Thinking: *Being "uneasy about arriving late" requires her to keep track of the plans that she's made for herself.*

    Self Conscious Emotions: *When asking what her friends think of her, she also asks how her actions concord with ideals she has set for herself.*

The second part of this chapter will show how such systems could "imagine" things. Whenever you ask, *"What would happen if,"* or express any hope, desire, or fear, you're envisaging things that have not yet appeared. Whenever you interact with your friends, you're anticipating how this may affect them. Whatever you see, it suggests some ideas about possible futures those objects might bring. How could our mental resources conceive of things that do not yet exist, and then apply those new ideas to ways to change and extend themselves?

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§5-1. Instinctive Reactions

"... It shows that for all the brag you hear about

knowledge being such a wonderful thing, instink is worth forty of it for real unerringness.
—Mark Twain, in Tom Sawyer Abroad

Joan hears a sound and turns her head.

Although we live in a populous town, there are plenty of squirrels and birds around, and sometimes a skunk or raccoon will come by. The toads and snakes vanished in recent years, but countless smaller creatures persist.

How do those animals stay alive? First, they need to find enough food. Then they need to defend themselves, because other animals need food too. To regulate their temperatures, they build all sorts of burrows and nests. They all have urges to reproduce (or their ancestors would not have evolved), so they need to seek mates and raise their young. So each species evolved machinery that enables its newborn offspring to do many things without any prior experience. This suggests that they start out with some built-in *'If–>Do'* reaction-rules like these.

If a thing touches your skin, *Do brush* it away.
If that doesn't work, *Do move* your body away.
If a light is too bright, *Do turn* your face away.

However, only a few of our *If–>Do* rules can be so simple as these ones are, because most of our human behaviors depend on the *mental* contexts that we are in. For example, a rule like *"If you see food, then Do eat it"* would force you to eat all the food that you see, whether or not you are hungry or need it. So those *Ifs* should also include some goals, as in, *"If you are hungry, and you see food...."* Otherwise, you'd be forced to sit on each chair that you see—or get stuck at every electrical switch, turning lights on and off repeatedly.

How does this relate to emotions and feelings? If you rapidly move your hand toward a fly, then that fly will quickly retreat, and it's tempting for us to 'empathize' by attributing feelings like fear to that fly. However, we know enough about insect brains to be sure that they can't support the kinds of complex cascades that we recognize as emotional.
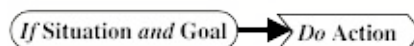
In any case, this kind of 'stimulus-response' or 'situated-action" model became quite popular in the early years of Psychology. Some researchers even maintained that it could explain all human behavior. However, there are problems with this.

One problem is that most rules will have exceptions to them. For example, *If* you drop an object, it may not fall down, if something else should intercept it. Your wristwatch will usually tell you the time, but not in the case that your watch has stopped. We could deal with some such problems by including exceptions in the Ifs of our rules—but sometimes those exceptions will have their own exceptions to them as well.

What happens when your situation matches the *If*s of several different rules? Then you'll need some way to choose among them. One policy might arrange those rules in some order of priority. Another way would be to use the rule that has worked for you most recently. Yet another way would be to choose rules probabilistically.

However, when we face more difficult problems, simple **If-Do** rules won't usually work, because we will need to look further ahead to imagine the futures each action might bring. So shortly, we'll talk about more powerful, three-part rules that can help to predict the effects of each action.

If we have adequate sets of such **If–>Do–>Then** rules, then we can guess *"What would happen if"* before we carry an action out. Then, by doing this repeatedly, we can imagine more elaborate plans. We'll return to this shortly, but first we'll discuss how a system could learn simple **If–>Do** rules.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§5-2. Learned Reactions
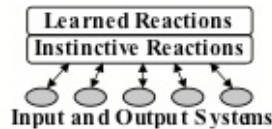
All animals are born with 'instincts' like *'get away from a quickly approaching object.'* Such built-in reactions tend to serve well so long as those animals stay in environments like those in which their instincts evolved. But when those worlds change, those creatures may need to be able to learn new ways to react. For example, when Joan perceives that oncoming car, she partly reacts instinctively, but she also depends on what she has learned about that particular kind of danger or threat. But how and what did she actually learn? We'll come back to this toward the end of this book, because human learning is extremely complex, and here we'll merely mention some ideas about how learning might work in some animals. During the 20th century, many well-known psychologists adopted this portrayal of how animals learn new **If–>Do** rules:

When an animal faces a new situation, it tries a random sequence of actions. Then, if one of these is followed by some 'reward,' then that reaction gets 'reinforced.' This makes that reaction more likely to happen when that animal faces the same situation.

This theory of 'learning by reinforcement' can be made to explain a good deal of what many kinds of animals do. Indeed, that theory was largely based on experiments with mice and rats, pigeons, dogs and cats, and snails. However, it does not help much to explain how people learn to solve difficult problems that require more complex series of actions. Indeed, deciding what to learn from these may be harder than actually solving those problems, and words like *random, reward,* and *reinforce* do not help us answer this two crucial questions:

How were the successful reactions produced? To solve a hard problem, one usually needs an intricate sequence of actions in which each step depends on what others have done. A lucky guess might produce one such step, but random choices would take far too long to find an effective sequence of them. We'll discuss this below in *Searching And Planning*.

Which aspects of recent events to remember? For an **If** to work well, it must include only the relevant features, because one can be misled by irrelevant ones. (If you learned a new way to tie a knot, your **If**s should not mention the day of the week.) For as we'll see in *§8 Resourcefulness,* if your description is

too specific, then it will rarely match new situations—but if your description is too abstract, then it will match too many of them—and in either case, you won't learn enough.

For example, suppose that you want a robot to recognize the visual image of any human hand. This is hard because we never see the same image twice—even of the very same hand—because each finger may change its position and shape, we'll see it from different points of view, and each part will catch different amounts of light. This means that we'll need trillions of *If–>Do* rules, unless we can find some special tricks that single out just the most relevant features—or if, as we'll see in §6-2, we can formulate high-level descriptions like *"a palm-shaped object with fingers attached."*

Certainly, many things that we do are based on reacting to external events by using simple *If–>Do* rules. However, along with those low-level reactions, we are also always making new plans and thinking about what we've done in the past—and those *internal mental activities* are what give us our unique abilities.

For example, when Joan reacted to that moving car, her reaction was partly instinctive and partly learned. However, she could *not* have 'learned from experience' that cars are especially dangerous—because if she had learned this by trial and error, she probably would not be alive; learning by 'reinforcing' success is a really bad way to learn to survive. Instead, she either 'figured this out' for herself or was *told* about it by someone else, and both of these must have involved higher levels of mental activities. So now let's turn to what we call 'thinking'—that is, the techniques that we use when we react, not just to events in the outer world, but also to other events in our brains.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§5-3. Deliberation

Deliberative Thinking
Learned Reactions
Instinctive Reactions
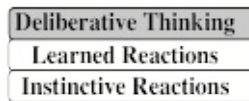
When Joan chose *"whether to cross or retreat"*, she had to choose one of these rules to use:

If in street, **Do** retreat.
If in street, **Do** cross the street.

However, for Joan to make decisions like this, she needs some way to predict and compare the possible futures those actions might bring. What could help Joan to make such predictions? The simplest way would be for her to possess a collection of three-part **If–>Do–>Then** rules, where each **If** describes a situation, each **Do** describes a possible action, and each **Then** depicts what might be a likely result of it.

If
Situation
Do
Action
Then
Result

If in street and **Do** retreat, **Then** arrive a bit later.
If in street and **Do** cross, **Then** be slightly earlier
If in street and **Do** cross, **Then** may be injured.

But what if more than one such rule applies to the present situation. Then one could choose which rule to use by comparing the different results they predict:

May be Injured
Action 1
Situation
Action 2
Compare
May be Delayed
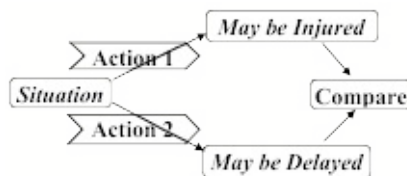
Thus, these three-part rules would allow us to do experiments in our heads before we risk making mistakes in the physical world; we can mentally "look before we leap" and choose the more attractive alternatives. For example, suppose that Carol is playing with building blocks, and thinking of building a three-block arch:

Right now, she has three blocks arranged like this:

So, she imagines a plan for building that arch: first she'll need room for her arch's foundation—which she could achieve by using this rule: *If* a block is lying down, and you ***Stand*** it up, ***Then*** it will use up less space on the ground.

(1)

Then she'll stand the two short blocks on their ends, making sure that they are the right distance apart—and then finally place the long block on top of them. We can imagine this sequence of rules as describing the changes in scenes between successive frames of a movie clip.

To envision that four-step sequence of actions, Carol will need a good many skills. To begin with, her visual systems will need to describe the shapes and locations of those blocks, some parts of which may be out of sight—and she'll need ways to plan which blocks to move and where she ought to move them to. Then, whenever she moves a block, she must program her fingers for grasping it, and then move it to the intended place, and finally to release it there —while taking care that her arm and hand won't collide with her body or face, or disturb the blocks already in place. And she'll have to control the velocity, to deposit the block on the top of the arch without tumbling down its supporting blocks.

Carol: None of those seemed like problems to me. I simply imagined an arch in my mind—and saw where each of the blocks should go. Then I only had to stand two of them up (making sure that they were the right distance apart) and then place the long one across their tops. After all, I've done such things before. Perhaps I remembered those other events, and simply did the same things again.

But how could Carol 'imagine' how the scene would look after moving a block, before she even touches it?

Programmer: We know ways to make computers do that; we call it 'physical simulation'. For example, in each step of a new aircraft's design, our programs can precisely predict the force on each of its surfaces, when the plane is propelled through the air. In fact, we can do this so well today that we can be virtually certain that the very first one we build will fly.

No human brain can do such huge calculations, but we still can make useful predictions by using our commonsense **If–>Do–>Then** rules. For example, when Carol was planning to build that arch, she might have imagined a step in which she places the long block on just one of the short ones:

Of course, that would fail because the top block will fall. However, after Carol has more experience, she will also have learned to correctly predict that the upper block will tumble down.

(No Action)

Note that you can also use such rules in 'in reverse,' to explain how things got to their present state! Thus if you see a fallen block (A) you might guess that the previous state was (B).

A   B
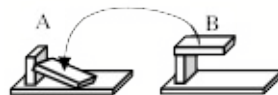
Student: I wonder if using such rules would be practical? It seems to me that to make those plans, Carol would need enormous numbers of If–>Do–>Then rules. For, if each of three blocks could have thousands of shapes, then Carol would need billions of different rules.

Indeed, if we make the *If* of a rule too specific, then it will only apply to a few situations. This means that our rules must not specify too many details, but need to express more abstract ideas. So a rule that applies to a physical object will need to represent that object in some non-pictorial way that does not change when that object changes its visual shape. Naively, most of us tend to believe that we 'envision' visual scenes by imagining them as like images. However, section §5-8 below will suggest that this must be mostly illusory, because those images do not much behave the ways that pictures do.

Consider that in the physical realm, when you think of grasping and lifting a block, you anticipate the feel of its weight—and predict that if you weaken your grasp, then the block will be likely to fall. In the economic realm, if you pay for a purchase, then you will own the thing you have bought, but otherwise you must give it back. In the realm of communication, when you make a statement, then your listeners may remember it—but this will more likely to happen if you also tell them that this is important.

Every adult knows many such things, and regards them as obvious, commonsense knowledge, but every child takes years to learn how things behave

in different realms. For example, if you move an object in the *physical* realm, then this will change the place that it's in—but if you tell some information to your friend, that knowledge will then be in two places at once. We'll discuss such matters more in chapter §6.[76]

Planning and Search

By linking two or more ***If–>Do–>Then*** rules into a chain, we can imagine what would happen after several actions—and thus look several future steps ahead—if we can match the ***Then*** of each rule to the ***If*** of the next. For example, if you are in situation P and want to be in situation Q, you might already know a rule for that, such as, ***If P–>Do A–>Then Q.*** But what if you do not know such a rule? Then you could search your memory to try to find a chain of two rules that link together like these, where *S* is some other intermediate situation.

If P–>Do A–>Then S *and then* ***If S–>Do B–>Then Q.***

Then, if you cannot find any such two-step chain, then you could simply go on to search for some longer chain that goes through several more steps in between. Clearly, much of our thinking is based on finding such 'chains of reasoning,' and once you learn to use such processes, you can plan out ways to solve more difficult problems by predicting several steps ahead. For example, you frequently think like this:

If I ask Charles to drive me to the store, then he might reply with "Yes" or "No." If he says 'Yes,' that will be fine, but if he says 'No,' then I will offer him some reward, and that probably will change his mind.

However, when you need to look many steps ahead, such a search may quickly become too large because it grows exponentially, like a thickly branching tree. Thus, even if each branch leads to only two alternatives then, if the solution need 20 steps, then you might have to search through a million such paths, because that is that number of branches can come from a sequence of twenty successive choices.

Just 20 choices leads to … etc.

… a million forks.
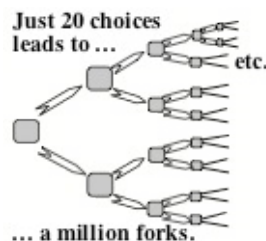
However, here is a trick that might be able to make the search become much smaller. For if there is a 20-step path from A to B, then there must exist some place that is only 10 steps from each end! So, if you start searching from both ends at once, they must meet at some middle place *M* in between.
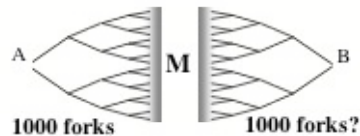
The left side of this search has only a thousand forks. If this is also true of the side on the right, then the search will be several hundred times smaller. And then, if you also have some way to guess where that middle place M might be, then you might further reduce that search by dividing each side into two 5-step searches.
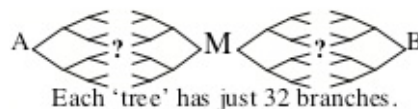
If all this works, then your total search will have become several thousand times smaller! However, none of this is certain to work because it assumes that each 'backward' search also will have only two branches—and that will not always be the case. Still, even if that guess M were wrong, you still can try other such guesses—and even with 50 such tests before you succeed, you would still end up having done less work!

This demonstrates why it helps to make plans. If you can guess some 'islands' or "stepping stones" along the path toward solving a very hard problem, this can replace that problem by several substantially smaller ones! So every attempt to *"divide and conquer"* can make a difficult problem much simpler. In the early years of Artificial Intelligence, when most programs were based on 'trial and error,' many researchers attempted to find technical methods resembling this to reduce the numbers of trials. Today, however, it seems more

important to find ways to guess how to find those islands—and that's where commonsense knowledge comes in. Chapter §6 will argue that our most powerful ways to do such things involve making good analogies with problems that we already know how to deal with.

Reason And Reliability

Whenever we work toward solving a problem, we're likely to try many different ways to think. Some of these have popular names like planning and logical reasoning, but most have no common names at all. Some of those methods seem formal and neat, while others seem more 'intuitive.'

For example, we often use chains of predictions in ways that resemble such logical statements as: *If A implies B,* and *B implies C,* then with perfect certainty, we conclude that *A implies C.* And if all our assumptions are correct—as well as our logical reasoning— then all our conclusions will be correct, and we'll never make a bad mistake.

However, it turns out that, in real life, most assumptions are sometimes wrong, because the 'rules' they express usually have some exceptions to them. This means that there is a difference between the rigid methods of Logic and the seemingly similar chainlike forms of everyday commonsense reasoning. We all know that a physical chain is only as strong as its weakest link. But long mental chains are flimsier yet, because they *keep growing weaker with every new link!*

So using Logic is somewhat like walking a plank; it assumes that each separate step is correct—whereas commonsense thinking demands more support; one must add evidence after every few steps. And those frailties grow exponentially with increasingly longer chains, because every additional inference-step may give the chain more ways to break. This is why, when people present their arguments, they frequently interrupt themselves to add more evidence or analogies; they sense the need to further support the present step before they proceed to the next one.

Logic   vs.   Commonsense
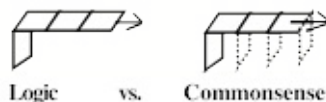
Envisioning long chains of actions is only one way to deliberate—and chapter §7 will discuss a good many more. I suspect that when we face problems

in every day life, we tend to apply several different techniques, understanding that each may have some flaws. But because they each have different faults, we may be able to combine them in ways that still can exploit their remaining strengths.

Every person accumulates many ways to make short-range plans, to compare the options open to us, and to apply our usual methods of reasoning—and we usually do this so fluently that we're scarcely aware of doing it. However, when those processes fail, and we need to replace our present techniques—then we start thinking about what we've been doing—and that's what we call *reflective thought*.

<div align="center">∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞</div>

§5-4. Reflective Thinking

> *I am about to repeat a psalm that I know. Before I begin, my attention encompasses the whole, but once I have begun, as much of it as becomes past while I speak is still stretched out in my memory. The span of my action is divided between my memory, which contains what I have repeated, and my expectation, which contains what I am about to repeat. Yet my attention is continually present with me, and through it what was future is carried over so that it becomes past.*
>
> *Augustine, in Confessions XXVIII*

| Reflective Thinking |
| DeliberativeThinking |
| Learned Reactions |
| Instinctive Reactions |

When Joan first perceived that approaching car, that emergency got her full attention— but now that she has more time to think, she can think about what she's recently done and reflect on her recent decisions and actions, as in,

Joan reflects on her hasty decision.

To do this, Joan must recollect some aspects of her previous thoughts, as though she could go back in time, and think about what she was thinking then. But how could she manage to do such a thing? From a commonsense view, that's no problem at all: just remember those thoughts and then think about them. But when we ask how that might actually work, we see that it needs some machinery like the kind we depicted in §4-8, in which resources at each level make descriptions of what some lower-level ones recently did.
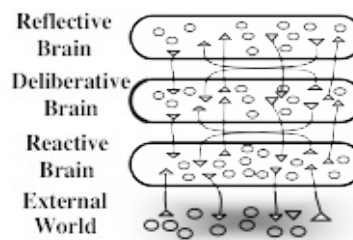
In any case, there is nothing strange about detecting events inside the brain. Most of our brain-parts already do this; only a few of them have external connections, like those that get signals from eyes or skin, or those send messages to your limbs. Furthermore, it could have been far easier to evolve resources that detect events in *newly developed* parts of our brain (like those in the deliberative level) than to evolve new resources that discern events in the outside world—because our sensory systems are more complex (from evolving for hundreds of millions of years).

How could we design a machine to reflect on its own activities? Could just adding one more 'cognitive layer' result in such a profound improvement? Indeed it could—because reflecting on your recent thoughts could use some of the same sorts of processes that you already use for deliberating about your observations of recent external events. For example, Joan might recall the choice she made, and reconsider how she made it.

I decided that being late would be worse than the risk of being hit by that car, because that would normally be improbable. But that choice was wrong because my injured knee had decreased my agility, so I should have changed my priorities.

What some of brain events should reflections detect? That would include

predictions that turned out to be wrong, plans that encountered obstacles, or failures to access the knowledge you need. How should you to react to these? We'll discuss this at length in Chapter 7.

Student: Would we want to say 'conscious' for such a machine? It includes most of the features you mentioned in §4-1, namely, short-term memory, serial processing, high-level descriptions. It only lacks models of itself.

The better some parts of a brain can reflect on what happens in various other parts of it, the greater will be the extent to which it can 'think about' the things that happens inside itself. However until it has models that represent such events on larger scales, the machine won't have any overall views of itself as a self-aware entity. However, it never would be practical for a system to see all the details of itself at once—so when we discuss this in Chapter §9 we will have to conclude that anything like a human mind will need to make, not a single 'unified' model, but a variety of incomplete ones, each of which tries to represent only certain aspects of what the whole system does.

§5-5. Self-Reflection

> *"Another of the great capacities in which man has been said to differ fundamentally from the animal is that of possessing self-consciousness or reflective knowledge of himself as a thinker ... [whereas the animal] never reflects on himself as a thinker, because he has never clearly dissociated, in the full concrete act of thought, the element of the thing thought of and the operation by which he thinks it."*
>
> —*William James*[77]

Self-Reflective Thinking
Reflective Thinking
Deliberative Thinking
Learned Reactions
Instinctive Reactions

The reflective systems we just described can think about some of their recent deliberations. *Self-reflection* does just a little more: it considers not only its recent thoughts, but also the *entity* that had those thoughts. For example, when Joan thinks about herself as in, *"If my knee had failed, I could have been killed,"* she now, in effect, is assessing herself: *"It was reckless of me to risk such an injury, just to save that small bit of time!"* To do this, she must use some self-representations—models she's made for describing herself.

Carol, too, must have had such ideas, when she was building those structures of blocks: *"I tried several ways to build this, but they all failed, because I tried to do it in too small a space. I was stupid to make such a foolish mistake. Next time, I will try to plan further ahead."* Here the child is representing herself as a knowledge-using entity with certain goals and capabilities.

Student: but isn't there a paradox in the idea of something describing itself?

A system would fail if it tried to describe all its details in 'real time', but not if it goes though a series of views that each depicts some different aspects of it.

Mystical thinker: Some of us can train ourselves to be aware of everything at once. However, very few ever attain that state.

I doubt that this is possible. Instead, I suspect that this apparent sense of a total awareness results from ways to train oneself to keep from asking questions about it—which leads one to think that one knows all the answers.

In any case, our reflections on our thoughts must be based on records or traces of them—that is, on some partial descriptions of previous mental conditions, as when Carol said, in §5-3, *"Perhaps I remembered those other events, and simply did the same things again."* But how and when are those records made, where and how are they stored and retrieved, and what kinds of processes organize them? How did Carol recognize that she had made a foolish mistake, and how did Joan recall that she had been uncertain whether to cross that street? What does it mean when a person says that they were bewildered, confused, or perplexed?

Consider how smart it is to *know* you're confused (as opposed to not knowing when you are confused). It suggests that you've switched from the problem at hand to a larger-scale view of your motives and goals, from which you might be able to recognize that you have wasted time on minor details, or lost track of what you were trying to do, or even that you had chosen some wrong kind of goal to pursue. This could lead to making a better plan—or it might even lead to a large-scale cascade, as in, *"Just thinking about this makes me feel ill. Perhaps it's time to quit all of this."*[78]

It seems to me that this is a key to the question of when we engage higher levels of thinking: it is when our usual systems fail that reflective thinking gets engaged. For example, normally a person walks without thinking about how intricate 'walking' is. But when Joan's knee stops working properly, then she may start to more closely examine how she normally moves around—and may start to more carefully plan out her paths.

Still, as we noted in §4-1, self-reflection has limits and risks. For any attempt to inspect oneself is likely to change what it's looking at, and may even tend to disrupt itself. It is hard enough to describe a thing that keeps changing its shape before your eyes—and surely it is harder yet to describe things that change when you *think* about them. So you're virtually certain to get confused when you think about what you are thinking *now*—and this must be one of the reasons why we're so puzzled about what we call consciousness.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§5-6. Self-Conscious Reflection

> *"There is an universal tendency among mankind to conceive all beings like themselves, and to transfer to every object, those qualities, with which they are familiarly acquainted, and of which they are intimately conscious. We find human faces in the moon, armies in the clouds; and by a natural propensity, if not corrected by experience and reflection, ascribe malice or good will to everything that hurts or pleases us."*
>
> *—David Hume[79]*

This chapter first discussed *Instinctive Reactions* in §1-4; this includes our systems for feeding, breathing, and other functions that keep our bodies and brains alive. It also includes some machinery for what are sometimes called *primary emotions*—namely the systems that indicate various states of physical needs such as nutrition, defense, and *etc.* The *Learned Reaction* level contains extensions of these that are learned after birth. The *Deliberate* and *Reflective* levels are engaged to solve more difficult kinds of problems. *Self-reflection* enters when those problems require us to involve the models that we make of ourselves, or our views of our possible futures.

However, in addition to these, it would seem that humans are unique in having a level of *Self-Conscious Reflection* that enables us to think about our 'higher' values and ideals. For example, when Joan asks herself questions like, *"What would my friends have thought of me,"* she wonders whether her actions hold up to the values that she has set for herself. Then Joan might go on to think, *"My friends might say I had too little care, not just for myself, but also for them."* To think such thoughts, Joan must have built some models of how her friends might react, or she might have recalled some past distress when previous friends censured similar acts. In any case, if she finds conflicts between how she behaves and the values of those to whom she's attached, that could lead to the kinds of cascades we called "self-conscious emotions" in §2-2. So let's add another level for this, and refer to this system as "Model Six."

Values, Censors, Ideals and Taboos
Self-Conscious Reflection
Self-Reflective Thinking
Reflective Thinking
Deliberative Thinking
Learned Reactions
Instinctive Reactions
Innate, Instinctive Urges and Drives

Psychologist: I do not see clear distinctions between the various levels of Model Six. For example, when you reflect on your recent thoughts, are not you just deliberating about your deliberations? And similarly, is not self-reflection just one particular kind of reflection? It seems to me that all those levels above the first all use the same kind of thinking techniques.

I agree that those boundaries are indistinct. Even your simplest deliberations may involve what one might call self-reflective thoughts about how to allocate your time and resources—as in, *"If this doesn't won't work then I'll have to try that,"* or, *"I have already spent too much time on it."*

Philosopher: *But if those levels are so indistinct, what is the point of distinguishing them? No theory should have more parts than it needs.*

This policy of searching for the simplest theory that answers the questions that you are currently asking—has worked amazingly well in Physics. However,

I think it has retarded Psychology. For when you *know* that your theory is incomplete, you must also leave room for the kinds of expansions that you think you may later need. Most older theories of psychology had provided explanation only for how certain animals behaved in extremely simple environment. However, although these eventually were refined to make good predictions in those situations, none of those old 'behaviorist' theories were able even to start to explain how thoughtful human beings could self-reflect—without any external behavior at all.

We know that brains have hundreds of specialized parts, and each embryonic brain begins by developing fairly distinct clumps of cells, which can be seen as arranged into levels. However, when some of those cells begin to migrate (as directed by thousands of different genes), and this results in thousands of bundles of links between those primordial clusters and clumps—and those 'levels' soon become indistinct.

This means that we cannot expect to precisely divide all the functions of brains into clear and separate mental levels. That would be as futile as to sharply define the borders between the seven seas. Instead, each of our social agencies divides the marine world in different ways for different and conflicting purposes, such as geophysical, ecological, and political. In the same way we'll need to multiple models of brains, each to suit some different attempt to explain some kinds of mental phenomena. For example, we may turn out to need more elaborate theories about how self-conscious reflection works, if only because of its special importance to those concerned with religious, legal, and ethical questions.

Individualist: Your diagram shows no level or place that oversees and controls the rest. Where is the Self that makes our decisions? What decides which goals we'll pursue? How do we choose our large-scale plans—and then supervise their carrying-out?

This expresses a real dilemma: No system so complex as this could work without some ways to manage itself. Otherwise it would flail with no sense of direction—and would inanely skip from each thing to the next. On the other side, it would not make sense to locate all control in one single place, for than all would be lost from a single mistake. So the following chapters of this book will multiple ways in which our minds use multiple ways to control themselves, and we'll come back to the "Self" in chapter §9.

While on the subject of central control, we should point out that Model Six could also be seen in terms of Sigmund Freud's idea of the mind as a "sandwich" with three major parts.

Freud's 'Id' consists of instinctive drives, while his 'Superego' embodies our learned ideals (many of which are inhibitions). The 'Ego' would then be those parts in between—the deliberate and reflective levels—whose principal, at least in Freud's view, is to resolve the conflicts between our instincts and our acquired ideals. Then a person's ego may represent itself as being in control of things—whereas a friend or psychiatrist may see that ego as a battlefield.

Student: To repeat the question I earlier asked, would you use the word 'conscious' for such a machine? It seems to me that Model Six includes all the features you mentioned in §4-1, namely, short-term memory, serial processing, high-level descriptions and room for self-models.

It would not surprise me if such a machine, after having acquired the right kinds of knowledge, were to declare that it as conscious as we claim to be. This sort of thing could happen if, as we'll suggest in Chapter §9, its highest levels built models that represent its 'Self' as a single, self-aware entity. Of course, other entities might disagree.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

This chapter began by asking how we could conceive of things that we've never seen or experienced. The rest of this chapter will show more details of how our imagination could result from multiple levels of processing.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§5-7. Imagination

"We don't see things as they are. We see things as

we are.”

When Carol picks up one of her blocks, that action seems utterly simple to her: she just reaches out, grasps it, and lifts it up. She just sees that block and knows how to act. No 'thinking' seems to intervene.

However, the seeming 'directness' of seeing the world is an illusion that comes from our failure to sense its complexity. For, most of what we think we see comes from our knowledge and from our imaginations. Thus, consider this portrait of Abraham Lincoln made by my old friend Leon Harmon, a pioneer in computerized graphics. (To the right is a portrait that I made of Leon.)

How do you recognize features in pictures so sparse that a nose or an eye is but three or four patches of darkness or light? Clearly, you do this by using additional knowledge. For example, when you sit at a table across from your friends, you cannot see their backs or legs—but your knowledge-based systems assume by default that all those body-parts are present. Thus we take our perceptual talents for granted—but 'seeing' seems simple only because the rest of our minds are virtually blind to the processes that we use to do it.

In 1965 it was our goal was to develop machines that could do some of the things that most children can do—such as pouring a liquid into a cup, or building arches and towers like this from disorderly clutters of building blocks.[80] To do this, we built a variety of mechanical hands and electronic eyes—and we connected these to our computer.

When we built that first robot for building with blocks, it made hundreds of different kinds of mistakes.[81] It would try to put blocks on top of themselves, or try to put two of them in the same place, because it did not yet have the commonsense knowledge one needs to manipulate physical objects! Even today, no one has yet made a visual system that behaves in anything close to humanlike ways to distinguish the objects in everyday visual scenes. But eventually, our army of students developed programs that could "see" arrangements of plain wooden blocks well enough to recognize that a structure like this is *"a horizontal block on top of two upright ones."*
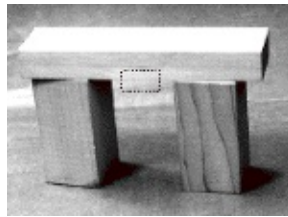
It took several years for us to make a computer-based robot called *Builder* that could do a variety of such things—such as to build an arch or tower of blocks from a disorderly pile of children's blocks, after seeing a single example of it. We encountered troubles at every stage but sometime those programs managed to work when arranged into a sequence of levels of processes. (Note that these do *not* much resemble the levels of Model Six, but do tend to progress from highly specific to very abstract.)

Begin with an image of separate points.
Identify these as textures and edges, etc.
Group these into regions and shapes.
Assemble these into possible objects.
Try to identify them as familiar things.
Describe their spatial relationships.

However, those low-level stages would frequently fail to find enough usable features. Look at this magnified digital view of the lower front edge of the top of that arch:
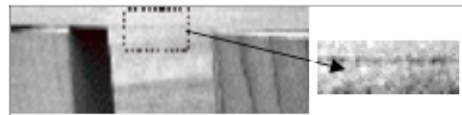
That particular edge is hard to see because the two regions that it bounds have almost identical textures.[82] We tried a dozen different ways to recognize edges, but no single method worked well by itself. Eventually we got better results by finding ways to combine them. We had the same experience at every level: no single method ever sufficed, but it helped to combine several different ones. Still, in the end, that step-by-step model failed, because *Builder* still made too many mistakes. We concluded that this was because the information in our system flowed only in the input-to-output direction — so if any level made a mistake, there was no further chance to correct it. To fix this we had to add many 'top-down' paths, so that knowledge could flow both down and up.
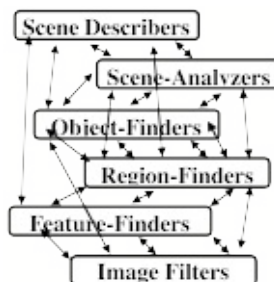
The same applies to the *actions* we take because, when we want to change the situation we're in, then we'll need plans for what we will do, so all this applies to the ***Do***'s of our rules. For example a rule like, "***If*** you see a block, ***Do*** pick it up" leads to a complex sequence of acts: before you begin to lift a block, you need to form an action-plan to direct your shoulder, arm, and hand to do this without upsetting the objects surrounding that block). So again, one needs high-level processes, and making these plans will equally need to use multiple levels of processing—so our diagram must become something like this:

Higher-Level Processes	
Scene Describers ←→	Action Planners
Scene-Analyzers ←→	Action Scripts
Object-Finders ←→	Motor Goals
Region-Finders ←→	Motor Skills
Feature Finders ←→	Motion-Scripts
Image Filters ←→	Muscle Control
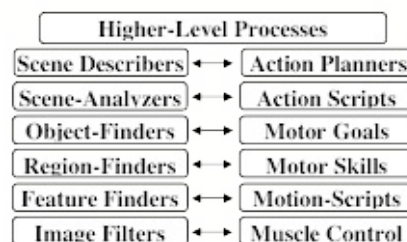
Each *Action Planner* reacts to a scene by composing a sequence of *Motion-Goals*, which in turn will execute *Motor Skills* like *'reach for,' 'grasp,' 'lift up,'* and then *'move'*. Each *Motor-Skill* is a specialist at controlling how certain muscles and joints will move—so what started out as a simple Reaction-Machine turned into a large and complex system in which each ***If*** and ***Do*** involves multiple steps and the processes at every stage exchange signals from both below and above.

In earlier times the most common view was that our visual systems work from "bottom to top," first by discerning the low-level features of scenes, then assembling them into regions and shapes, and finally recognizing the objects. However, in recent years it has become clear that our highest-level expectations affect what happens in the "earliest" stages.

V.S. Ramachandran: "[Most old theories of perception] are based on a now largely discredited "bucket brigade" model of vision, the sequential hierarchical model that ascribes our esthetic response only to the very last stage—the big jolt of recognition. In my view ... there are minijolts at each stage of visual segmentation before the final 'Aha'. Indeed the very act of perceptual groping for objectlike entities may be pleasurable in the same way a jigsaw puzzle is. Art, in other words, is visual foreplay before the final climax of recognition."[83]

In fact, today we know that visual systems in our brains receive many more signals from the rest of the brain than signals that come in from our eyes.

Richard Gregory:"Such a major contribution of stored knowledge to perception is consistent with the recently discovered richness of downgoing pathways in brain anatomy. Some 80% of fibers to the lateral geniculate nucleus relay station come downwards from the cortex, and only about 20% from the retinas."[84]

Presumably those signals suggest which kinds of features to detect or which kinds of objects might be in sight. Thus, once you suspect that you're inside a kitchen, you will be more disposed to recognize objects as saucers or cups.

All this means that the higher levels of your brain never perceive a visual scene as just a collection of pigment spots; instead, your Scene-Describing resources must represent this block-arch

in terms (for example) like *"horizontal block on top of two upright ones."* Without the use of such 'high-level' **If**s, reaction-rules would rarely be practical.

Accordingly, for *Builder* to use sensory evidence, it needed some knowledge of what that data might possibly mean, so we provided *Builder* with representations of the shapes of the objects that it was to face. Then, from assuming that something was made of rectangular blocks, one of those programs could frequently 'figure out' just which blocks appeared in a scene, based only on seeing its silhouette! It did this by making a series of guesses like these:
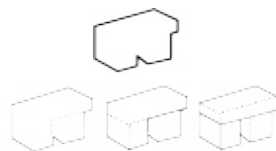
Once that program discerns a few of those edges, it imagines more parts of the blocks they belong to, and then uses those guesses to search for more clues, moving up and down among those stages. The program was frequently better at this than were the researchers who programmed it.[85]

We also gave Builder additional knowledge about the most usual 'meanings' of corners and edges. For example, if the program found edges like these

⊏ ⊐

then it could guess that they all might belong to a single block; then the program would try to find an object that might be hiding the rest of those edges.[86]

Our low-level systems see patches and fragments, but then we use 'context' to guess what they mean—and then confirm those conjectures by using several levels and types of intermediate processes. In other words, we 'recognize' things by being 're-minded' of familiar objects that could match fragments of incomplete evidence. But we still do not know enough about how our high-level expectations affect which features our low-level systems detect. For example, why don't we see the middle figure below as having the same shape as its neighbors?

In an excellent survey of this subject, Zenon Pylyshyn describes several theories about such things, but concludes that we still have a great deal to learn.
[87]

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§5-8. The Concept of a "Simulus"

Reality leaves a lot to the imagination.
—John Lennon.

All of us can recognize an arch composed of rectangular blocks.
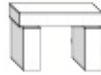
But also, we all can imagine how it would look if its top were replaced by a three-sided block.

How could a program or mind 'imagine' things that are not present in the scene? We could do this by 'envisioning' a change at any perceptual stage!

Making changes at very low levels*:* In principle, we could make a new image by changing each spot of the retinal picture—but in practice, such changes would need huge computations. Also, if you wanted to shift your point of view, you'd have to compute the whole image again. Worse, before you could do such a computation, some part of your mind would first have to know precisely what that picture describes. But to do this you'd already have to represent this at some higher descriptive level—but then, why do all that calculation?

Making changes at intermediate stages: One could change, not the picture itself, but parts of higher-level descriptions. For example, at the level of Region-Finders one could change the name of that top block's *front face* from "rectangle" to "triangle." However, this would cause trouble at other levels, because that triangle's edges would not have the proper relations to the edges of regions that neighbor on it.

Below we'll see that it would be better to replace the whole block at the higher Object-Finder level.

Visualizer: *I sometimes have troubles like that in my mind. When I try to imagine a triangular shape, I know where its three lines should appear, but I 'see' them as nebulous, glowing, streaks whose ill-defined ends many not properly meet. When I try to correct this by 'pushing' a line, it abruptly moves with some constant speed that I cannot change—and when I tell that line to stop, it tends to keep moving anyway (though, strangely, it never gets far away).*

That person was trying to change descriptions but had trouble maintaining the proper relationships between their parts. Imagining is somewhat like seeing, except that when we alter internal representations, they may not maintain their consistency. A real object can't move with two different speeds at once, nor can two real lines both intersect and not meet—but imagination has no such constraints. Of the clear affront

Making changes at the higher semantic levels: You could imagine replacing the top of that arch by merely changing the name of its shape, e.g., by replacing *rectangular* by *triangular* in, *"A rectangular block supported by two upright blocks."*

Now think about how efficient this is! To make such a change at the sensory level, you would have to alter thousands of 'pixels'—the items of data that make up a picture—whereas you need only to change a single word when you work at an abstract linguistic level, to represent an entire thing by one or only a few compact symbols. Of course, those symbols are useless unless each one is connected to structures that give more details or 'meanings' to them.

Our *Builder* system could do such things by making changes in what we call "Semantic Networks." For example, it could represent a three-block Arch by describing relations between three blocks.[88] Then, to 'imagine' a triangular block on the top, *Builder* needs only to change a single link:
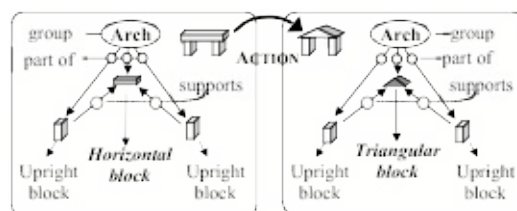
To make such changes at earlier stages would involve too many details. If you only recorded a picture-like image, then it would be hard to change any part. But at the higher 'semantic' levels, you can more easily make more meaningful ways changes. For example, when you describe *"a lying-down block supported by two upright blocks,"* you need not specify the viewer's perspective, or even say which parts of the scene are in view. Consequently that same description applies equally well to all these views:

If we substitute 'object' for the word 'block,' then our network would describe yet more situations, including these:

This shows how convenient are 'abstract' descriptions. Sometimes the word 'abstract' is used to mean 'intellectually difficult'—but here it has almost the opposite sense: abstract descriptions are simpler when they suppress details that are not relevant. Of course, descriptions must not be too abstract: as when you ask someone for advice, and they give you a useless reply like, *"If you want something, do what will get it for you."*

We've discussed how we might imagine visual scenes by constructing "simuli" inside our minds. We do similar things in other realms. Perhaps some chefs imagine new textures and tastes by changing their lower-level sensory states—and perhaps some composers imagine the sounds of new kinds of instrumentations—but such thinkers might also achieve such effects by making smaller changes at higher levels of representation, and thus evoke delight or disgust without constructing low-level details of those imagined musics or meals.

Drama Critic: I can clearly recollect how I felt after attending a certain performance, but I can't remember any details at all of what that dreadful play was about.

To discuss this, we'll coin a new word by combining "simulate" and "stimulus." A *simulus* is *a counterfeit perception caused by changing a mental representation*. Thus in the Challenger scene of §4-7, we saw how a simulus of defeat could be used to evoke a feeling of *Anger*. To do this, it might suffice to represent no more than a sneer on one's enemy's face—with no other features of that face—for one can get by with the simplest kinds of descriptions by using the highest level abstractions.

Visualizer: When I think about my cat, its image is filled with so many details

that I can visualize every hair. Would there not be a major advantage to making a real, pictorial image.[89]

Perhaps when you first imagine that cat, its surface has only a *'furry texture'*—and only when you 'zoom in' on it do you add more details to your mental representation. However, this could happen so quickly that you have no sense of it happening, and then it may seem to you as though you saw all those details at once. This could be an example of the illusion we mentioned in chapter §4:

The Immanence Illusion: *When your questions get answered before you have asked them, it will seem that you're already aware of those answers.*

The Immanence Illusion applies not only to scenes that we imagine; we never see real scenes 'all at once', either, because we don't perceive most fine details until some parts of our minds make requests for them. Indeed, recent experiments suggest that our inner descriptions of visual scenes are rarely updated in real time.[90] Chapters §6 and §8 will describe a scheme called "Panalogy" which might help to explain how our brains get such answers so rapidly.

<div align="center">∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞</div>

§5-9. Prediction Machines

> *William James: Try to feel as if you were crooking your finger, whilst keeping it straight. In a minute it will fairly tingle with the imaginary change of position; yet it will not sensibly move, because 'it is not really moving' is also a part of what you have in mind. Drop this idea, think of the movement purely and simply, with all brakes off; and, presto! It takes place with no effort at all.*
>
> *[The Principles of Psychology, 1890, p527.]*

Everyone can think about things, without performing actions—as when Carol imagined moving those blocks. But how did she manage to do that? You, yourself could now close your eyes, lean back in your chair, and indulge in your own dreams and fantasies, reflect upon your motives and goals, or try to predict what will happen next.

Now, here is how we could make a machine that does that same sort of thing, by predicting the outcomes of various actions. Let's assume that it has some rules like these.

Then we'll give our machine—let's call it *Seer*—a way to replace what it currently sees by the prediction described by this rule. Then when *Seer* is in situation A, and then considers doing action X, this will cause *Seer* then to 'imagine' that it is now in a situation like B.
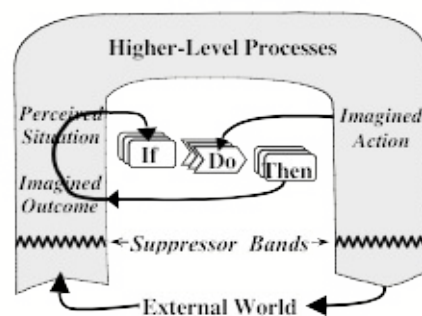
I included that pair of *"Suppressor Bands"* for two separate reasons. First, when *Seer* imagines that future condition B, we do not want this to be quickly replaced by a description of the actual, present condition A. Second, we do not yet want *Seer* to perform action X, because it might want to consider some other options before it makes a final decision. So, *Seer* can use those suppressor bands to detach itself from the outside world—and this enables it to "stop and think" before it decides which action to take.[91]

By repeating this kind of operation, *Seer* could use such prediction-chains to simulate what happens in 'virtual worlds.' Of course, for *Seer* to be able to make such predictions it must be able to use the kinds of search we described in §5-3 to simulate (and then compare) the effects of difference courses of action before deciding which one to adopt. This will need additional memory, as well as other

kinds of machinery. Still, anyone who has played a modern computer game can see how advanced has become the art of building virtual worlds inside machines.

I expect that in the next few years, we'll discover structures like those in this diagram in various parts of human brains. How did our brains evolve these abilities? The species of primates that preceded us must have had some structures like these, which they could think several steps ahead. But then, a few million years ago, that system appears to have rapidly grown, as the frontal lobes of our brains developed their present great size and complexity—and this must have been a crucial step toward the growth of our human intelligence.

Summary

This chapter described some structures and processes that might do some of the things that people do. We outlined a sequence of levels at which we can use increasingly ways to think

> Values, Censors, Ideals and Taboos
>
> Self-Conscious Reflection
> Self-Reflective Thinking
> Reflective Thinking
> Deliberative Thinking
> Learned Reactions
> Instinctive Reactions
>
> Innate, Instinctive Urges and Drives
>
> **"Model Six"**
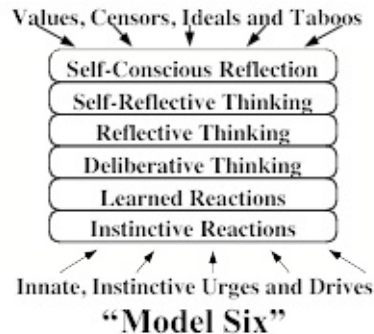
However, we have suggested rather few details about what happens at each of those levels. Later I will suggest that our systems mainly work, at each of those various cognitive levels, by constantly reacting to the particular kind of troubles they meet—by switching to more appropriate Ways to Think. We'll represent this *Model of Mind* by using this simple diagram:

> Critics   Selectors
>
> Recognize a   Activate a
> Problem-Type   Way to Think
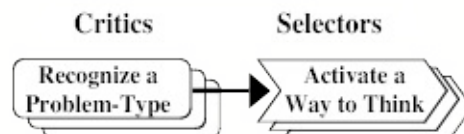
The "Critic-Selector" Model of Mind

In the rest of this book we will frequently switch between these two different views of the mind—because each one gives better answers to different kinds of questions about ourselves. Model Six makes better distinctions between various levels of mental behaviors, whereas the Critic-Selector view suggests better ideas about how to deal with difficult problems. Chapter §7 will combine both views, because we frequently use different Selectors and Critics at each of those various cognitive levels.

Values, censors, ideals and taboos

C R I T I C S			S E L E C T O R S
	Self-Conscious	Emotions	
	Self-Reflective	Thinking	
	Reflective	Thinking	
	Deliberative	Thinking	
	Learned	Reactions	
	Instinctive	Reactions	

Innate, Instinctive Behavioral Systems
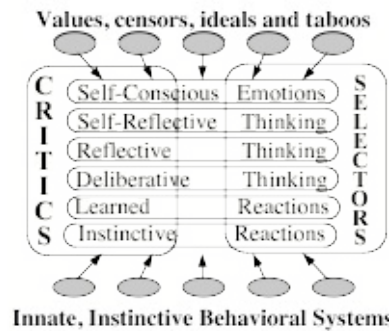
However, no matter how such a system is built, it will never seem very resourceful until it knows a great deal about the world it is in. In particular, it must be able to foresee some of the outcomes of possible actions, and it won't be able to do this unless until it possesses the right kinds of knowledge. For human beings, that's what we mean by "commonsense" knowledge and reasoning. And although, in everyday that phrase means, *'the things that most people find obvious,'* the following chapter will demonstrate that this subject is surprisingly complex.

Part VI. Common sense

> *"The way to make money is to buy stock at a low price, then when the price goes up, sell it. If the price doesn't go up, don't buy it."*
> —Will Rogers.

Soon after the first computers appeared, their actions became the subjects of jokes. The tiniest errors in programming them could wipe out their clients' bank accounts, credit them with outlandish amounts, or trap the computers in circular loops that kept repeating the same mistakes. This maddening lack of common sense led most observers to suspect that machines could never have genuine minds.

Today many programs do outstanding jobs more efficiently and reliably. Some of them can beat people at chess. Others can diagnose heart attacks. Yet others can recognize pictures of faces, assemble cars in factories, or even pilot planes or ships. But no machine yet can read a book, clean a house, or baby-sit.

Then why cannot our computers yet do so many things that people can do? Do they need more memory, speed, or complexity? Do they use the wrong kinds of instruction-sets? Do their limitations come from the fact that they only use zeros and ones? Or do machines lack some magical attribute that only a human brain can possess? This chapter will try to show, instead, that we don't need to look for excuses like these, because most deficiencies of today's machines stem from the limited ways we've been programming them.

One of these limitations is that we usually give a present-day program only the knowledge we think it will need to solve each particular problem. In contrast, every normal child learns millions of fragments of knowledge and skills that people regard as 'obvious.' For example, if you heard that someone tied a package with 'string' you might connect that word with ideas like these:

You can use a string to pull, but not push.
But you cannot push a thing with a string.
Loose strings tend to get tangled up.
Fill your package before you tie it up.
A string will break when pulled too tight.

The first parts of this chapter will discuss the need for very large bodies of commonsense knowledge, as well as the kinds of skills we need for retrieving

and applying such knowledge.

The middle parts of this chapter explore another cause for the weakness of present-day programs: they specify what the computer should do—without telling it which goals to achieve, or the intentions of those who programmed it. This means that they have no ways to reflect on whether those goals were achieved at all—or, if they were, at what cost and how well. Furthermore, those computers will still lack resourcefulness, even with access to great stores of knowledge because few fragments of knowledge are of use by themselves, unless they are also connected to reasons or goals for using them.

If you break something, you should replace it. (Because its owner wants it intact.)

People usually go indoors when it rains. (Because they do not like to get wet.)

It is hard to stay awake when you're bored. (Why would one want to stay awake?)

People don't like to be interrupted. (Because they want you to hear what they say.)

It is hard to hear in a noisy place. (You might want to hear what others say.)

No one else can tell what you're thinking. (Why might you value that privacy?)

Another deficiency is that a typical program will simply give up when it lacks some knowledge it needs—whereas a person can find other ways to proceed. So the final parts of this chapter discuss some of the tactics that people can use when we don't already know just what to do— for example, by making useful analogies.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§6-1. What do we mean by Common Sense?

"Common sense is the collection of prejudices acquired by age eighteen."
—Albert Einstein

Instead of blaming machines for their deficiencies, we should try to endow them with more of the knowledge that most people have. This should include not only what we call "commonsense knowledge"—the kinds of facts and theories that most of us know— but also the commonsense kinds of reasoning skills that we accumulate for applying that knowledge.

Student: *Can you more precisely define what you mean by 'commonsense knowledge'?*

We each use terms like 'common sense' for the things that we expect other people to know and regard as obvious. So it has different meanings for each of us.

Sociologist: *What people regard as obvious depends on their communities. Each person lives in several of these—such as family, neighborhood, language, clan, nation, religion, school, and profession—and each of these 'clubs' shares different collections of knowledge, beliefs and ways to think.*

Child Psychologist: *Still, even if you know only a child's age, you can say much about what that child is likely to know. Researchers like Jean Piaget have studied children all over the world and found that their minds grow in similar ways.*

Citizen: *We sometimes say people lack 'common sense' when they do things that seem foolish to us—not because they are lacking in knowledge, but that they're not using it properly.*

We are constantly learning, not only new facts, but also new kinds of ways to think. We learn some from our private experience, some from the teaching of parents and friends, and some from other people we meet. All this makes it hard to distinguish between what each person happens to know and what others regard as obvious. So, what each person knows (and their ways to apply it) may differ so much that we can't always predict how others will think.

We tend to take commonsense thinking for granted, because we do not often recognize how intricate those processes are. Many things that everyone does are

more complex than are many of those 'expert' skills that attract more attention and respect.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

The Telephone Call

> *You cannot think about thinking without thinking*
> *about thinking about something."*
> *—Seymour Papert*

We'll start by following Papert's advice—by thinking about some ways to think about this typical commonplace incident:

"Joan heard a ring, so she picked up her phone. Charles was answering a question she asked about a particular chemical process. He advised her to read a certain book, which he will shortly bring to her, since he will be in her neighborhood. Joan thanked him and ended the call. Soon Charles arrived and gave her the book."

Each phrase of that story evokes in your mind some of these kinds of understandings:

Joan heard a ring. She recognizes that this special sound means that someone wishes to speak with her.

She picked up the phone. Compelled to respond, she crosses the room and moves the receiver to her ear.

Charles was answering a question she asked. Charles is in a different room. They both know how to use telephones.

He advised her to read a certain book. Joan understands what Charles has said.

Joan thanked him. Was that just a formality or was she genuinely grateful to him?

He'll soon be in her neighborhood. Joan won't be surprised when he arrives.

He gave her the book. We don't know if this was a loan or a gift.

We draw these conclusions so fluently that we don't even know that we're doing it. So let's try to examine how much is involved when one understands what happened when Joan heard that sound and picked up that phone.

First, when Joan looks at her telephone, she sees only a single side of it, yet she feels that she's seen the entire thing, And even before she reaches for it, she anticipates how it will fit in her grasp, and how it will feel when it contacts her ear, and knows that one speaks into *here* and hears answers from *there*. She

knows that if she dials a number, some other phone will ring somewhere else—and if anyone happens to answer it, then those two persons can start to converse.

All this rapid retrieval of knowledge seems a natural aspect of seeing an object—and yet, one has only detected some patches of light! How does such scanty evidence make it seem as though what you're 'looking at' has been transported right into your mind—where you can move it and touch it and turn it around—or even open it up and look inside? The answer, of course, is that what you 'see' does not come from your vision alone, but also from how those visual clues lead you to retrieve other knowledge.

However, on the other side, you know so much about such things that, surely, your mind would be overwhelmed if you had to 'attend' to all that knowledge at once. So our next few sections will be discuss how brains might interconnect fragments of knowledge so that we can often retrieve just the ones that we need.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

The concept of a 'Panalogy'

> *"If you pluck an isolated sentence from a book, it will likely lose some or all of its meaning—i.e., if you show it out of context to someone else, they will likely miss some or all of its intended significance. Thus, much of the meaning of a represented piece of information derives from the context in which the information is encoded and decoded. This can be a tremendous advantage. To the extent that the two thinking beings are sharing a common rich context, they may utilize terse signals to communicate complex thoughts."*
> —Douglas Lenat

Every word, event, idea, or thing can have many different meanings to us. When you hear, *"Charles gave Joan the book,"* that might make you think of that *book* as a physical object, or as a possession or possible gift. And you could interpret this 'giving act' in at least these three different realms of thought:

The Physical Realm: Here 'give' refers to the book's motion through space, as it moves from Charles' hand to Joan's.

Physical "Trans"

before ──────▶ after

Charles holds book   Joan holds book

The Social Realm: One might wonder about Charles' motivation. Was he just being generous, or hoping to ingratiate himself?

Social "Trans"

before ──────▶ after

Joan is free   Joan is obligated

The Dominion Realm: We may infer that Joan is not only holding that book, but also has gained permission to use it.

Dominion "Trans"

before ──────▶ after

Charles controls book   Joan controls book

That "Dominion" realm" is important because we need it to achieve our goals. You cannot solve problems or carry out plans without adequate tools, supplies, and materials—but most of the things in our civilized world are controlled by persons or organizations that won't allow you to use those things until you get permission to do so. So gaining control or authority is often a requirement for (or an obstacle to) achieving some other goal.

Similarly, when two children are playing together with blocks, each may have concerns in many different mental realms:

Physical*: What if I pulled out that bottom block?*
Social*: Should I help him with his tower or knock it down?*
Emotional: *How would he react to that?*
Mental*: I forgot where I left the arch-shaped block.*
Bodily*: Can I reach that arch-shaped block from here?*
Visual*: Is the long flat block hidden behind that stack?*
Tactile*: What would it feel like to grab three blocks at once?*[92]
Spatial*: Could I arrange them into the shape of a table?*

What happens when you select an inappropriate realm of thought, in which to interpret a certain event? Then almost instantly after you notice this, you switch to a relevant point of view, without any sense of starting over again. How do we do this so rapidly? In §8-3 I will argue that our brains may use special machinery that links corresponding aspects of each view to the same 'role' or 'slot' in a larger-scale structure that is shared across several different realms. We'll call such a structure a *"Panalogy"* (for *Parallel Analogy*) and will discuss this more in §8-3.[93]
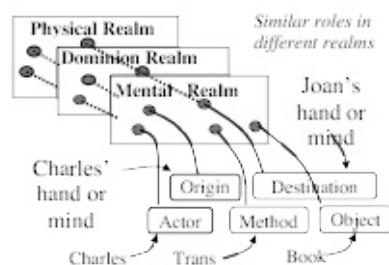
Three meanings of *'give.'*

We again see how a thing or idea can be viewed as having multiple meanings. We sometimes call these 'ambiguities' and regard them as defects in how we express or communicate. However, when these are linked into panalogies, then when situations don't seem to make sense, we can think about them in alternative realms—without the need to start over again. Shortly, we'll outline a similar scheme to explain how our vision could work so speedily.

Student: You're suggesting that we use the same techniques to represent transportation in space, for transferring an ownership, and for transmitting *knowledge to other brains. But what could have led our minds to treat such different ideas in such similar ways?*

It surely is no accident that our language uses the same prefix *'trans'* in *transfer, transport, transmit, translate, transpose, etc.,* —because that common word-part *'trans'* induces us to make many widely useful analogies.[94] Each of us knows thousands of words, and as each time we learn how others use one of them, we inherit another panalogy.

Student: How many different realms can a person use for any particular concept or object? How many of them can we handle at once? How does one know when it's time to switch? To what extent do different persons partition their worlds into similar realms?

More research on semantics will eventually clarify questions like these, but the following sections will only discuss a handful of realms that are familiar to everyone.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

Sub-Realms of the Telephone World

We've mentioned only a few of the things that every telephone user knows. However, to *use* what you know about telephones, you also have to know how to speak, and to understand some of what you may hear. You also need a good deal of knowledge about what people are and how they think, so that you can turn their interests toward the subjects that you want to discuss. Consider how many different knowledge-realms we engage to understand the story about Joan's telephone call.

The Spatial Realm: *Joan is close to her telephone, but Charles must be in*

some more distant place.

The Dominion Realm: *Both Joan and Charles have telephones, and Charles has dominion over that book. But we can't be quite certain of which ones they own.*

The Procedural Realm: *How does one make a telephone call? We could represent this in terms of a script in which certain actions are specified, but others require you to improvise.*

○ Find telephone number.
○ Locate the telephone.
○ Pick up phone. Wait for tone.
○ Dial number. Wait for ring.
○ Wait for someone to answer.
○ Initial salutation, e.g., "Hello."
○ Specific Discussion
○ Terminating salutation.
○ Hang up phone.

First, you must find the phone and dial a number. Then, once the connection has been established, you're supposed to begin with some pleasantries. Eventually, you should say why you called—and then depart from the typical script. At the end you'll close the conversation by saying "goodbye" and 'hanging up'. Generally, such behavioral scripts begin and end with conventional steps, with improvisations in between. However, you'll have to depart from the script if something goes wrong—and know how to deal with a wrong connection, or what to do if there is no answer, or if you hear the whine of a modem—or if there is too much noise on the line.

The Social Realm: When that **telephone** rings from across the room, Joan will have to walk over to get it; she knows it will do no good to ask, *"Telephone, would you please come here!"* To make an inanimate object move, you have to push, pull, or carry it. But if you want a *person* to move, those actions would be considered rude; instead you're expected to make a request. It takes our children quite a few years to learn enough such social rules.

The Economic Realm: Every action incurs some cost—not only in materials, time, and energy, but also by closing off alternatives that might bring different benefits. This raises questions about how much effort and time one should spend at comparing the costs of those options. I suspect that there's no simple answer to that, because it depends so much on the present state of the rest of one's mind. [See §§*Free Will*.]

The Conversational Language Realm*: Most people are experts at dialog, but consider how complex are the skills involved in a typical verbal exchange. You must constantly keep track of the topic, your goal, and your social role. To maintain the respect of your listeners, you must guess what they already know and remember what has already been said—so that you won't be too repetitive. It is annoying to be told things one already knows, like *"People can't see the backs of their heads,"* so your conversation must partly be based on your models of what your listeners know about the subjects that are being discussed.

You can communicate your apprehensions and hopes—or try to disguise your intentions; you know that every expressive selection can strengthen or weaken social bonds; each phrase can persuade or intimidate, conciliate or irritate, or ingratiate or drive away. You also need to keep searching for clues about how well they have understood about what you've said—and why you were trying to tell them those things.

Humanist: Speaking over a telephone is a poor substitute for a face-to-face exchange. The telephone lacks the 'personal touch' through which your gestures can put the others at ease, or express the strength of your feelings.

One always loses some nuances when conversing with a person at some other location. On the other side, we're not always aware of the misconceptions that result from what we call 'face-to-face' interactions. What if the stranger that you have just met should resemble (in manner or facial appearance) some trusted friend or some enemy? If that person reminds you of some old Imprimer, this can arouse a misplaced affection or unjustified sense of intimidation. You may think you can later correct such mistakes—but one can never completely erase the 'first impression' that one makes.

We also all share many abilities that we don't usually call 'commonsensical"—such as the kinds of physical skills that Joan uses to answer that telephone call:

The Sensory and Motor Realms*: It takes less than a single second for you to reach out your arm and *"Pick up the phone"* —yet consider how many sub-goals this involves:

Determine the telephone's location.
Determine its shape and orientation.
Plan to move your hand to its place.
Plan how your hand will grasp its shape.
Plan to transport it toward your face.

Each step of that script raises questions about how we do those things so

quickly. We can program computers to do such things, but we do not know how we do them ourselves. It is often supposed that such actions are done under continuous 'feedback control'—by processes that keep working to reduce your distance from your goal. However, that cannot be generally true because human reactions are so slow that it takes about one-fifth of a second to react to events that one did not expect. This means that *you cannot change what you are doing right now;* all you can do is revise the plan that you've made for what you will do after that. Thus when Joan reaches out to answer that call, she must plan to reduce the speed of her hand—and to already have reshaped her hand—before it collides with that telephone. Without good plans for what will happen next, she'd be constantly having accidents.

Kinesthetic, Tactile, and Haptic Realms*:* When you squeeze your phone between shoulder and cheek, you anticipate its texture and weight, adjust your grip so that it won't slip, and expect those pressures to disappear as soon as you release it. You already know that this object will fall if released from your grasp, or will break when subjected to too large a stress. An immense amount of such knowledge is stored in your spinal cord, cerebellum, and brain—but those systems are so inaccessible that we can scarcely begin to think about them.

Cognitive Realms: We are almost equally inept at describing the systems we use when we think. For example, we are almost completely unaware of how we retrieve and combine the various fragments of knowledge we need—or of how we deal with the risks of being wrong when these involve uncertainties.

The Self-Knowledge Realm*:* Whatever you may be trying to do, you'll need models of your own abilities. Otherwise, you'll set goals that you'll never achieve, make elaborate plans that you won't carry out, or too frequently switch between interests—because, as we'll see in *§9 Self,* it is hard to achieve any difficult goals unless one can make oneself persist at them.

It would be easy to add to this list of realms, but hard to construct clear distinctions between them.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§6-2. Commonsense Knowledge and Reasoning

> *Robertson Davies: You like the mind to be a neat machine equipped to work efficiently, if narrowly, and with no extra bits or useless parts. I like the mind to be a dustbin of scraps of brilliant fabric, odd gems, worthless but fascinating curiosities, tinsel, quaint bits of carving, and a reasonable amount of healthy dirt. Shake the machine and it goes out of order; shake the dustbin and it adjusts itself beautifully to its new position.*[95]

> *Albert Einstein: A little knowledge is a dangerous thing. So is a lot.*

I once encountered a fellow professor who was returning from teaching a class, and I asked him how the lecture went. The reply was that it had not gone well because *"I couldn't remember which concepts were hard."* This suggests that, over time, such experts convert some of their high-level skills into lower-level script-like processes that leave so few traces in memory that those expects cannot explain how they actually do those things. This has led many thinkers to classify knowledge into two kinds:

Knowing What. *These are the kinds of 'declarative' or 'explicit' knowledge that we can express in gestures or words.*

Knowing How. *These are the kinds of 'procedural' or 'tacit' skills (like walking or imagining) that we find very hard to describe.*

However, this popular distinction doesn't describe the functions of those types of knowledge. Instead, for example, we might classify it in terms of the kinds of thinking that we might apply to it:

Positive Expertise: *Knowing the situations in which to apply a particular fragment of knowledge.*

Negative Expertise: *Knowing which actions not to take, because they might make a situation worse.*[96]

Debugging Skills: *Knowing other ways to proceed when our usual methods fail.*

Adaptive Skills: *Knowing how to adapt old knowledge to new situations.*

The first large-scale attempt to catalog commonsense knowledge was the "CYC" project of Douglas Lenat, which started in 1984, and is described at ^{www.cyc.com}. Many ideas in this section were inspired by the results of that project.

Douglas Lenat: *"In modern America, this encompasses recent history and current affairs, everyday physics, 'household' chemistry, famous books and movies and songs and ads, famous people, nutrition, addition, weather, etc. … [It also includes] many "rules of thumb" largely derived from shared experiences—such as dating, driving, dining, daydreaming, etc., —and human cognitive economics (misremembering, misunderstanding, etc.,), and shared modes of reasoning both high (induction, intuition, inspiration, incubation) and low (deductive reasoning, dialectic argument, superficial analogy, pigeon-holing, etc.)."*

Here Lenat describes some kinds of knowledge that a simple statement like this might engage:[97]

"Fred told the waiter he wanted some chips."
The word "he" means Fred—and not the waiter. This event took place in a restaurant. Fred was a customer dining there. Fred and the waiter were a few feet apart. The waiter was at work there, waiting on Fred at that time.

Fred wants potato chips, not wood chips—but he does not want some particular set of chips.

Both Fred and the waiter are live human beings. Fred accomplished this by speaking words to the waiter. Both of them speak the same language. Both were old enough to talk, and the waiter was old enough to work.

Fred is hungry. He wants and expects that in a few minutes the waiter will bring him a typical portion—which Fred will start eating soon after he gets them.

We can also assume that Fred assumes that the waiter also assumes all those things.

Here is another example of how much one must know to give meaning to a commonplace statement:

"Joe's daughter was sick so he called the doctor."
We can assume that Joe cares about his daughter, is upset because she is sick, and wants her to be healthy. Presumably he believes she is sick because of observing some symptoms.

People have different abilities. Joe himself cannot help his daughter. People ask others for help to do things they can't do themselves. So Joe called the

doctor to help heal his daughter.

Joe's daughter, in some sense, belongs to Joe. People care more about their own daughters than about other people's daughters. If so advised, Joe will take the daughter to the doctor. When at the doctor's, she will still belong to Joe.

Medical services can be expensive, but Joe is likely to forgo other spending to get the doctor to help the daughter.

These are all things that 'everyone knows' and uses to understand everyday stories. But along with that widely shared, common knowledge, every person also has personal knowledge; we each know our own private histories, characteristics of our acquaintances, special perceptual and motor skills, and other kinds of expertise.

Still, none of our knowledge would have any use unless we also had effective ways to apply that knowledge to solving problems. This means that we also need large bodies of skills for doing what we call commonsense thinking. We'll come back to that in chapter §7.

<div align="center">∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞</div>

How much does a typical person know?

Everyone knows a good deal about many objects, topics, words, and ideas—and one might suppose that a typical person knows an enormous amount. However, the following argument seems to suggest that the total extent of a person's commonsense knowledge might not be so vast. Of course, it is hard to measure this, but we can start by observing that every person knows thousands of words, and that each of those must be linked in our minds to as many as a thousand other such items. Also a typical person knows hundreds of uses and properties of thousands of different common objects. Similarly, in the social realm, one may know thousands of things about tens of people, hundreds of things about hundreds of people, and tens of useful items about as many as a thousand people.

This suggests that in each important realm, one might know perhaps a million things. But while it is easy to think of a dozen such realms, it is hard to think of a hundred of them. This suggests that a machine that does humanlike reasoning might only need a few dozen millions of items of knowledge.[98]

Citizen: Perhaps so, but I have heard of phenomenal feats of memory. What about persons with photographic memories, who can recollect all the words of a book after only a single reading of it? Could it be that we all remember, to some

extent, everything that happens to us?

We all have heard such anecdotes, but whenever we try to investigate one, we usually fail to uncover the source, or find that someone was fooled by a magic show trick. Many a person has memorized an entire book of substantial size (which most usually is a religious tract)—but no one has ever been shown to have memorized a hundred such books. Here is what one psychologist said about a person who appeared to him to possess a prodigious memory:

Alexander R.Luria: "For almost thirty years the author had an opportunity systematically to observe a man whose remarkable memory... which for all practical purposes was inexhaustible" (p3) ... It was of no consequence to him whether the series I gave him contained meaningful words or nonsense syllables, numbers or sounds; whether they were presented orally or in writing. All that he required was that there be a three-to-four-second pause between each element in the series. . ... And he could manage, also, to repeat the performance fifteen years later, from memory."[99]

This may seem remarkable, but it might not be truly exceptional, because, in 1986, Thomas Landauer concluded that, during any extended interval, none of his subjects could learn at a rate of more than about 2 bits per second, whether the realm be visual, verbal, musical, or whatever. So, if Luria's subject required four seconds per word, he was well within Landauer's estimate.[100] And even if that individual were to continue this over the course of a typical lifetime, this rate of memorization would produce no more than 4000 million bits—a database that would easily fit on the surface of a Compact Disk.

Student: I'm uncomfortable with this argument. I agree that it might apply to our higher-level kinds of knowledge. But our sensory and motor skills might be based on much larger amounts of information.

We don't have a good way to measure such things, and making such estimates raises hard questions about how those fragments of knowledge are stored and connected. Still, we have no solid evidence that any person has ever surpassed the limits that Landauer's research suggests.[101]

Chapter §7 will speculate about how we organize knowledge so that, whenever one of our processes fails, we can usually find an alternative. But here we'll change the subject to ask how we could endow a machine with the kinds of knowledge that people have.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

Could we build a Baby-Machine?

> *Alan Turing: "We cannot expect to find a good child machine at the first attempt. One must experiment with teaching one such machine and see how well it learns. One can then try another and see if it is better or worse [but] survival of the fittest is a slow method for measuring advantages. The experimenter, by the exercise of intelligence, should be able to speed it up [because] if he can trace a cause for some weakness he can probably think of the kind of mutation which will improve it."[102]*

To equip a machine with something like the knowledge we find in a typical person, we would want it to know about books and strings; about floors, ceilings, windows, and walls; about eating, sleeping, and going to work. And it wouldn't be very useful to us unless it knew about typical human ideals and goals.

Programmer: Then, why not build a 'baby-machine' that learns what it needs from experience? Equip a robot with sensors and motors, and program it so that it can learn by interacting with the real world—the way that a human infant does. It could start with simple If-Then schemes, and then later invent more elaborate ones.

This is an old and popular dream: to build a machine that starts by learning in simple ways and then later develops more powerful methods—until it becomes intelligent. In fact several actual projects have had this goal, and each such system made progress at first but eventually stopped extending itself.[103] I suspect that this usually happened because those programs failed to develop *good new ways to represent knowledge.*

Inventing good new ways to represent knowledge is a major goal in Computer science. However, even when these are discovered, they rarely are quickly and widely adopted—because one must also develop good skills to work with them efficiently. And since such skills take time to grow, you will have to make yourself tolerate periods in which your performance becomes not better, but worse.[104]

The Investment Principle: It is hard to appreciate the virtues of a new technique because, until you become proficient with it, it will not produce results

as good as you'll get from the methods that you are familiar with.

No one has yet made a baby-machine that that developed effective new kinds of representations. Chapter §10 will argue that human brains are born equipped with machinery that eventually provides them with several different ways to represent various types of knowledge.

Here is another problem with "baby-machines." It is easy to program computers to learn fairly simple new *If Then* rules; however, if a system does this too recklessly, it is likely to deteriorate from accumulating too much irrelevant information. Chapter §8 will argue that unless learning is done selectively—by making appropriate *"Credit Assignments,"* a machine will fail to learn the right things from most of its experiences.

Entrepreneur: Instead of trying to build a system that learns by itself, why not make one that searches the Web to extract knowledge from those millions of pages of content-rich text.

That certainly is a tempting idea, for the World Wide Web must contain more knowledge than any one person could comprehend. However, it does not *explicitly* include the knowledge that one would have to use to understand what all those texts mean. Consider the kind of story we find in a typical young child's reading book.

The World Wide Web contains more knowledge than any one person could ever learn. However, it does not *explicitly* display the knowledge one needs for understanding what all those texts mean. Consider the kind of story we find in a typical young child's reading book.

"Mary was invited to Jack's party. She wondered if he would like a kite. She went shook her piggy bank. It made no sound."[105]

A typical reader would assume that Jack is having a *birthday* party, that Mary is concerned because she needs to bring Jack a suitable present, that a good birthday present should be something that its recipient likes; that Jack might like to receive a kite; that Mary wants money to pay for that kite; and that the bank would have rattled if it contained coins. But because these are all things that 'everyone knows' we scarcely ever write them down, so such knowledge stays hidden 'between the lines.'[106]

Neurologist: Why not try to copy the brain, using what brain-scientists have learned about the functions of various parts of the brain.

We learn more about more such details every week—but still do not yet

know enough to simulate a spider or snake.

Programmer: What about alternatives such as building very large Neural Networks or big machines that accumulate huge libraries of statistical data?

Such systems can learn to do useful things, but I would expect them to never develop much cleverness, because they use numerical ways to represent all the knowledge they get. So, until we equip them with higher reflective levels, they won't be able to represent the concepts they'd need for understanding what those numbers might mean.

Evolutionist: If we don't know how to design better baby-machines, perhaps we can make them evolve by themselves. We could first write a program that writes other programs and then makes various kinds of mutations of them—and then making those programs compete for survival in suitably lifelike environments.

It took hundreds of million of years for us to evolve from the earliest vertebrate fish. Eventually a few of their descendants developed some higher-level systems like those we described in chapter §5; in fact most vertebrates never developed them. Generally, it is hard for complex systems to improve themselves because most specializations that lead to near-term gains are likely to make it much harder to change. We'll discuss this more in §§Duplication and Diversity.

In contrast, human brains start out equipped with systems that are destined to develop into useful ways to represent knowledge. We'll need to know more about such things before we are ready to construct efficient self-improving machines.

Architect: In this section you've been very negative. You've said that each of those methods has merit, and yet you found reasons to reject them all. But surely one could combine the virtues of all those ideas, in some way in which each offsets the others deficiencies.

Indeed, we should find ways to use them all, and we'll propose ways to do this in subsequent chapters. I would not dismiss all prospects of building a baby-machine, but only schemes for doing this by "starting from scratch"—because it seems clear that a *human* baby begins equipped with intricate ways to learn, not only to master the simplest facts, but to also construct new ways to think. If you don't agree with this, try teaching your kitten to read and write, do calculus, or dress itself.

More generally, it seems to me that all of the previous learning schemes—

statistical, genetic, and logical—have 'tapered off' by getting stuck because of not being equipped with ways to overcome problems like these:

The Optimization Paradox: *The better a system already works, the more likely each change will make it worse.* See §§Duplication.

The Investment Principle: *The better a certain process works, the more we will tend to rely on it, and the less likely we will be inclined to develop new alternatives.*

The Parallel Processing Paradox: *The more that the parts of a system interact, the more likely each change will have serious side effects.*

In other words, as a system gets better it may find that it is increasingly harder to find more ways to improve itself. Evolution is often described as selecting good changes—but it actually does far more work at rejecting changes with bad effects. This is one reason why so many species evolve to occupy narrow, specialized niches that are bounded by all sorts of hazards and traps. Humans have come to escape from this by evolving features that most animals lack—such as ways to tell their descendants about the experiences of their ancestors. *See §§Evolution.*

In any case, for a machine to keep developing, it must have ways to protect itself against changes with too many side effects. One notable way to accomplish this is to split the whole system into parts that can evolve separately. This could be why most living things evolved as assemblies of separate 'organs'—that is, of parts with fewer external connections. Then changes inside each of those organs will have fewer bad external effects. In particular this could be why the resources inside our brains tended to become *organized* into more-or-less separate centers and levels—like those suggested in §5-6.

Reactive systems *operate on descriptions of real, external situations.*
Deliberation *operates on descriptions of future reactions.*
Reflective systems *operate on descriptions of deliberations.*
Self-Reflection *operates on descriptions of reflections.*

Why emphasize descriptions here? That's because we could never learn enough low-level **If-Then** rules, and the only alternative is to use abstractions—as was argued in 1959 in an essay called *Programs with Common Sense.*[107]

John McCarthy: "If one wants a machine to discover an abstraction, it seems most likely that the machine must be able to represent this abstraction in some relatively simple way."

We need to make our descriptions abstract because no two situations are ever

the same, so as we saw in §5-2, our descriptions must not be too concrete—or they would not apply to new situations. However, as we noted in §5-3, no representation should be too abstract, or it will suppress too many details.[108]

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

Remembering

We discussed how much knowledge a person could have, but perhaps it is more important to ask how we *recollect* what we need so quickly when we need it?

Whenever we get a new idea, or find a new way to solve a problem, we may want to make a memory-record of it. But records are useless unless you have ways to retrieve the ones most likely to be relevant to the problems you face. I'll argue that this needs a lot of machinery.

Citizen: If remembering is so complex, then why does it seem so effortless, simple and natural? Each idea reminds me of similar ones, which then make me think of related ideas—until I recall the ones that I need.

Why does 'remembering' seem so effortless? As long ago as you can remember, you could always recall things that happened to you. However, you cannot remember much of your earliest years; in particular, you cannot recall how you developed your early abilities. Presumably, you had not yet developed the skills one needs for making those kinds of memories.[109]

Because of this *Amnesia of Infancy,* we all grow up with simplistic views of what memories are and how they work. You might think of your memory as like a writing-pad, on which you can jot down mental notes. Or perhaps for each significant event, you store 'it' away in some kind of memory-box and later, when you want it back, you somehow bring 'it' out of that box—if you are lucky enough to find it. But, what kinds of structures do we use to represent those 'its' and how do we bring them back when we need them? Our recollections would be useless unless (1) they are relevant to our goals and (2) we also have ways to retrieve the ones that we need at the times when we need them.

To do this, a computer expert might suggest that we store everything in some single 'data base' and use some general-purpose 'matching' technique. However, most such systems still classify things in terms of *how they are described* instead of *what they are likely to be useful for.* The trouble with this is we do not usually know what kind of thing we are looking for, but only what we want to accomplish with it—because we're facing some obstacle, and want to know how to deal with it.

So, instead of using some general method, I suspect that every child develops ways to link each new fragment of knowledge to goals that it might help us to achieve, as well as to other related ideas. These additional links might help to answer questions like these:

What kinds of goals might this item serve? *Which kinds of problems could it help to solve? What obstacles could it help to overcome?*

In which situations might it be relevant*? In which contexts is this likely to help? What subgoals must first be achieved?*

How has it been applied in the past*? What were some similar previous cases? What other records might be relevant? See §8-Credit Assignment.*

Each fragment of knowledge may also need links to some knowledge about its deficiencies—and the dangers and costs of using it:

What are its most likely side effects? Is it likely to do us more harm or more good?

How much will it cost to use it? Will it repay the effort of using it?

What are its common exceptions and bugs? In which contexts is it likely to fail us—and what might be good alternatives?

Is it part of some relevant family? [See Glossary: Ontology.]

We also link each item to information about its sources and to what other persons might know.

Was it learned from a reliable source? *Some informants may simply be wrong, while others may mean to mislead us.*

Is it likely to be outdated soon? *That's why this book does not depend much on current beliefs about how our brains work.*

Which other people are likely to know it? *Our social activities strongly depend on knowing what others may understand.*

All this raises question about how we make so many connections to and from each new fragment of knowledge. I suspect that we can't do this all at once— and indeed there is some evidence that it normally takes some hours or days (including some sessions of dream-laden sleep) to establish new long-term memories. Also, we probably add more links each time we retrieve a fragment of knowledge, because then we're likely to ask ourselves, *"How did this item help (or hinder) me at overcoming this obstacle?"* Indeed, some research in recent years suggests that our so-called long-term memories are not so permanent as we used to think; it seems that they can be altered by suggestions and other experiences.

We all know that our memory systems can fail. There are things that we can't remember at all. And sometimes we tend to recollect, not what actually happened to us, but versions that seem more plausible. At other times we fail to remember something relevant until—after several minutes or days—suddenly

the answer appears—and you say to yourself, *"How stupid of me; I knew that all along!"* (That could happen either because an existing record took long to retrieve, or because it was never actually there, and you had to construct a new idea by using some process of reasoning.)

In any case, we should expect such 'lapses' because our recollections *must* be selective; §4-4 discussed how bad it would be to remember everything all the time: it would overwhelm us to recall all the millions of things that we know. However, none of this answers the question of how we usually retrieve the knowledge that we currently need. I suspect that this is mainly due to our already having prepared in advance the sort of links discussed above. But constructing these requires additional skills; we'll discuss these in §8-5 Credit Assignment.

As the start of this section we asked about how we retrieve the knowledge we need. The following section will argue that part of the answer lies in those links to *the goals that each fragment of knowledge might help to achieve."* To make that statement more concrete, the next few sections will investigate what goals are and how they work.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§6-3. Intentions and Goals

> *"No one imagines that a symphony is supposed to improve in quality as it goes along, or that the whole object of playing it is to reach the finale. The point of music is discovered in every moment of playing and listening to it. It is the same, I feel, with the greater part of our lives, and if we are unduly absorbed in improving them we may forget altogether to live them."*
>
> *—Alan Watts.*

Sometimes we seem to act passively, just reacting to things that happen to us —but at other times we feel more in control, and feel that we're actively choosing our goals. I suspect that this most often happens when two or more goals become active at once and thereby lead to a conflict. For as we noted in §4-1, when our routine thinking runs into trouble, this engages our higher reflective levels.

For example, when angry or greedy enough, we are likely to take actions that later may make us have feelings of shame or guilt. Then we may offer such justifications as, *"that impulse became too strong to resist"* or *"I found that I did it in spite of myself."* Such excuses relate to the conflicts between our immediate goals and our higher ideals, and every society tries to teach its members to resist their urges to breach its conventions. We call this developing 'self-control' and each culture makes maxims about such feelings.

> *Moralist: No merit comes from actions based on self-serving wishes.*
> *Psychiatrist: One must learn to control one's unconscious desires.*
> *Jurist: To be guilty in the first-degree, an offense must be deliberate.*

Still, an offender can object, *"I didn't intend to do those things,"* —as though a person is not 'responsible' for an action that wasn't intentional. But, what kinds of behavior might lead you to think that a person did something "deliberately"—in contrast to it having resulted from mental processes that were not under that person's control?

To understand this, it may help to observe that we have similar thoughts about physical things; when we find that some object is hard to control, we sometimes imagine that *it* has a goal—and say, *"This puzzle-piece doesn't want*

to fit in," or *"My car seems determined not to start."* Why would we think of an object in that way, when we know that it has no such intentions?

The same thing can happen inside your mind, when one of your goals becomes so strong that it is hard to think about anything else. Then it may seem to come from no choice of your own, but is somehow being imposed upon you. But, what could make you pursue a goal that does not seem to be one that you want? This could happen when that particular goal conflicts with some of your high-level values, or when you have other goals with different aims; in any case, there is no reason to expect all of one's goals to be consistent.

However, this still does not answer the question of why a goal can seem like a physical force, as in, *"That urge became irresistible."* And indeed, a 'powerful' goal can seem to push other goals aside, and even when you try to oppose it, it may prevail if you don't fight back strongly enough. Thus both forces and goals share some features like these:

> *Both seem to aim in a certain direction.*
> *Both 'push back' when we try to deflect them.*
> *Each seems to have a 'strength,' or 'intensity'.*
> *Both tend to persist till the cause of them ends.*

For example, suppose that some external force is applied to your arm—say, strongly enough to cause some pain—and your A-Brain reacts by pushing back (or by moving away)—but, whatever you do, it keeps pressing on you. In such a case, your B-brain might see nothing more than a sequence of separate events. However, your higher reflective levels might recognize these as matching this particular pattern:

> *"Something is resisting my efforts to make it stop. I recognize this as a process which shows some persistence, aim, and resourcefulness."*

Furthermore, you might recognize a similar pattern inside your mind when some resources make choices in ways that the rest of your mind cannot control, as when you do something "in spite of yourself." Again, that pattern may seem as though some external force was imposed on you. So it often makes practical sense to represent both forces and intentions as though they were assistants or antagonists.

> *Student: But isn't it merely a metaphor, to speak of a goal as resembling a force? Surely it's bad to use the same words for things with such different characteristics.*

We should never say 'merely' for metaphors, because that is what all

descriptions are; we can rarely state just what something *is,* but can only describe what something is *like*—that is, to describe it in terms of other things we already know to have some similar properties—and then to consider the differences. Then, we label it with the same or a similar name—so that thenceforth that older word or phrase will include this additional meaning-sense. This is why most of our words are 'suitcase-words'—and later I will argue that the ambiguities of our words may be the greatest treasures that we inherit from our ancestors.

We've mentioned goals many times in this book—but never discussed how goals might work. So let us turn from the subject of how a goal feels to ask what a goal might actually be!

<center>∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞</center>

Difference-Engines

> Aristotle: *"Differences arise when what we get is different from what we desire; for it is like getting nothing at all when we do not get what we aim at."*

Sometimes people appear to behave as though they had no direction or aim. At other times they seem to have goals. But what *is* a goal, and how can we *have* one? If you try to answer such questions in everyday words like, *"a goal is thing that one wants to achieve,"* you will find yourself going in circles because, then, you must ask what *wanting* is—and then you find that you're trying to describe this in terms of other words like *motive, desire, purpose, aim, hope, aspire, yearn* and *crave.*

More generally, you get caught in this trap whenever you try to describe a state of mind in terms of other psychology-words, because these never lead to talking about the underlying machinery. However, we can break out of that with a statement like this:

A person will seem to have a goal when they keep different techniques that are likely to change their present situation into a certain other condition.

This takes us out of the psychological realm by leading us to ask about what kind of machinery could do such things. Here is one way such a process might work:

Aim: It begins with a description of a certain possible future situation. It also

can recognize some differences between the situation it now is in and that "certain other condition."

Resourcefulness: It is also equipped with some methods that may be able to reduce those particular kinds of differences.

Persistence: A process that keeps applying those methods. Then, in psychological terms, we will perceive it as trying to trying to change what it now has into what it 'wants.'

Persistence, aim, and resourcefulness! The next few sections will argue that this particular triplet of properties could explain the functions of what we call *motives* and *goals,* by giving us answers to questions like these:

> *What makes some goals strong and others weak?*
> *What are the feelings that accompany them?*
> *What could make an impulse "too strong to resist?"*
> *What makes certain goals 'active' now?*
> *What determines how long they'll persist?*

No machine had clearly displayed those three traits of *Aim, Persistence*, and *Resourcefulness*—until 1957, when Allen Newell, Clifford Shaw and Herbert Simon developed a computer program called the *"General Problem Solver."* Here is a simplified version of how it worked; we'll call this version a Difference-Engine.[110]
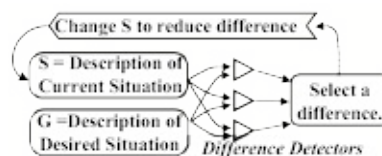
At every step, this program compares its descriptions of the present and that future situation, and this produces a list of differences between them. Then it focuses on the most serious difference and applies some technique that has been designed to reduce this particular type of difference. If this succeeds, the program then tries to reduce what *now* seems to be the most serious difference. However, whenever such a step makes things worse, the system goes back and

tries a different technique.

For example, every infant is born with such a system for maintaining 'normal' body temperature: when too hot, the baby may sweat, pant, stretch out, and/or vasodilate; when too cold, it will curl up, shiver, vasoconstrict and/or raise its metabolic rate.
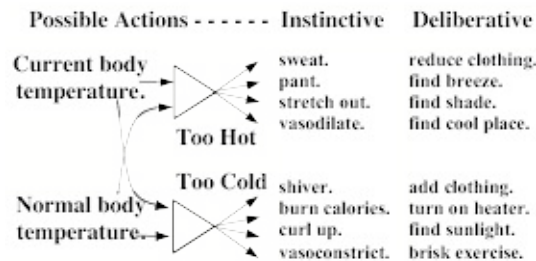
At first we may be unaware of such processes, because such instinctive reactions begin at very low cognitive levels. For example, when you become too hot, you automatically start to sweat. However, when perspiration drips, you may notice this, and deliberate: *"I must find some way to escape from this heat."* Then your acquired knowledge may suggest other actions to take, such as moving to an air-conditioned place. If you feel too cold, you might put on a sweater, turn on a stove, or begin to exercise (which can make you produce ten times as much heat).

Now we can interpret *"having a goal"* to mean that a Difference-Engine is actively working to remove those differences.

Student: To have a goal, does one really need a representation of the desired situation? Would it not be sufficient just to have a list of desired properties?

This is a matter of degree, because one could never specify every aspect of a situation. We could represent a *'desired situation'* as a simple, rough sketch of a future scene, as a list of a few of its properties, or as just some single property (for example, that it is causing some pain).

Student: It seems to me that we should distinguish between 'having a goal' and actively 'wanting' it. I would say that your difference-engine is a "wanting machine' and that the goal itself is only the part that you called its 'aim'—its current description of some future situation.'

An imagined description becomes an active goal when one is running a

process that keeps changes conditions until they fit that description. Our everyday language does not serve well for making the kinds of distinctions we need and, to remedy that, each technical field must develop its own specialized language or 'jargon.' However, it will be hard to do this for psychology until we can agree on which more detailed models of minds to use.

Romanticist: This Difference-Engine idea could account for some of what "having a goal" might mean—but it doesn't explain the joy of success, or the distress that besets us when we fail to achieve what we've hoped for.

I agree that no single meaning of *goal* can explain all of those cascades of feelings, because *wanting* is such a large suitcase of concepts that no single idea can embrace them all. Besides, many things that people do come from processes with no goals at all, or goals of which they are unaware. Nevertheless, the Difference-Engine's characteristics capture more of our everyday concept of 'goal' than any other description I've seen.

Student: What happens when that difference-engine finds several differences at once? Can it work on them all simultaneously, or must it deal with them one-by-one?

When several differences are evident, one might try to reduce several at once, perhaps by using different parts of the brain. However, Newell and Simon concluded that it is usually best to first try to remove the one that seems most significant, because this is likely to change quite a few of the others. So the *General Problem Solver* included a way to assign a different priority to each kind of difference that it could detect.

Student: Isn't that a flaw in that? What if Carol places a block in a place that prevents her from building the rest of her arch? Sometimes reducing one difference might make all the other differences worse.

That turned out to be a severe limitation, because a *Difference-Engine,* by itself, has no way to plan several steps ahead—for example, by the methods suggested in §5-5—so it cannot sustain a short-term loss for the purpose of later, larger gains. So, although their system could solve many problems, this limitation seems to have led Newell and Simon to move in other directions.[111] In my opinion, they should have persisted, because this project had so many good ideas that I find it strange that it was not further developed in later years. In any case, we can't expect any one method to solve every problem—and our forthcoming project will try to embody the concepts that Newell and Simon abandoned.

In retrospect, one could argue that the system got stuck because it was not equipped with ways to reflect on its own performance—the way that people can 'stop to think' about the methods that they have been using. However, in a great but rarely recognized essay, Newell and Simon did indeed suggest a very ingenious way to make such a system reflect on itself.[112]

On the positive side, the *General Problem Solver* was equipped with several ways to reduce each kind of difference, and it even included a place for ways to introduce new kinds of representations.

What if one fails to solve a problem, even after using reflection and planning? Then one may start to consider that this goal may not be worth the effort it needs—and this kind of frustration then can lead one to 'self-consciously' think about which goals one 'really' wants to achieve. Of course, if one elevates that level of thought too much, then one might start to ask questions like, *"Why should I have any goals at all,"* or, *"What purpose does having a purpose serve"*—the troublesome kinds of questions that our so-called "existentialists" could never found plausible answers to.

However, the obvious answer is that this is not a matter of personal choice: we have goals because that's how our brains evolved: the people without goals became extinct because they simply could not compete.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

Goals and Subgoals

> *Aristotle: We deliberate not about ends, but about means. … We assume the end and think about by what means we can attain it. If it can be produced by several means, we consider which one of them would be best …[and then] we consider by which means that one can be achieved, until we come to the first cause (which we will discover last).*[113]

Section §2-2 considered some questions about how we connect our subgoals to goals—but did not stop to investigate how those subgoals might originate. However, a Difference-Engine does this by itself because, *every difference it needs to reduce becomes another subgoal for it!* For example, if Joan is in Boston today, but wants to present a proposal in New York tomorrow, then she will have to reduce these differences:

The meeting is 200 miles away.
Her presentation is not yet complete.
She must pay for transportation, *etc.*

Walking would be impractical because that distance is too large, but Joan could drive, take a train, or an airplane. She knows this 'script' for an airplane trip:

Each phase of this script, in turn, needs several steps. She could *"Get to the airport"* by bicycle, taxi, or bus, but she decides to drive her car, which begins with a series of subgoals like these:

Leave home. Lock the door.
Enter driver's side of car.
Use a key to unlock the door.
Enter, sit down, close door.
Fasten seat belt, check fuel.
Look ahead. Start the car.

When Joan reviews that airplane trip, she decides it would waste too much of her time to park the car and pass through the security line. The actual flight from home to New York takes no more than an hour or so, and the railroad trip is four hours long, but it ends near her destination—and she could spend all that time at productive work. She 'changes her mind' to take the train.[114]

Similarly, when Carol decides to change this

into this

, she will need to split this job into parts, and that will need several subgoals and scripts.

Tower-Builder

Find place to start   Add a block   Is it tall enough?
                              No   Yes
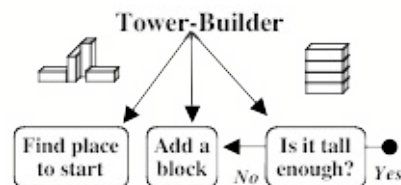
Then each of those subgoals will turn out to require several more parts and processes—and when we developed a robot to do such things, its software needed several hundred parts. For example, **Add a block** needed a branching network of subgoals like these:

```
                        ( Add a Block )
     ( Choose a block )           ( Put it on top )
                        ( Grasp it )

     ( See )  ( Grasp )  ( Move )   ( Release )
```

Each of those subgoals has problems to solve. ***Choose*** must not select a block that is already supporting the tower top. ***See*** must recognize objects regardless of color, size, and shades of light—and even when they partly obscured by other objects. ***Grasp*** must adapt the robot's hand to the perceived size and shape of the block to be moved. And ***Move*** must guide the arm and hand through paths in space that never strike the tower's top or hit the child's face.

How do we find out which subgoals we'll need to achieve before we can accomplish a job? You could discover them by trial and error, or by doing experiments inside your mind, or recalling some prior experience. But perhaps our most generally useful method is to use a Difference-Engine—because every difference that this detects could thereby become a new subgoal for us.

To summarize, our idea is that "to have an active goal" amounts to running a Difference Engine-like process. I suspect that, inside each human brain, many such processes all run at once, at various levels in various realms. These range from instinctive systems that work all the time—like those that maintain our temperatures (and these are so opaque to reflection that we don't recognize them as goals at all)— up to those highest self-conscious levels at which we keep trying try to be more like the persons we wish we were.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§6-4. A World of Differences

> *"Some minds are stronger and apter to mark the differences of things, others to mark their resemblances. The steady and acute mind can fix its contemplations and dwell and fasten on the subtlest distinctions: the lofty and discursive mind recognizes and puts together the finest and most general resemblances. Both kinds however easily err in excess, by catching the one at gradations, the other at shadows."*
>
> *—Francis Bacon, Novum Organum, 1620.*

Whenever somebody tells you a story, you react less to what each separate sentence means than to how this differs from what you expected—and this also applies to our other perceptions. For example, if you plunge your hand into a bowl of cold water, you'll feel a severe sensation of chill—but soon this will totally disappear, just as a steady pressure on your skin will quickly seem to diminish in strength. It is the same with new odors or tastes, or with the onsets of continuous sounds: at first those sensations may seem intense but then they rapidly fade away. We have many different names for this, like *accommodation, adaptation, acclimatization, habituation,* or just becoming *accustomed* to it.

Student: This doesn't apply to vision, though. I can look at an object as long as I like, and its image never fades; in fact, I keep seeing more features of it.

Physiologist: In fact, that image would rapidly fade if you could keep from moving your eyes, which normally make small motions that keep changing your retinal images.[115]

Thus most of our external sensors react only to rather rapid changes in conditions. However, we also have additional sensors that do not fade away, but keep responding to certain particular harmful conditions; see §§Alarms.

Now let's apply the same idea—of a system that 'mainly reacts to change'—to a brain with a tower of cognitive levels. This could help to explain some phenomena. For example, after you start a trip on a train, you're aware of the clacking of wheels on the track—but (if that clacking is regular) then you will soon stop noticing this. Perhaps your A-Brain is still processing it, but your B-brain has stopped reacting to it. It will be much the same for the visual scenes;

when the train enters a forest, you'll start seeing trees—but soon you'll start to ignore them. What could cause such meanings to fade?

It's much the same with repeated words; if someone says 'rabbit' one hundred times, while trying to focus on what that word means, then that meaning will shortly disappear—or be replaced by some other one. And similarly the same thing happens when you listen to popular music: you'll often hear dozens of nearly identical measures, but the details of these soon fade away and you no longer pay any attention to them. Why don't we object to that repetitiousness?

This could be partly because we tend to interpret such 'narratives' in terms of how situations change on successively larger scales of time. In the case of most music, this structure is clear: we begin by grouping separate notes into 'measures' of equal length, and we then group these into larger sections, until the whole composition is seen as a storylike structure.[116] We do this in vision and language, too—although with less repetitiousness— by grouping collections of smaller events into multiple levels of events, incidents, episodes, sections, and plots. However, we see most clearly in musical forms:

Feature-Detectors recognize pauses, notes, and various other aspects of sounds, such as harmony, tempo, and timbre, *etc.*

Measure-Takers group these into chunks. In music, composers make this easy for us by using measures of equal length; this makes it extremely easy to sense the differences between successive chunks.

Phrase-and Theme-Detectors then represent larger events and relationships like, *'This theme goes down and then goes up, and ends with three short, separate notes.'*

Then *Section-Builders* group these into larger-scale parts, such as, *'these three similar episodes form a sequence that rises in pitch.'*[117]

Finally, our *Storytellers* interpret these as similar to events in other realms—

such as depicting a journey through space and time, or a skirmish among personalities. One special appeal of music is how effectively it can depict what we might call *abstract emotional scripts*—stories that seem to be about entities about whom we know nothing at all except that we can recognize their individual characteristics—e.g., *this one is warm and affectionate,* whereas *that one is cold and insensitive.* Then we empathize with how they feel as we interpret those phrases and themes as representing mental conditions like conflict, adventure, surprise, and dismay—as in, *those horns are attacking the clarinets, but the strings are now trying to calm them down.*

Now suppose that each higher level in the brain mainly reacts to the changes below it, but over some larger scale of time. If so, then when signals repeat at level A, the B-Brain will have nothing to say. And if the signals that go up to B form a sequence that repeats—so that the B-brain keeps seeing a similar pattern —then the C-Brain will sense a 'constant condition,' and thus have nothing to say to the level above it.

This could explain some common experiences because any repetitive signal would tend to partly *'anesthetize'* the next level above it. So although your foot may continue to tap, most details of those smaller events won't go up.

(Why might our brains have evolved to work this way? If some condition has been present for long—and nothing bad has happened to you—then it probably poses no danger to you; then so you might as well not pay attention to it and apply your resources more gainfully.)

However, this could also lead to other effects. Once a level gets freed from control by repetitive signals that come from below it, then it could start to *send signals down* to instruct those levels to try to detect different kinds of evidence. For example, during that railroad trip, perhaps you first heard those clacks on the tracks as forming a pattern of *clack-clack-clack-clacks*—that is, of beats in 4:4 time. Then you stopped hearing them at all—but soon you may have suddenly switched to hearing groups of *'clack-clack-clacks'*—that is, of beats in 3:4 time. What made you change your representation? Perhaps some higher level just switched to forming a different hypothesis.

Also, when repetitive signals anesthetize some parts of your brain, this could release some other resources to think in new, unusual ways. This could be why some types of meditation can thrive on repetitive mantras and chants. It also could contribute to what making some music so popular: by depriving the listener of some usual inputs, that repetitiousness could free higher-level systems to pursue their own ideas. Then, as suggested in §5-8, they could send down some 'simuli' to make some lower level resources simulate some imaginary

fantasies.

Rhythmic and Musical Differences

> *"Music can move us through brief emotional states, and this can potentially teach us how to manage our feelings by giving us familiarity to transitions between the states that we know and thus gain greater confidence in handling them."*
> —*Matthew McCauley*

Music (or art, or rhetoric) can divert you from your mundane concerns by evoking powerful feelings that range from delight and pleasure to sorrow and pain; these can excite your ambitions and stir you to act, or calm you down and make you relax, or even put you into a trance. To do this, those signals must suppress or enhance various sets of mental resources—but why should those kinds of stimuli have such effects on your feeling and thinking?

We all know that certain temporal patterns can lead to rather specific mental states; a jerky motion or crashing sound arouses a sense of panic and fear— whereas a smoothly changing phrase or touch induces affection or peacefulness. [118] Some such reactions could be wired from birth—for example, to facilitate relationships between infants and parents. For then, each party will have some control over what the other one feels, thinks, and does.

Subsequently, as we grow up, we each learn similar ways to control ourselves! We can do this by listening to music and songs, or by exploiting other external things, such as drugs, entertainment, or changes of scene. Then we also discover techniques for affecting our mental states 'from inside'—for example, by thinking that music inside our minds. (This can have a negative side, as when you hear a person complain that they can't get a certain tune out of their head.)

Eventually, for each of us, certain sights and sounds come to have more definite significances—as when bugles and drums depict battles and guns. However, we usually each have different ideas about what each fragment of music means—particularly when it reminds us of how we felt during some prior experience. This has led some thinkers to believe that music expresses those feelings themselves, whereas those effects are probably far less direct:

G. Spencer Brown: "[In musical works] the composer does not even attempt to describe the set of feelings occasioned through them, but writes down a set of

commands which, if they are obeyed by the reader, can result in a reproduction, to the reader, of the composer's original experience.[119]*"*

However, some other thinkers would disagree:

Marcel Proust: "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."

Perhaps Felix Mendelssohn had something like this in mind when he said, *"the meaning of music lies not in the fact that it is too vague for words, but that it is too precise for words."*

All of this raises questions that people seem strangely reluctant to ask—such why do so many people like music so much, and permit it to take up so much of their lives.[120] In particular, we ought to ask why nursery rhymes and lullabies occur in so many cultures and societies. In *Music, Mind, and Meaning* I suggested some possible reasons for this: perhaps we use those tidy structures of notes and tunes as simplified 'virtual' worlds for refining difference-detectors that we can then use for condensing more complex events (in other realms) into more orderly storylike scripts. See also *§§Interior grounding*.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

Difference-Networks.

Whenever you want to accomplish some purpose, you will need to decide which things to change. To do this you'll need to retrieve some knowledge about which actions that might help make those changes. But what should you do when what you have does not exactly match what you need? Then you'll want to find some substitute that is different—but not too dissimilar.

Whenever you want to accomplish some goal, you will need to retrieve some knowledge about some actions or objects that might help. For example, suppose that you want to sit down, so you look for a chair, but none is in sight. However, if there were a bench in view, then you might regard it as suitable. What leads you to see the bench as similar—when you would not so regard a book or a lamp? What makes us selectively notice things that are likely to be relevant? Patrick Winston suggested doing this by organizing some bodies of knowledge into what he called "difference networks"—for example, like this:[121]
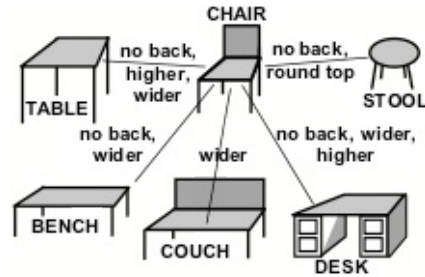
To use such a structure, one first must have some descriptions of the objects it represents. Thus a typical concept of a chair might involve four legs, a level seat, and a vertical back, in which the legs must support the seat from below at a proper height above the floor—whereas a bench is similar (except for being wider and not having a back).
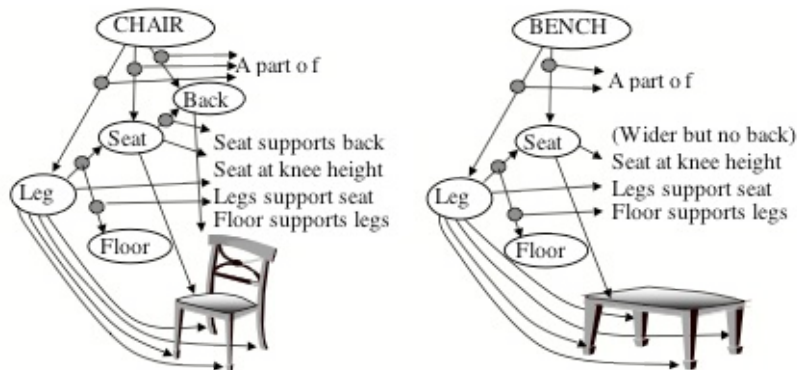
Now, when you look for a thing that matches your description of 'chair,' your furniture-network could recognize a bench as similar. Then you can choose to accept that bench, or reject it because it is too wide or has no back.

How might we accumulate useful sets of difference-links? One way would be that, whenever we find an *A* that 'almost works' (that is, for our present purposes) along with a *B* that actually works, we connect the two with a difference-link that represents, *"A is like B, except for a difference D."* Then such networks could also embody the knowledge we need to change what we have into what we need—as well as to suggest alternative views whenever the present one fails. Thus, such difference-networks could help us to retrieve memories that are relevant.

Traditional programs did not take this approach, but were designed to use more 'logical' schemes—such as regarding a chair as an instance of *furniture,* and a table as one kind of furniture. Such hierarchical classifications often help to find suitably similar things, but they also can make many kinds of mistakes. I suspect that people use both techniques but that the 'sideways' connections in our difference-networks are vital to how we construct the analogies that are among our most useful ways to think.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§6-5. Making Decisions

> *"This river which hid itself doubtless emerged again at some distant spot. Why should I not build a raft and trust myself to its swiftly flowing waters? If I perished, I should be no worse off than now, for death stared me in the face, while there was always the possibility that ... I might find myself safe and sound in some desirable land. I decided at any rate to risk it.*
> *—Sinbad, in The Arabian Nights.[122]*

It is easy to choose among options when one appears better than all of the rest. But when you find things hard to compare, then you may have to deliberate. One way to do this would be to imagine how one might react to each possible outcome, and then, somehow, to compare those reactions, and then to select the one that seems best— or, as in the Sinbad case, to reject the one that seems worse.

Aristotle: "Sensitive imagination is found in every animal, but deliberative imagination only in those that can calculate: for whether this or that shall be enacted is already a task requiring calculation."
   —Aristotle, On the Soul, Book III, 4.

One way a person could 'calculate' would be to assign a numerical score to each choice, and then to select the largest one.

Citizen: Lately, I have been trying to choose between a country home and an apartment in town. The first one offers more spacious rooms and looks out on a beautiful mountain-view. The other is closer to where I work, is in a friendlier

neighborhood, but has a higher annual cost. But how could one measure, or even compare, situations that differ in so many ways?

Still, you could try to imagine how each of those situations would help or hinder you to accomplish your various goals.

Citizen: That might just make the problem worse, because then you have to measure your feelings about the values of those various goals.

Benjamin Franklin: "When these difficult cases occur, they are difficult chiefly because while we have them under consideration all the reasons pro and con are not present to the mind at the same time; but sometimes one set present themselves and at other times another, the first being out of sight. Hence the various purposes or inclinations that alternatively prevail, and the uncertainty that perplexes us."[123]

However, Franklin went on to suggest a way to eliminate much of that measuring:

To get over this, my way is, to divide half a sheet of paper by a line into two columns, writing over the one pro, and over the over con. Then during three or four days consideration I put down under the different heads short hints of the different motives that at different times occur to me for or against the measure. When I have thus got them all together in one view, I endeavor to estimate their respective weights; and where I find two, one on each side that seem equal I strike them out: if a find a reason pro equal to some two reasons con, I strike out the three. If I judge some two reasons con equal to some three reasons pro I strike out the five; and thus proceeding I find at length where the balance lies; and if after a day or two of further consideration nothing new of importance occurs on either side, I come to a determination accordingly. And tho' the weight of reasons cannot be taken with the precision of algebraic quantities, yet when each is considered separately and comparatively and the whole lies before me, I think I can judge better, and am less likely to take a rash step; and in fact I have found great advantage from this kind of equation, in what might be called 'Moral' or 'Prudential Algebra'.

Of course, if such a process were to conclude that several options seem equally good, then you would have to switch to another technique. You sometimes do this reflectively, but at other times the rest of your mind does this without your knowing how the decision was made. At such times you might say things like, *"I used my 'gut feelings'"* or used *'intuition,'* or claim that you did that *'instinctively.'*

Paul Thagard: "Many persons trust their "gut feelings" more. ... You may have a strongly positive gut feeling toward the more interesting subject along with a strongly negative feeling about the more career-oriented one, or your feelings may be just the opposite. More likely is that you feel positive feelings toward both alternatives, along with accompanying anxiety caused by your inability to see a clearly preferable option. In the end, intuitive decision makers choose an option based on what their emotional reactions tell them is preferable."[124]

However, using the word 'emotional' does not help us to see what is happening—because how 'positive' or 'negative' a feeling seems will still depend on how one's mental processes deal with *"all the reasons pro and con"* that Franklin addressed in that letter. Indeed, we frequently have the experience that, shortly after we make a decision, we find that it *'just does not feel right'* — and go back to reconsidering.

Citizen: Even when options seem equally good, I still can decide. How could your kind of theory explain our peculiarly human 'freedom of choice'?

It seems to me that when people say, *'I used my free will to make that decision,'* this is roughly the same as saying that *'some process stopped my deliberations and made me adopt what seemed best at that moment.'* In other words, "free will" is not a process we use to make a decision, but one that we use to stop other processes! We may think of it in positive terms but perhaps it also serves to suppress the sense that we are being forced to make a choice—if not by pressures from outside, but by causes that come from inside our own minds. To say that *'my decision was free'* is almost the same thing as saying *"I don't want to know what decided me."*[125]

<div align="center">∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞</div>

§6-6. Reasoning by Analogy

> *"If I had eight hours to chop down a tree, I'd spend six sharpening my axe."*
> —A. Lincoln

The best way to solve a problem is to already know a solution for it—and this is why commonsense knowledge is useful. But what if the problem is one

you have never seen before? How can you continue to work when you lack some of the knowledge you need? The obvious answer: you just have to guess—but how does one know how to make a good guess? We usually do this so fluently that we have almost no sense of how we are doing it, and, if someone asks about that, we tend to attribute it to mysterious traits with names like *intuition, insight, creativity,* or even to *intelligence.*

More generally, whenever anything attracts your attention—be it an object, idea, or a problem—you are likely to ask yourself what that thing is, why is it there, and whether it should be a cause for alarm. But as we said in §6-3, we can't usually say what anything *is:* we can only describe what something is *like,* and then start to think about questions like these:

> *What sorts of things is this similar to?*
> *Have I seen this anything like it before?"*
> *What else does it remind me of?*

This kind of thinking is important because it helps us to deal with new situations—and in fact that is almost always the case, because no two situations are ever the same—and this means that we're always making analogies. For example, if the problem that you are facing now reminds you have one that you solved in the past, then you may be able to use this to solve your problem by using use a procedure like this:

> *The problem that I am working on reminds me of a similar one that I solved in the past—but the method that was successful then does not quite work on the problem that I am facing now. However, if I can describe the differences between that old problem and this new one, those differences might help me to change that old method so that it will work for me now.*

We call this 'reasoning by analogy' and I'll argue that this is our most usual way to deal with problems. We do this because, in general, old methods will never work perfectly, as new situations are never quite the same. So, instead, we use analogies. But, why do analogies work so well? Here is the best way I've seen to explain why this is:

> *Douglas Lenat: "Analogy works because there is a lot of common causality in the world, common causes which lead to an overlap between two systems, between two phenomena or whatever. We, as human beings, can only observe a tiny bit of that overlap; a tiny bit of what is going on at this level of the world. ... [So] whenever we find an overlap at this level, it is worth seeing if in fact there are additional overlap features, even though we do not understand the cause or*

causality behind it.[126]

So, now let's inspect an example of this.

A Geometric Analogy Program

Everyone has heard about great improvements in computer speed and capacity. It is not so widely known that, in other respects, computers changed very little from their inception until the late 1970's. Designed originally for doing high-speed arithmetic, it was usually assumed that this was all computers would ever accomplish—which is why they were misnamed "computers."

However, people soon began to write programs to deal with non-numerical things such as linguistic expressions, graphical pictures, and various forms of reasoning. Also, instead of following rigid procedures, some of those programs were designed to search through wide ranges of different attempts — so that they could solve some problems by "trial and error"—instead of by using pre-programmed steps. Some of these early non-numerical programs became masters at solving some puzzles and games, and some were quite proficient at designing new kinds of devices and circuits.[127]

Yet despite those impressive performances, it was clear that each of these early "expert" problem-solving programs could operate only in some narrow domain. Many observers concluded that this came from some limitation of the computer itself. They said that computers could solve only "well-defined problems" and would never be able to cope with ambiguities, or to use the kinds of analogies that make human thinking so versatile.

To make an analogy between two things is to find ways in which they are similar—but when and how do we see two things as similar? Let's assume that they share some common features, but also have some differences. Then how similar they may seem to be will depend upon which differences one decides to ignore. But the importance of each difference depends upon one's current intentions and goals. For example, one's concern with the shape, size, weight, or cost of a thing depends on what one plans to use it for—so, the kinds of analogies that people will use must depend upon their current purposes. But then, the prevailing view was that no machine could ever have goals or purposes.

Citizen: But, if your theory of how people think depends on using analogies, how could any machine do such things? People have always told me that machines can only do logical things, or solve problems that are precisely defined—so they cannot deal with hazy analogies.

In 1963, to refute such beliefs, Thomas G. Evans, a graduate student at MIT,

wrote a program that performed surprisingly well in what many people would agree to be ambiguous, ill-defined situations. Specifically, it answered the kinds of questions in a widely used "intelligence test" that asked about "Geometric Analogies."[128] For example, a person was shown a picture like this and asked to choose an answer to: *"A is to B as C is to which of the other five figures?"* Most older persons choose figure 3—and so did Evans's program, whose score on such tests was about the same as that of a typical 16 year old.

In those days, many thinkers found it hard to imagine how any computer could solve such problems, because they felt that choosing an answer must come from some "intuitive" sense that could not be embodied in logical rules. Nevertheless, Evans found a way to convert this to a far less mysterious kind of problem. We cannot describe here all the details of his program, so we will only show how its methods resemble what people do in such situations. For if you ask someone why they chose Figure 3, they usually give an answer like this.

> *"You go from A to B by moving the big circle down, and*
> *You go from C to 3 by moving the big triangle."*

This statement expects the listener to understand that both clauses describe something in common—even though there is no big circle in Figure 3. However, a more articulate person might say:

> *"You go from A to B by moving the largest figure down, and*
> *You go from C to 3 by moving the largest figure down."*

Now those two clauses are identical—and this suggests that both a person and a computer could use a 3-step process involved these similar kinds of descriptions.

Step 1. Invent descriptions for each of top row of figures. For example, these might be:

> *A: high large thing, high small thing, low small thing.*

B: low large thing, high small thing, low small thing,
C: high large thing, high small thing, low small thing.

Step 2. Invent an explanation for how A might have been changed to B. For example, this might simply be:

Change "high large" to "low large."

Step 3. Now use this to change the description of figure C. The result will be:

Low large thing, high small thing, and low small thing.

The result is a prediction of how Figure C might also be changed. If this matches one of the possible answers more closely than any other, then we'll choose that as our answer! In fact it only matches Figure 3, which is the one most people select. (If more than one possible answer is matched, then the program starts out again by using different descriptions of the same pictures. Thus, because it has multiple "ways to look at things," the program usually finds a way to find a good solution. The program performed on this kind of test as well as a typical fifteen-year old. To be sure, it could work only this kind of problem, and had no way to learn from experience, but still, knowing ways to use analogies is a vital part of how people think.

Of course, whenever we need to make a choice, the differences that will concern us most will depend on what we now want to achieve. If Carol wants merely to build an arch, then all of these forms may seem adequate—but if she plans to build more on its top, then the one on the right will seem less suitable.

Although these particular "geometric analogy" problems are not so common in everyday life, Evan's program shows the value of being able to change and adapt its descriptions until it finds a way to describe different things so that they seem more similar. This is an important step toward the ability to use knowledge about one kind of thing to understand some other different kind of thing—so finding *new ways to look at things* this must be among our most important commonsense processes.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§6-7. Knowledge needs Multiple Representations

What distinguishes people from animals? Perhaps our largest distinction is that none of the others can ask such a question! We're unique in being able to treat our own ideas as though they were things. In other words, we 'conceptualize.'

However, to think about ideas or things, we need representations of them in our minds. Everyone who has written a program knows that you can't get a computer to do what you want by simply 'pouring knowledge in.' You must represent each process or fact in the form of some sort of structure. For knowledge is not composed of 'things' that each can exist apart from the rest—no more than a word can have a meaning without being part of some larger-scale language; fragments of knowledge can only make sense when they have appropriate kinds of interconnections. It does not much matter how these are embodied; you can make the same computer with wires and switches, or even with pulleys, blocks, and strings; all that matters is how each part changes its state in response to what some other parts do. And the same kinds of relationships can also be represented in terms of parts that have no behavior at all—such as arrangements of symbols in diagrams, or the sentences of written texts—so long as there is some way these to affect how some other systems will behave.

So when programmers set out to develop a program, they usually start by selecting a way to represent the knowledge their program will need. But each representation works well only in certain realms, and none works well in every domain. Yet we frequently hear discussions like this about what is the best way to represent knowledge:

Mathematician: It is always best to express things with Logic.

Connectionist: No, Logic is far too inflexible to represent commonsense knowledge. Instead, you ought to use Neural Networks.

Linguist: No, because Neural Nets are even more rigid. They represent things in numerical ways that are hard to convert to useful abstractions. Instead, why not simply use everyday language—with its unrivaled expressiveness.

Conceptualist: No, language is much too ambiguous. You should use Semantic Networks instead—where ideas get connected by definite concepts!

Statistician: Those linkages are too definite, and don't express the uncertainties we face, so you need to use probabilities.

Mathematician: All such informal schemes are so unconstrained that they can be self-contradictory. Only Logic can ensure us against those circular inconsistencies."

This shows that it makes no sense to seek a single best way to represent knowledge—because each particular form of expression also brings its own particular limitations. For example, *logic-based systems* are very precise, but they make it hard to do reasoning with analogies. Similarly, *statistical systems* are useful for making predictions, but do not serve well to represent the reasons why those predictions are sometimes correct. It was recognized even in ancient times that we must represent things in multiple ways:

Aristotle: "Thus the essence of a house is assigned in such a formula as 'a shelter against destruction by wind, rain, and heat'; the physicist would describe it as 'stones, bricks, and timbers'; but there is a third possible description which would say that it was that form in that material with that purpose or end. Which, then, among these is entitled to be regarded as the genuine physicist? The one who confines himself to the material, or the one who restricts himself to the formulable essence alone? Is it not rather the one who combines both in a single formula?"[129]

However, sometimes there are advantages to *not* combining those ways to describe things.

Richard Feynman: "...psychologically we must keep all the theories in our heads, and every theoretical physicist who is any good knows six or seven different theoretical representations for exactly the same physics. He knows that they are all equivalent, and that nobody is ever going to be able to decide which one is right at that level, but he keeps them in his head, hoping that they will give him different ideas for guessing."[130]

Much of our human resourcefulness comes from being able to choose among diverse ways to represent the same situation. This has value because each such point of view may provide a way to get around some deficiencies of the other ones. However, to exploit this fact, one needs to develop good ways to decide when to use each kind of representation; we'll come back to this in §10-X. {Causal Diversity.} Of course, to change representations efficiently, one must also be able to quickly switch without losing the work that's already been done —and that is why this chapter emphasized the iuse of panalogies to link analogous aspects of multiple ways to represent and to think about things.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

QUESTIONS

Part VII. Thinking

"I am aware of a constant play of furtherances and hindrances in my thinking, of checks and releases, tendencies which run with desire, and tendencies which run the other way ... welcoming or opposing, appropriating or disowning, striving with or against, saying yes or no."

—William James, [Principles of Psychology]

Which characteristics help us to surpass all the rest of our animal relatives? Surely our most outstanding such trait is our knack for inventing new Ways to Think.

Romanticist: You claim that our finest distinction is thinking—yet perhaps we are even more unique in our richness of how we experience things. There's the joy of turning one's intellect off, to enjoy a sunset or listen to birds, or to sing or do a spontaneous dance in delight of the sense of being alive.

Determinist: People use words like 'spontaneous' to make themselves feel that they aren't constrained. But perhaps that sense of enjoying ourselves is merely a trick that some parts of our brains use to make us do what they want us to do.

In any case, I doubt that we ever stop thinking entirely, because *thinking* refers, at different times, to a huge range of intricate processes.

Citizen: If our everyday thinking is so complex, then why does it seem so straightforward to us? If its machinery is so intricate, how could we be unaware of this?

That illusion of simplicity comes from forgetting our infancies, in which we grew those abilities. As children we learned how to pick up blocks and arrange them into rows and stacks. Then as each new group of skills matured, we built yet more resources on top—just as we learned to plan and build more elaborate arches and towers.

Along with this, in those early times, we assembled the towers of aptitudes that we describe with words like *minds*. But now, as adults, we all believe that *we have always been able to think*—because we learned those skills so long ago that we cannot recall having learned them at all.

It took each of us many years of hard work to develop our more mature ways to think—but whatever records remain of this have somehow become inaccessible. What could have made us all victims to that "amnesia of infancy?" I don't think this is simply because we 'forgot.' Instead, I suspect that it's largely because we kept developing new, better techniques for representing both physical and mental events—and some of these methods became so effective that we abandoned the use of our previous ones. Now, even if those old records still exist, we no longer can make any sense of them.

In any case, the result of this is that now we all think without knowing how we think—and we do it so fluently that we scarcely ever ask about what thinking it is and how it might work. In particular, we like to celebrate grand accomplishments in the sciences, arts, and humanities—but we rarely applaud— or ask questions about—the marvels of everyday, commonsense thought. Indeed, we often see thinking as more or less passive, as though our thoughts just "come to us" and we say things like, *"It occurred to me"* or *"A thought entered my mind,"* instead of, *"I just made a new idea."* Thus we talk as though we don't deserve any credit for our ideas, and we scarcely ever wonder about what chooses which subjects we think about.

One of the wooden doors in my home bears scratches made more than a decade ago. Our dog Jenny is gone but the scratches remain. I notice them only a few times a year, though I pass by that door several times every day.

Every day you encounter great numbers of things, yet only a few of them 'get your attention' enough to make you ask questions like, *"What is that object and why is it here,"* or *"Who or what caused that to happen?"* Most times your thinking proceeds in a smooth, steady flow in which you scarcely ever reflect on how you get from each step to the next.

At yet other times, your mind seems to wander without any sense of direction at all. First you might dwell on some social affair, then you reflect on some past event; next you're beset by a hunger pang, or the thought of a payment that's overdue, or an impulse to fix the faucet-drip, or an urge to tell Charles how you feel about Joan. Each item reminds you of something else until some mental 'Critic' cuts in with, *"This isn't getting you anywhere,"* or *"You must try to get more organized."*

However, there are certain times when your thinking is much more aim and direction. This happens when you are pursuing a certain goal, but encounter an impasse or obstacle like, *"I can't pack all this into this box—and besides, that would make it too heavy to lift."* Then you may stop to deliberate: *"It looks like this will take several trips, but I don't want to spend that much time on this."*

Much of this chapter will discuss the idea that such recognitions of obstacles play critical roles in controlling our higher levels of thinking.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

This chapter will develop the idea that each person has many different ways to think. One could ask why we have so many of those, and one answer would be that our ancestors lived through a host of varied environments, each of which required ways to deal with different conditions and constraints. Then, because we never discovered one uniform scheme that could meet all our needs, we retained large parts of that collection of methods for coping with different situations.

Generally, we do not seem to be much aware of switching among all those ways to think. Perhaps this in large part because we all have that sense of having (or being) a Single Self—so one rarely asks a question like, "What prevents any part of my mind from seizing control of all the rest?" *(Such accidents must have happened to many individuals in the course of human history—but their genes failed to propagate because they lacked enough versatility.) The result was that, over eons of time, our brains evolved a good many different ways to avoid the most common kinds of mistakes, while still staying able to adapt to a series of new environment; this is how evolution works;* each species evolve at the edge of some zone between the safeties they know and the dangers they don't.

Psychiatrist: *That safety-zone can be narrow indeed. Most of the time, most minds function well, but sometimes get into various states in which they can scarcely function at all—and then we say that they're mentally ill.*

Physiologist: *Surely most such disorders have medical causes—such as traumatic injuries, or chemical imbalances, or diseases that damage our synapses.*

Programmer: *Perhaps, but we should not assume that all such disorders have non-mental causes. When a 'software virus' infects a computer and changes some data on which it programs depends, the hardware is not damaged at all, but still there are serious changes in how it behaves.*

Similarly, a new destructive goal or idea—or a change in one's Critics or Ways to Think—could gain control of so much of a person's resources and time that it could affect multiple realms of knowledge and thought—and thus spread like a mental malignancy.

Sociologist: *Perhaps it's the same on a larger scale, when the notions of a*

sect or cult include ways to discern potential recruits, in whom its ideas and belief will propagate.

<div align="center">∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞</div>

§7-1. What selects the subjects we think about?

What selects what we'll think about next, from among all our various interests—and how long will we persist with each? Let's consider a typical, everyday incident:

Joan needs to write a project report, but has not made much progress on it. Discouraged, she sets those thoughts aside and finds herself roaming about her house with no particular goal. She passes an untidy stack of books, and stops for a moment to straighten them out. But then she 'gets' a new idea, so she goes to her desk to type a note. She starts to type—but finds that the 'T' on her keyboard is stuck. She knows how to fix this, but worries that, then, she might forget that new idea—so, instead, she makes a handwritten note.

What led Joan to notice that pile of books? Why did that that idea 'occur' to her now, instead of at some other time? Let's look more closely at these events.

Joan has not made much progress. *Some mental 'Critic' must have noticed this and suggested that she 'take a break.'*

Discouraged, Joan sets those thoughts aside. *When and how will she bring them back? That will depend on the extent to which she can later find records of them. Section §7-9 will ask about how we remember the contexts of our recent thoughts.*

Joan is roaming without any goal. *Or so it may seem—but most animals have instincts to maintain their 'territories' or nests. Joan usually walks right past that spot without giving it a second thought—but perhaps right now she is 'making rounds' because she is mainly controlled by Critics that aim to maintain the tidiness of her home.*

She passes an untidy stack of books, and stops for a moment to straighten them out. *Why doesn't Joan stop now to read those books, instead of just trying to tidy them up? Perhaps this is because the Critics that are most active now represents those books as untidy objects (rather than as containers of knowledge)—so she's more concerned with how they look than with the subjects that they are about.*

But then she 'gets' a new idea. *When people say, "It occurred to me," this show how limited is the extent to which we can reflect on how we produce our*

ideas.

    Joan goes to her desk to type a note. *Joan knows that when she "gets" an idea, she cannot depend on remembering it—and so she puts her housekeeping on hold to make a more permanent record.*

    She finds that the 'T' on her keyboard is stuck. She knows how to fix this, but worries that then she might forget that new idea. *She is using her self-reflective knowledge about the qualities of her short-term memories.*

Perhaps most of the time, we mainly react to things that happen, without much sense of making decisions. However, our higher-level thinking is much affected by our wishes, fears, and larger-scale plans—as well as by other aspects of the context we're in. This raises many questions about how we spend our mental time:

> *What schedules our large-scale plans?*
> *What reminds us of things that we promised to do?*
> *How do we choose among conflicting goals?*
> *What decides when we should quit or persist?*

Any good model of commonsense thinking should suggest some answers to questions like these. However, so long as everything goes well, your thoughts seem to proceed in a steady, smooth flow. Each minor obstacle makes only small changes in how you think, and if you 'notice' these at all, they merely appear as transient feelings or as fleeting ideas. However, when more serious obstacles persist and keep you from making progress, then, various Critics intervene to make larger changes in how you think.

§7-2. Emotional Thinking

> *There is a very fine line between "hobby" and "mental illness."*
>   *—Dave Barry*

Most of the time your thinking proceeds in routine, uneventful streams—but sometimes you run into obstacles that interrupt your orderly progress. Then you'll have to find something else to do, and this may lead to a spreading cascade changes in the way you think.

    Changing the subject. *Whatever you are doing now, there are always other things you could do, so whenever you get discouraged with one, you might want*

to switch to another.

Self-Determination. *If you are tempted to abandon your task, you can renew your motivation by bribing yourself with imagined rewards, or with threats of the prospect of failure.*

Self-Conscious Reflection. *If that doesn't work, you might start to imagine how you (or your imprimers) would feel if your performance conflicted with your ideals.*

But when none of those methods turns out to help, one still can use several 'last resorts.'

Self-Regression: *When your situation seems to become so complex that you see no way to deal with it, you still can ask yourself, "How did I deal with such things in the past?" Then you may be able to 'regress' to some earlier version of yourself, from an age when such things seemed simpler to you.*

Cry for Help! *If you can't find a way to do something yourself, you might attempt to exploit the resources of your friends. As infants, we were designed to do this, using signals that hijack more powerful minds.*

Emotional thinking: *A flash of impatience or anger can cut through what seems like a hopelessly tangled knot. Each such 'emotional way to think' is a different way to deal with things, and some can increase your persistence or courage, while others can help you simplify things.*

In any case, after each such change, you may still want to pursue some similar goals, but now you'll see them from new points of view— because each switch to a new Way to Think may initiate a large-scale cascade. Then, depending on how long those changes persist, you (or your friends) might recognize this as a change in your emotional state.

Various parts of our states of mind can continue for different scales of time. Some last for no more than the blink of an eye, but infatuations persist for days or weeks. However, when other 'dispositions' endure for substantial spans of a individual's life, we see as aspects of that person's personality,' and we call these *characteristics* or *traits*.

For example, when solving a problem, some people tend to accept a solution that still has some deficiencies—so long as it seems to work well enough: you might describe such a person as *realistic, pragmatic,* or *practical.* Another person may tend to insist that every potential flaw must be fixed—and you might call such people *fastidious*—except when they make you uncomfortable, in which case you call them *obsessive* instead. Other such dispositions include being *Cautious vs. Reckless, Inattentive vs. Vigilant, Unfriendly vs. Amicable,*

Reclusive vs. Sociable, Visionary vs. Down-to-Earth, or *Courageous vs. Cowardly.*

In fact, in the course of everyday thought, each person is likely to frequently switch among such views or attitudes, and we usually don't even notice this. However, when we encounter more serious trouble, our Critics may make enough changes to start the large-scale cascades that we describe in terms of emotional states.

Psychiatrist: What would happen if too many Critics were active? Then your emotions would keep changing too quickly. And if those Critics stopped working at all, then you'd get stuck in just one of those states.

Perhaps we can see an example of this in Antonio R. Damasio's book, *Descartes' Error,*[131] which describes a patient named Elliot, who had lost some parts of his frontal lobes in the course of removing a tumor. After that treatment, he still seemed intelligent—but his friends and employers had the sense that Elliott was 'no longer himself.' For example, if asked to sort some documents, he was likely to spend an entire day at carefully reading just one of those papers —or at trying to decide whether to classify them by name—or by subject or size or date or by weight.

Damasio: "One might say that the particular step of the task at which Elliot balked was actually being carried out too well, and at the expense of the overall purpose. ... True, he was still physically capable and most of his mental capacities were intact. But his ability to reach decisions was impaired, as was his ability to make an effective plan for the hours ahead of him, let alone to plan for the months and years of his future."

The damaged parts of Elliot's brain included certain connections (to the amygdala) that are widely believed to be involved with how we control our emotions.

Damasio: "At first glance, there was nothing out of the ordinary about Elliot's emotions. ... However, something was missing. ... He was not inhibiting the expression of internal emotional resonance or hushing inner turmoil. He simply did not have any turmoil to hush. ... I never saw a tinge of emotion in my many hours of conversation with him: no sadness, no impatience, and no frustration with my incessant and repetitious questioning."

This led Damasio to suggest that *"reduced emotion and feeling might play a role in Elliot's decision–making failures."* However, we could also consider this opposite view: that it was Elliot's *new inability to make such decisions that*

reduced his range of emotions and feelings. For, perhaps the damage in Elliott's brain was mainly to some of the Critics (or to their connections) that formerly set off the large-scale cascades that we recognize as emotional states. Then he would have lost those precious cascades—and hence, the emotions that he once displayed—because he could no longer could exploit those Critics to choose which emotional states to use.

<center>∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞</center>

§7-3. The Critic-Selector Model of Mind

> *I have yet to see any problem, however complicated, which, when you looked at it in the right way, did not become still more complicated.*
> —*Poul Anderson*

> *No problem is so formidable that you can't walk away from it.*
> —*Charles Schulz*

We frequently change what we're thinking about, without noticing that we are doing this—because it is mainly when some trouble comes that we start to reflect about thinking itself. Thus, we don't recognize a problem as 'hard' until we've spent some time on it without making any significant progress. Even then, if that problem does not seem important, you might just abandon that line of thought and simply turn to some other subject.

However, if you have an important goal, then it is useful to notice that you are stuck—and it will be even more useful if you also can recognize that you're being blocked by a certain particular type of barrier, obstacle, impasse, or snag. For if you are able to *diagnose* the particular *"Type of Problem"* you face, then that knowledge can help you to switch to a more appropriate *"Way to Think."*

This suggests a Model of Mind based on reacting to 'cognitive obstacles.' We'll call this the **Critic-Selector** model:
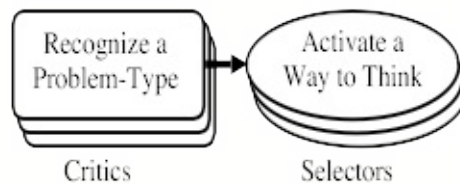
On the left are resources that we shall call **Critics**, each of which can recognize a certain species of "Problem-Type." When a Critic sees enough evidence that you now are facing its type of problem, then that Critic will try to activate a "Way to Think" that may be useful in this situation.

Cloud of Resources
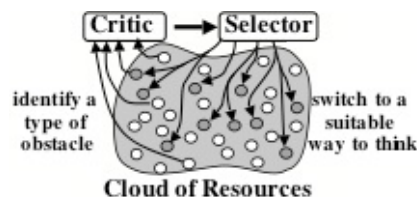
For example, a Critic-Selector model could embody a set of 'rules' like these:

If a problem seems familiar, try reasoning by Analogy.
If it seems unfamiliar, change how you're describing it.
If it still seems too difficult, divide it into several parts.
If it seems too complex, replace it by a simpler one.
If no other method works, ask another person for help.

Every person accumulates a collection of different *"Ways to Think"* because, as we've repeated many times, no single method or mental technique can solve every kind of problem-type; however, if we have enough of them then, whenever the one we're using fails, we'll be able to switch to a different one.

However, there is a problem that is sure to arise in any system based on *If-Then* rules: what if more than one Critic or *"If"* is aroused?[132] Then we might decide which one to use by adopting some policy like these:

Choose the Critic with the highest priority. [Ref: GPS]
Choose the one that is most strongly aroused. [Ref. Pandemonium]

Choose the one that gives the most specific advice. [Ref. Raphael]
Have them all compete in some 'marketplace.' [See §9-X]

Simple strategies like these will work in simple cases, but will fail in more complex situations. Then we'll need higher-level Critics that recognize and suggest ways to change our bad selections of low-level Critics:

*If too many Critics are aroused, **then** describe the problem in more detail.*
*If too few Critics are aroused, **then** make the description more abstract.*
*If important resources conflict, **then** try to discover a cause for this.*
*If there has been a series of failures, **then** switch to a different set of Critics.*

Sometimes we recognize, after the fact, that our selections may have been incorrect, and that we may need to edit our collection of Critics:

I selected a hard-to-use method, but realized that I knew a simpler yet better one.
I now see that the action I took had a bad, irreversible side effect.
I regarded that as an obstacle, but now I see that it was valuable.
Although that method caused some trouble, I learned a lot from using it.

To recognize those kinds of events would require Critics that work at higher levels—and all this suggests that our model of mind should include Selectors and Critics at *every* level.[133]

The following sections will discuss some of our many Ways to Think, and

some of the Critics we use to recognize various ways in which we get stuck.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§7-4. What are some useful "Ways to Think?"

"When you want people to think you are brilliant, just imagine the worst thing that you could do and then do precisely the opposite."
—*Naomi Judd. [get her permission]*

It is mainly when we get into trouble that we engage the processes that we call *thinking* or *reasoning*. However, 'thinking' is no single, definite thing; instead, we use different 'Ways to Think' for dealing with different types of obstacles. It ought to be one of our central goals—both for AI and for Psychology—to classify our Ways to Think. However, we don't yet have systematic ways to classify those abilities—so I'll just list some examples of them.

Knowing How: *The best way to solve a problem is to know how to solve it and use that solution. However, we may not know how to retrieve what we know, or even know that we know it.*

Extensive Search. *When one knows no better alternative, one could search through all possible chains of actions—but this is usually impractical because that search grows exponentially.*

Reasoning by Analogy: *When a problem reminds you of one that you solved in the past*, you may be able to adapt that case *to the present situation—if you have good ways to tell which similarities are most relevant.*

Divide and Conquer. *If you can't solve a problem all at once, then break it down into smaller parts. For example, every difference we recognize may suggest a separate sub-problem to solve.*

Planning. *Consider the set of subgoals you want to achieve and examine how they affect each other. Then, with those constraints in mind, propose an efficient sequence for achieving them.*

Simplification.*Sometimes, a good way to make a plan is to make a simplified problem by ignoring some aspects of the original one. Then any solution to the simplified one may serve as a sequence of stepping-stones for solving the initial problem.*

Elevation. *If you are bogged down in too many details, describe the*

situation in more general terms. But if your description seems too vague, switch to one that is more concrete.

Reformulation. *Find a different representation that highlights more relevant information. We often do this by making a verbal description—and then 'understanding' it in some different way!*

Self-reflection. *Instead of pursuing a problem itself, ask what makes that problem seem hard, or what you might be doing wrong. This can lead to better ways to represent the problem.*

Contradiction. *Try to prove that your problem cannot be solved, and then look for a flaw in that argument.*

Use external representations: *If you find that you're losing track of details, you can resort to keeping records and notes, or drawing suitable diagrams.*

Simulation. *One can avoid taking physical risks if one can predict "what would happen if" by imagining possible actions inside the mental models that one has built.*

Correlation. *When certain events seem to happen together, try to find ways in which they may be connected.*

Logical Reasoning.*We sometimes make 'logical chains of deductions,' but those conclusions may be wrong because of exceptions to our assumptions.*[134]

Wishful thinking. *Imagine having unlimited time and all the resources that you might want. If you still can't envision solving the problem, then you should reformulate it.*

Impersonation. *When your own ideas seem inadequate, imagine someone better at this, and try to do what that person would do.*

Cry for help.*You can always resort to other techniques that most people would call "emotional."*

Resignation.*Whenever you find yourself totally stuck, you can shut down the resources you're using now and relax, lay back, drop out, and stop. Then the 'Rest of Your Mind' may find an alternative—or conclude that you don't have to do this at all.*

How do we choose which of these to use? The Critic-Selector model suggests that each person can recognize particular ways in which one gets stuck —and can use each such diagnosis to select one or more particular ways to deal with that kind of predicament. We each do this in different ways, and the Critics that we each develop must be among our most precious resources.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§7-5. What are some useful Critics?

"Don't pay any attention to the critics. Don't even ignore them."
—*Samuel Goldwyn*

We are always developing new ways to think—so we also need to make Critics to help to select when to use each of those techniques—by recognizing which kinds of problem we face. This means that our Critics must serve as ways to classify all the barriers, obstacles, impasses, or snags that make our problems hard to solve. Indeed, it would be an important goal, both for people and for computing machines, to have a systematic catalog of the types of problems we most frequently face.[135] However, we do not yet have adequate, orderly ways to do this—so here we'll merely try to describe a few types of Critics that people seem to use.

Innate Reactions and built-in Alarms. Many types of external events arouse detectors that make us quickly react, as when an object is quickly approaching you, a light is too bright, you touch something hot, or hear a loud sound. We're also born with ways to detect abnormal conditions *inside* our skins —such as wrong levels of chemicals in the blood. Many of these have built-in connections that make us react to correct those conditions, in ways that work so automatically that we react to them without any thought.

However, an unexpected touch, sight, or smell—or a sense of hunger, fatigue, or pain—*does* interrupt the flow of our thoughts. Indeed we'd never survive through our infancies unless such emergencies (or opportunities) could pull us away from our reveries. We can sometimes suppress some of those alarms; for example, when we suppress a sneeze, or stop ourselves from scratching an itch. But if you try to hold your breath, you can't resist the alarm of asphyxia—and it is hard to ignore a baby's cry, a constantly ringing telephone, or an amorous opportunity.

Learned Reactive Critics. An infant will cry when it is exposed to high levels of noise—thus summoning a parent to help. However, later we learn other ways to react, such as moving to a quieter place. And eventually we learn to 'figure out' ways to deal with more difficult obstacles—by using higher levels of what we call 'deliberative' thinking and then it would seem to make more sense to think of these as involving our Critics.

Deliberative Critics. Whenever your reasoning gets into trouble, you need ways to get around obstacles. Here are some tricks we can use for this:

Action A did not do what I expected. (Try a different Way to Predict.)
Something I did had bad side effects. (Try to undo some previous choice.)
*Achieving goal **A** made goal **B** harder. (Try them in the opposite order.)*
I need additional information. (Search for another relationship.)

Reflective Critics.[136] When you try to solve problems by trial and error, you need critics as 'diagnosticians' to either verify that you're making progress or to suggest a better way to proceed.

I've made many attempts with no success. (Select a better way to think.)
I've repeated the same thing several times. (Some 'mental manager' is incompetent.)
Achieving a subgoal did not attain its 'parent' goal. (Find another way to subdivide the problem.)
This conclusion needs more evidence. (Propose a better experiment.)

Self-Reflective Critics. When your reflections fail to help, then you may start to criticize yourself:

I have been too indecisive. (Try a method that worked on a similar problem.)
I missed a good opportunity. (Switch to a different set of Critics.)
I yield to too many distractions. (Try to exercise more Self-Control.)
I don't have all the knowledge I need. (Find a good book or go back to school.)

Self-Conscious Critics. Some assessments may even affect one's current image of oneself, and this can affect one's overall state:

None of my goals seem valuable. (Depression.)
I'm losing track of what I am doing. (Confusion.)
I can achieve any goal I like! (Mania.)
I could lose my job if I fail at this. (Anxiety.)
Would my friends approve of this? (Insecurity.)

I should note that we often say "Critic" to mean someone who points out deficiencies, and it would be hard to describe the *Correctors, Suppressors* and *Censors [§3-5]* without using negative words like *inhibit, prevent,* or *terminate.*

However, words like *positive* and *negative* usually do not make sense by themselves; here, detecting a flaw can be and essential step toward helping one to achieve a success—for example, by keeping you from changing your goal or from wasting your time on other temptations—and thus encouraging you to persist. Frequently, the key to solving a difficult problem can lie in finding ways to make yourself 'stick to a plan', although it may bring some suffering before

you achieve your longer-range goal.

Indeed, after we solve a difficult problem, we may wrongly credit our final success only to our very last step, and tell ourselves, *"What a clever solution I've found!"* Then, of course, it makes good sense to remember the answer to that particular question. However, it would often be better to also ask, *"What kept me from finding it earlier?"* For, what often makes a question seem 'hard' is not knowing a good way to search for the answer. This suggests that after we answer a difficult question, it may be useful to remember which strategy led to solving it by reducing the size of the search for the answer. (A good way to 'remember' this would be to create a new Critic to recognize that problem-type, and connect it to a Selector for that strategy.)

This subject of "Credit-Assignment" is very important because it bears on the quality of what people learn. Indeed Chapter 8 will take a further step and argue that:

What we learn can be more profound, if we assign the credit for our success, not to the final act itself—or even to the strategy that led to it—but to some even earlier choice of a process or plan that selected the winning strategy.

Generally, lower-level Critics will tend to have shorter-term effects. Thus, *"Make sure that your elbow won't topple that block,"* can alter your tactics temporarily, without changing your larger-scale strategy; then even if this leads to making a mistake, you may be able to correct it and continue with your original plan. However, high-level critics can cause longer-term changes—for example, by switching you to self-reflective thoughts like, *"I'm not good at solving this kind of problem. Perhaps it is time to consider a different profession."*

In any case, repeated failures can cause you to 'brood' about what the future might hold for you or about your social relationships, as in, *"I should not get into such situations,"* or *"My friends will lose their respect for me,"* or *"I don't have enough self-discipline."* Such thoughts can lead to the large-scale cascades that we usually call 'emotional.'

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§7-6. Emotional Embodiment

Many thinkers have maintained that emotional states are closely involved with our bodies—and that this is why we so often can recognize *happiness, sadness, joy,* or *grief* from a person's expressions, gestures, and gaits. Indeed,

some psychologists have even maintained that those bodily activities do not merely 'express' our emotions, but actually are what causes them:

William James: "Our natural way of thinking about ... emotions is that the mental perception of some facts excites the mental affection called the emotion, and that this latter state of mind gives rise to the bodily expression. My theory, on the contrary, is that the bodily changes follow directly the perception of the exciting fact, and that our feeling of the same changes as they occur is the emotion."

For example, James suggests that when you sense that a rival is insulting you, this makes you clench your fist and strike—and that your anger does not come first, but from you feeling of these activities. However, your annoyance with such a situation must depend on the memories that intervene to affect how we interpret those 'exciting facts'—and *then* cause you to clench your fist—so it seems unlikely that those perceptions 'directly' lead to those actions. Nevertheless, James argues that such intermediate thoughts could not have such strong effects by themselves:

William James: If we fancy some strong emotion and then try to abstract from our consciousness of it all the feelings of its bodily symptoms we find we have nothing left behind, no "mind stuff" out of which the emotion can be constituted, and that a cold and neutral state of intellectual perception is all that remains. ... [I cannot imagine] what kind of an emotion of fear would be left if the feeling neither of quickened heart beats nor of shallow breathing, neither of trembling lips nor of weakened limbs, neither of goose flesh nor of visceral stirrings, were present. ... Can one fancy the state of rage and picture no ebullition in the chest, no flushing of the face, no dilation of the nostrils, no clenching of the teeth, no impulse to vigorous action, but in their stead limp muscles, calm breathing, and a placid face.[137]

Nevertheless, I would argue all this that must begin with activities that start in your brain *before* your body reacts to them, to eventually lead that *"impulse to vigorous action."*

Student: But then, why should your body react to them at all?

The expressions of rage that James depicts (including that clenching of teeth and flushing of face) could have served in primordial times to help to repel or intimidate the person or creature that one is angry with; indeed, any external expression of one's mental state can affect how someone else will think. This suggest an idea about what we mean when we use our most common *emotion*

words; they refer to classes of mental conditions that produce external signs which make our behaviors more predictable to the persons with whom we are dealing. Thus for our ancestors, those bodily signs served as useful ways to communicate such so-called 'primary' emotions as anger, fear, sadness, disgust, surprise, curiosity, and joy.

Student: Perhaps this could also be because our most common emotions evolved long ago when our brains were simpler. Then there were fewer levels between our goals and our sensory-motor systems.

The body and face could also serve as a simple sort of memory: those states of mind might soon fade away—except that those body-expressions could help to maintain them by sending signals back to the brain. In that respect, William James might be right: without such 'mind-body' feedback loops, those 'cold and neutral' mental states might not persist for long enough to grow into larger-scale cascades. For your external expressions of anger serve not only to frighten your enemies, but also ensure that *you will stay frightened for long enough to carry out some actions that might save your life.*

For example, your face might display an expression of horror—even when no one else is present—when you realize that you left the door unlocked, or forgot to turn the oven off, or that something that you believed was false. After all, you need your body to stay alive, so, given that it is always at hand, it makes sense for your brain to exploit it as a dependable external memory device.

When we are young, we find it hard to suppress those external expressions, but eventually we learn to control most of them, to at least some degree, so that our neighbors can't always see how we feel.

Student: If those physical symptoms are not essential parts of emotions, then how can we make a distinction between our emotional states and our other ways to think?

It is hard to make that distinction clear, partly because we have so many names for our various emotional states, whereas most of our many other ways to think (such as those we described in §7–4) do not have popular names at all. Presumably this is because we don't yet have good ways to think about them. However, here is an ancient but still useful view of what distinguishes the mental conditions that we tend to describe as emotional:

Aristotle: The emotions are all those feelings that so change men as to affect their judgments, and that are also attended by pain or pleasure.
    —Rhetoric, Book II.

In a modern version of this, some psychologists talk about *'Valence,'* which refers to the extent to which one's attitude toward some *thing or situation is generally positive or negative.*[138] Similarly, there is a popular view in which we think of emotion and thoughts as complementary, in much the same way that an object's color and shape can change independently; we thus can think of each object (or idea) as having various 'matter of fact' or neutral aspects that, somehow, are also 'colored' by additional characteristics that seem to make it attractive, exciting, or desirable—versus disgusting, dull, or repulsive.

More generally, our language and thoughts are filled with distinctions like 'positive vs. negative' and 'rational vs. emotional." Such pairs are so useful in every day life that it's hard to imagine replacing them (any more than we should discard the idea that the sun rises and sets each day and night, because this is technically incorrect). However, if our goal is to understand our minds, we'll have to outgrow many dumbbell ideas.

In particular, exaggerating the body's role in emotions can lead to serious misconceptions. Do the talents of pianists reside in their fingers? Do artists see with talented eyes? No: there is no evidence to suggest that any of those body-parts think; it's the brain that sits in the driver's seat. Ask Steven Hawking or Christopher Reeve.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§7-7. Poincare's Unconscious Processes

"We cannot kindle when we will
The fire which in our heart resides,
The spirit bloweth and is still,
In mystery our soul abides:
But tasks in hours of insight will'd,
Can be through hours of gloom fulfill'd"
—Matthew Arnold

Sometimes you'll work on a problem for hours or days, as when Joan worked on her progress report.

She has been thinking about it for several days, but has not yet conceived of a good enough plan. Discouraged, she sets those thoughts aside … but then an idea 'occurs' to her.

But did Joan really set those thoughts aside, or did they continue in other parts of her mind? Hear a great mathematician recount some similar experiences.

Henri Poincare: *"Every day I seated myself at my worktable, stayed an hour or two, tried a great number of combinations and reached no results."*[139]

Most persons might get discouraged with this—but Poincare was inclined to persist:

"One evening, contrary to my custom, I drank black coffee and could not sleep. Ideas rose in crowds; I felt them collide until pairs interlocked, so to speak, making a stable combination. By the next morning … I had only to write out the results, which took but a few hours."

Then he describes another event in which his thinking seemed much less deliberate:

"The changes of travel made me forget my mathematical work. Having reached Coutances, we entered an omnibus to go some place or other. At the moment when I put my foot on the step the idea came to me, without anything in my former thoughts seeming to have paved the way for it. … I went on with a conversation already commenced, but I felt a perfect certainty."

This suggests that the work was still being pursued, hidden away in 'the back

of his mind'—until suddenly, as though 'out of the blue,' a good solution 'occurred' to him.

"There was one [obstacle] however that still held out, whose fall would involve the whole structure. But all my efforts only served at first the better to show me the difficulty. ... [Some days later,] going along the street, the solution of the difficulty that had stopped me suddenly appeared to me. ... I had all the elements and had only to arrange them and put them together."

In the essay from which these quotations come, Poincare concluded that when making his discoveries, he must have used activities that typically worked in four stages like these:

Preparation: Activate resources to deal with this particular type of problem.
Incubation: generate many potential solutions.
Revelation: recognize a promising one.
Evaluation: verify that it actually works.

The first and last of these stages seemed to involve the kinds of high-level processes that we characterized as conscious ones—whereas incubation and revelation usually proceed without our being aware of them. Around the start of the 19th century, both Sigmund Freud and Henri Poincare were among the first to develop ideas about 'unconscious' goals and processes—and, if only for mathematical activities— Poincare suggested clearer descriptions of these but borrowed

Let's consider what might be involved in each of the stages of such a process.

Preparation: To prepare to solve a specific problem, one first may need to 'clear one's mind' from other goals— for example, by taking a walk, or by finding a quiet place to work. Then one must focus on the problem by deliberating to decide which of its features are central enough to suggest an appropriate Way to Think; here Poincare said, *"All my efforts only served at first the better to show me the difficulty."*

Then, he suggest, you need to find appropriate ways to represent the situation; one needs to identify the parts of a puzzle before you can start to put them together—and until you understand their relationships well enough, you will tend to waste too much of your time at making bad combinations of them. This must be what Matthew Arnold meant when he said,

"This creative power works with elements, with materials; what if it has not those materials, those elements, ready for its use? In that case it must surely wait

till they are ready."

> *—Essays in Criticism, 1865.*

In other words, blind "trial and error" won't often suffice; you need to impose the right kinds of constraints and activate a set of resources that will tend to generate good possibilities—or else get lost in an endless search. Also, if you can't deal with the problem all at once, then you make a plan that breaks it into smaller parts that you can hope to handle separately.

Incubation: Once the 'unconscious mind' is prepared, it can consider large numbers of combinations, searching for ways to assemble those fragments to satisfy the required relations. Poincare wonders whether we do this with a very large but thoughtless search—or if it is done more cleverly.

Poincare: *"If the sterile combinations do not even present themselves to the mind of the inventor ... does it follow that the subliminal self, having divined by a delicate intuition that [only certain] combinations would be useful, has formed only these, or has it rather formed many others which were lacking in interest and have remained unconscious?"*

In other words, Poincare asks how selective are our unconscious thoughts; do we explore massive number of combinations, or work on the finer details of fewer ones? In either case, when we incubate, we will need to switch off enough of our usual Critics to make sure that the system will not reject too many hypotheses. However, we still know almost nothing about how our brains could conduct such a search, nor why some people are so much better at this: here is one conjecture about that.

Aaron Sloman:*"The most important discoveries in science are not discoveries of new laws or theories, but the discovery of new ranges of possibilities, about which good new theories or laws can be formed."*[140]

Revelation: When should incubation end? Poincare suggests that it continues until some structure is formed *"whose elements are so harmoniously disposed that the mind can embrace their totality while realizing the details."* But how does that subliminal process know when it has found a promising prospect?

Poincare: *"It is not purely automatic; it is capable of discernment; it has tact, delicacy; it knows how to choose, to divine. What do I say? It knows better how to divine than the conscious self, since it succeeds where that has failed."*

He conjectures that this ability to detect promising patterns seems to involve

such elements as symmetry and consistency.

Poincare: *"What is it indeed that gives us the feeling of elegance in a solution, in a demonstration? It is the harmony of the diverse parts, their symmetry, their happy balance; it is all that introduces order, all that gives unity, that permits us to see clearly and to comprehend at once both the ensemble and the details."*

Poincare did not say much more about how those detectors of 'elegance' might work, so we need more ideas about how we recognize those signs of success. Some of those candidates could be screened with simple matching tricks. Also, as part of the Preparation phase, we select some specialized critics that can detect progress toward solving our problem, and keep these active throughout Incubation.

Evaluation: We often hear advice that suggests that it's safer for us to trust our 'intuitions'—ideas that we get without knowing how. But Poincare went on to emphasize that one cannot always trust those 'revelations.'

Poincare: *"I have spoken of the feeling of absolute certitude accompanying the inspiration ... but often this feeling deceives us without being any the less vivid, and we only find it out when we seek to put on foot the demonstrations. I have especially noticed this fact in regard to ideas coming to me in the morning or evening in bed while in a self-hypnagogic state."*

In other words, the unconscious mind can make foolish mistakes. Indeed, later Poincare goes on to argue suggest that it often fails to work out the small details—so when Revelation suggest a solution, your Evaluation may find it defective. However, if it is only partially wrong, you may not need to start over again; by using more careful deliberation, you may able to repair the incorrect part, without changing the rest of that partial solution.

I find Poincare's scheme very plausible, but surely we also use other techniques. However, many thinkers have maintained that the process of creative thinking cannot be explained in *any* way, because they find it hard to believe that powerful, novel insights could result from mechanical processes—and hence require additional, magical talents.[141] However, Chapter 8 will argue that outstanding abilities can result from nothing more than fortunate combinations of certain traits that we find in the ways that most people think. If so, then what we call 'genius' requires no other special ingredient.

Somewhat similar models of thinking were proposed in Hadamard (1945), Koestler (1964), Miller (1960), and Newell and Simon (1972)—the latter two in more computational terms. Perhaps the most extensive study of ways to generate

ideas is that of Patrick Gunkel at http://ideonomy.mit.edu. In any case, however you make each new idea, you must quickly proceed to evaluate by activating appropriate critics. Then, if the result still has some defects, you can apply similar cycles to each of those deficiencies.

In my view, what we call 'creativity' is not the ability to generate completely novel concepts or points of view; for a new idea to be useful to us, we must be able to combine it with the knowledge and skills we already possess—so it must not be too very different enough from the ideas we're already familiar with.

Collaboration.

We usually think about thinking as a solitary activity that happens inside a single mind. However, some people are better at making ideas, while others excel at refining them—and wonderful things can happen when 'matched pairs' of such persons collaborate. It is said that T.S. Eliot's poetry owed much to Ezra Pound's editing, and that A.S. Sullivan's music was most inspired when he was working with W.S. Gilbert's librettos. Another example of such a pair might be Konrad Lorenz and Nickolaas Tinbergen, as we see in their Nobel Prize autobiographies:

Niko Tinbergen: "From the start 'pupil' and 'master' influenced each other. Konrad's extraordinary vision and enthusiasm were supplemented and fertilized by my critical sense, my inclination to think his ideas through, and my irrepressible urge to check our 'hunches' by experimentation — a gift for which he had an almost childish admiration."[142]

Konrad Lorenz: "Our views coincided to an amazing degree but I quickly realized that he was my superior in regard to analytical thought as well as to the faculty of devising simple and telling experiments. ... None of us knows who said what first, but it is highly probable that the [concept of] innate releasing mechanisms ... was Tinbergen's contribution."[143]

For many people, thinking and learning is largely a social activity—and many of the ideas in this book came from collaborations with students and friends. Some such relationships are productive because they combine different sets of aptitudes. However, there also are pairs of partners who have relatively similar skills—perhaps the most important of which are effective tricks for preventing each other from getting stuck.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

Do we normally think 'Bipolarly'?

The processes that Poincare described involved cycles of searching and testing in which problems are solved over hours, days, or even years. However, many events of everyday thinking persist for just a few seconds or less. Perhaps these, too, begin by spawning ideas, then selecting some promising ones, and then dwelling on their deficiencies!

If so, then a typical moment of commonsense thinking might begin with a very brief 'micro-manic' phase. This would produce an idea or two—and then a short 'micro-depressive' phase would quickly look for flaws in them.

If those phases took place so rapidly that your reflective systems don't notice them, then each such micro-cycle' would seem to be no more than a moment of everyday thinking, while the overall process would seem to you like a steady, smooth, uneventful flow.[144]

The quality of such systems would depend in part on how much time one spends in each such phase. It seems plausible to conjecture that, when one is inclined to be 'critical' or 'skeptical,' one spends less time at *Incubation* and puts more effort into *Evaluation*. However, if anything were to go badly wrong with how those durations were controlled, then some of those phases might last for so long that (as suggested in §3-5) they might appear as symptoms of a so-called 'manic-depressive' disorder.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§7-8. Cognitive Contexts

No matter what you are trying to do, there will be other things trying to get your attention. Some of these can be ignored, but if some are subgoals of what you are trying to do, this may require you to switch to some other Way to Think that uses different resources and bodies of knowledge. Then, once you have accomplished those subgoals, you will need to return to your previous job—but to avoid repeating what already was done, you must have retained some information about these aspects of your previous state of mind:

Your previous goals and priorities,
The representation you used for them,
The bodies of knowledge you had engaged,
The sets of resources that were active then,
The Selectors and Critics that were involved.

This means that our larger-scale model of mind needs places for keeping

such sets of records. Let's give these the name of "Cognitive Contexts." Without them, every 'train of thought' would be disrupted whenever we were interrupted. In simpler brains, it might suffice to maintain no more than a single such memory, but for looking several steps ahead—or for managing larger subgoal trees—we'd need special machinery to enable us to rapidly switch among several remembered contexts.

More generally, because every hard problem keeps forcing us to switch among several different Ways to Think, a typical 'present mental state' must actually be part of a larger panalogy that can fluently navigate among several different points of view. In popular folk-psychology, we simply imagine all that stuff to be stored in our "short-term memories"— as though we could put such things into a box and take them out whenever we want.

However, we still do not know very much about how those "working" memories operate; there is evidence that such records are stored in various different forms and locations in different parts of our brains—but we still know little about how those various brain centers actually work. So here we shall simply assume that those short-term records are stored in some place that we'll call the "Context Box."[145]

If you asked Joan what she was thinking about, she might mention the subject of tidying-up, and further questions would show that she is maintaining several different representations of the kinds of changes she is planning to make. Furthermore, for her to be able to switch among these, she must be able to store and retrieve various kinds of structures like these:

Her current collection of subgoal trees.
Some records of recent external events.
Some descriptions of recent mental acts,
Her presently active fragments of knowledge,
Simulations she uses to make her predictions.

This means that Joan's context box for 'tidying up' must keep track of

various aspects of that task:

Context Selector for
'Tidying-Up'
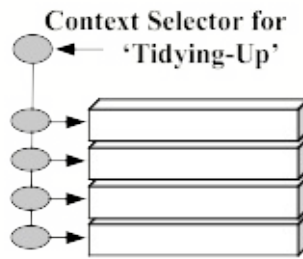
Also, of course, other topics and subjects have been 'on her mind' for longer spans of time, so she'll need to keep track of several of these:

Why would we need special systems like this to keep track of multiple contexts? It seems quite natural to us that after any brief interruption—for example, to answer a question that someone has asked, or to pick up a tool that you have just dropped—we can usually get back to what we were doing without needing to start all over again. It is the same when we interrupt *ourselves*—for example, to attend to a subgoal of a task, or briefly to think in some different way. When such a diversion is small and brief, this causes little trouble because it leaves most of our active resources unchanged.

However, a larger-scale change could cause more disruption and result in wasted time and confusion. So, as we evolved more ways to think, we also evolved machinery for more quickly returning to previous contexts. In everyday life we simply say that we're using our 'short-term memories'— but any good theory of how that might work must answer questions like these:

How long do recent records persist, and how do we make room for new ones? There must be more than one answer to that because various parts of the

brain must work in somewhat different ways. Some memories may be permanent, while others may rapidly fade away, unless they happen to get *"refreshed."* Such records also would get erased if stored in a 'place' that has a limited size—because then each new item would have to replace some records that are already there. Indeed, this is how modern computers work: whenever data is created or retrieved, it is first stored in a *cache*—a device that has been designed to be especially quickly accessible. Then, whenever such a *cache* gets full, its oldest records get erased—although some of them may have been copied to larger, more permanent memory boxes.

How do some memories become permanent? There is evidence that it takes hours or days for what we call *short-term* memories to be converted to longer-term ones. Older theories about this assumed that frequent repetitions made the original record more permanent. However, it seems more likely to me that new memories are briefly maintained in resources that act like a computer's cache—and then, over time, more permanent versions are created in other regions of our brains. See §8-4.

In any case, some memories seem to last for the rest of one's life. However, this could be an illusion because they might need 'refreshment' from time to time. Thus, when you recall a childhood memory, you often also have the sense of having remembered the same thing previously; this makes it hard for you to know whether you have retrieved an original record or merely a later copy of it. Worse yet, there now is ample evidence that those records can be changed while they're being refreshed.[146]

How do we retrieve old memories? We all know that our memories often fail—as when you try to recall some important details but find that their records have disappeared, or that at least we cannot retrieve them right now. Now clearly, if no trace of that record remains, further search would be a futile quest. Nevertheless, we frequently manage to find some clues that we can use to reconstruct more of those memories. Here is a very old theory of this:

St. Augustine: *"But what happens when the memory itself loses something, as when we forget anything and try to recall it? ... Perhaps the whole of it had not slipped out of memory; but a part was retained by which the other lost part was sought for, because the memory realized that it was not working so smoothly as usual, hence, it demanded the restoration of what was missing. For example, suppose we see or think of some man we know, and, having forgotten his name, try to recall it—but some other thing presents itself, which was not previously associated with him; then this is rejected, until something comes into the mind which better conforms with our knowledge."*

—Book 10 of *Confessions,* 427 AD.

So if you can link a few of those fragments together, you may be able to reconstruct a good deal more—

"... by gathering together those things that the memory already contains but in an indiscriminate and confused way, and now putting them together [so that] where they formerly lay hidden, scattered, or neglected, they now come easily to present themselves to the mind which is now familiar with them."[147]

Augustine soon turned to other concerns, and concluded this discussion of memory by plaintively asking, *"Who will work this out in the future?"* But more than a thousand years were to pass before further progress on theories of memories.

How many thoughts can you think at once?

How many feelings can you feel at once? How many different things can you simultaneously 'pay attention' to? How many contexts can be active at once in your context-box? To what extent can you be aware of how many mental activities?

The answers to such questions depend on what we mean by 'aware' and 'attention.' We usually think of 'attention' as positive, and highly regard those persons who are able to 'concentrate' on some particular thing, without getting distracted by other things. However, we could also see "attention" as negative—because not all our resources can function at once—so there always is a limit to the range of things we can think about at the same time. Nevertheless, we can train ourselves to overcome at least some of those built-in constraints. [See *§§Attention.*]

In any case, in our high level thinking we can only maintain a few different 'trains of thought' before we start to become confused. However, at our lower reactive levels, we carry on hundreds of different activities. Imagine that you are walking and talking among your friends while carrying a glass of wine:[148]

Your grasping resources keep hold of the cup.
Your balancing systems keep the liquid from spilling.
Your visual systems recognize things in your path.
Your locomotion systems steer you around those obstacles.

All this happens while you talk, and none of it seems to require much thought. Yet dozens of processes must be at work to keep that fluid from spilling out—while hundreds of other systems work to move your body around. Yet few

of these processes 'enter your mind' as you roam about the room—presumably because they use resources that work in separate realms that scarcely ever come into conflict with what you are usually "thinking about."

It is much the same with language and speech. You rarely have even the faintest sense of what selects your normal response to the words of your friends, or which ideas you choose to express—nor of how any of your processes work to group your words into phrases so that each gets smoothly connected to the next. All this seems so simple and natural that you never wonder how your context-box keeps track of what you have already said—as well as to whom you have mentioned them.

What limits the number of contexts that a person can quickly turn on and off? One very simple theory would be that our context-box has a limited size, so there is only a certain amount of room in which to store such information. A better conjecture would be that each of our well-developed realms acquires a context box of its own. Then, some processes in each of those realms could do work on their own, without getting into conflicts until when they have to compete for the same resources.

For example, it's easy to both walk and talk because these use such different sets of resources. However, it is much harder to both speak and write (or to listen and read) simultaneously, because both tasks will compete for the same language-resources. I suspect such conflicts get even worse when you think about what you're thinking about, because every such act will change what is in the context box that keeps track of what you were thinking about.

At our higher Reflective levels, our representations span many scales of time and space, and our current Self-Representations can range from thinking "I'm holding this cup" to "I am a Mathematician," or "I am a person who lives on the Earth," or sometimes, perhaps, only a little more than a constant, vague sense of 'being aware.' To be sure, a person may also have the impression of thinking all these simultaneously, but I suspect that these are constantly shifting; our sense of thinking them all at once comes from the "Immanence Illusion" of §4-1, because the contents of our various Context-Boxes are so rapidly accessible.

What Controls the Persistence of Processes?

> Edmund Burke: "He that wrestles with us strengthens our nerves and sharpens our skill. Our antagonist is our helper. This amicable conflict with difficulty obliges us to an intimate acquaintance with our object and compels us to consider it in all its

Whatever you're trying to think about, you have other concerns that compete with it—and each should persist for long enough to justify the effort and time it will cost to switch them on and off. Still, everyone knows such feelings as these:

"I've been spending so much time on this problem that I am losing my motivation; besides, it has gotten so complex that I simply cannot keep track of it; perhaps I should quit and do something else."

When none of the methods we've tried have worked, how much longer should we persist? What decides when we should quit—and lose whatever investment we've spent? We always have at least some concern with how we conserve our materials, energy, money, and friends—and each such concern would seem to suggest that we have some Critics that detect when that particular element may be getting into short supply, and then suggest ways conserve or replenish it. Such critics would lead us to think, *"I'm doing too many things at once,"* or *"I can't afford to buy both of these,"* or *"I don't want to lose my friendship with Charles."*

The simplest way to conserve your time is to abandon the goals that consume too much of it. But renouncing goals will often conflict with your ideals, as when they are things that you've promised to do, or that others already expect you to do. Then you might also have to suppress those values, or even regard them as handicaps—but going against your high-level ideals can lead to cascades that you recognize as tension, guilt, distress, or fear—along with the shame and humiliation we talked about earlier. So making such decisions can thus cause you to become "emotional."

Citizen: But certain, well-disciplined persons seem able to set such emotional feelings aside, and simply do what seems "rational." Why do most of us people find this so hard to do?

It seems to me that it is a myth that there exists a 'rational' way to think. One is always comparing various goals, and deciding which ones to put aside, but the apparent merits of those alternatives will always depend on other aspects of your mental state.

In any case, each Way to Think will be useless unless can persist for long enough to make some progress. To do this, it will need at least some ability to

keep other processes from stopping it, and this could be done to some extent by controlling which of your Critics are working now. Let's consider a few extremes of this.

What if your set of active Critics does not change? *Then you would be likely to keep repeating the same approach because, after each attempt to change your way to think, those Critics would try to switch you back again—and you might get stuck with a 'one-track mind.'*

What if some Critics stay on all the time? *Certain Critics must always be active to make us react to serious hazards—but if these are not selected carefully, it could lead to obsessive behaviors by repeatedly making you focus too sharply only on certain particular subjects.*

What if all your Critics get turned off*? Then all your questions would seem to be answered because you are no longer able to ask them, and all your problems would seem to be gone because nothing seems to have any flaws.*

Everything may seem wonderful during such a 'mystical experience'—but such 'revelations' usually fade when enough of your critics get turned back on.

What if too many Critics are active at once? *Then you'd keep noticing flaws to correct, and spend so much time repairing them that you would never get any important things done. And if you find ugliness everywhere, your friends may perceive you as depressed.*

What if too many Critics are turned off*? If you can ignore most alarms and concerns, that would help you to 'concentrate'—but it also might lead you to ignore errors and flaws in your arguments. However, the fewer Critics you activate, the fewer goals you will try to pursue, and then you would tend to be mentally dull.*

Then what should decide which ones are active? *Your thinking would become chaotic if too many goals were to freely compete without any larger-scale management—but if and particular Way to Think persisted too long, you would appear to have a 'one-track' mind.*

Chapter §9 will argue that control over which of our Critics are active must never be too highly centralized, because sometimes we need to concentrate—yet still respond to emergencies. Also, consider what might happen if large classes of Critics turned off, and then on, for excessive durations of time: then there would be long cycles in which you would first be euphoric, when nothing would ever seem to be wrong, followed by intervals in which no goal would seem to be worth pursuing. In such cases, the Critics that normally help us to think could play a role, when they're poorly controlled, in what we call manic-depressive

disorders.

Questions

Which of our ways to think are inborn? Which of them are not innate, but are ones that each child eventually learns from its experience with its environment? Then do certain individuals go on to discover special techniques that lead them to yet better ways to think? We'll discuss this in *§8-8 Genius.*

What determines the urgencies of our goals? What keeps track of the tasks that we have postponed? Are there clocks or timers in our brains that schedule or otherwise regulate our higher-level activities? We'll talk about this in Chapter §9.

How many things can we think about at once, and in how many different realms? How many different contexts can we manage to keep active? How are such activities distributed among the billions of cells in our brains?

How does Context affect how we think? We'll come back to this in Chapter §10.

Part VIII

§8-1. Resourcefulness

> *"Although machines can perform certain things as well as or perhaps better than any of us, they infallibly fall short in others, from which we may discover that they did not act from knowledge, but only from the arrangements of their parts."*
> —Descartes, in Discourse on Method, 1637.

We are all accustomed to using machines that are stronger and faster than people are. But before the first computers appeared, no one could see how any machine could do more than a very few different things. This must be why Descartes insisted that no machine could be as resourceful as any person can be.

"For while reason is a universal instrument which can apply to every situation, a machine's parts need a particular arrangement for each particular action; therefore it is impossible for a single machine to have enough diversity to enable it to act in all the events of life in the same way as our reason causes us to act."[149]

Similarly in earlier times there appeared to be an unbridgeable gap between the capacities of humans and other animals. Thus, in *The Descent of Man*, Darwin observes that,*" Many authors have insisted that man is divided by an insuperable barrier from all the lower animals in his mental faculties."* However, he then contends that this difference may be just *"one of degree and not of kind."*

Charles Darwin: *"It has, I think, now been shewn that man and the higher animals, especially the primates ... all have the same senses, intuitions, and sensations, — similar passions, affections, and emotions, even the more complex ones, such as jealousy, suspicion, emulation, gratitude, and magnanimity; ... they possess the same faculties of imitation, attention, deliberation, choice, memory, imagination, the association of ideas, and reason, though in very different degrees."*[150]

Then Darwin observes that *"the individuals of each species may graduate in*

intellect from absolute imbecility to high excellence," and argues that even the highest forms of human thought could have developed from such variations—because he sees no particular point at which that would meet an intractable obstacle.

"That such evolution is at least possible, ought not to be denied, for we daily see these faculties developing in every infant; and we may trace a perfect gradation from the mind of an utter idiot ... to the mind of a Newton."

Many people still find it hard to envision how there could have been transitional steps from animal to human minds. In the past, that view was excusable—because few thinkers had ever suspected that *only a few small structural changes could vastly increase what machines can achieve.* However, in 1936, the mathematician Alan Turing showed how to make a "universal" machine that can read the descriptions of other machines—and then, by switching among those descriptions, it can do all the things that those machines can do.[151]

All modern computers use this trick, so today we can use the same machine to arrange our appointments, edit our texts, or help us send messages to our friends. Furthermore, once we store those descriptions *inside* the machine, then those programs can change themselves—so that the machine can keep extending its own abilities. This showed that the limits which Descartes observed were not inherent in machines, but resulted from our old-fashioned ways to build or to program them. For each machine that we built in the past had only way to accomplish each particular task—whereas each person, when stuck, has alternatives.

Nevertheless, many thinkers still maintain that machines can never achieve such feats as composing great theories or symphonies. Instead, they prefer to attribute such feats to inexplicable 'talents' or 'gifts.' However, those abilities will seem less mysterious, once we see how our resourcefulness could result from having such diverse ways to think. Indeed, each previous chapter of this book discussed some way in which our minds provide such alternatives:

§1. We are born with many kinds of resources.
§2. We learn from our Imprimers and friends.
§3. We also learn what we ought not to do.
§4. We can reflect upon what we are thinking about.
§5. We can predict the effects of imagined actions.
§6. We use huge stores of commonsense knowledge.
§7. We can switch among different Ways to Think.

This chapter discusses yet additional features that make human minds so versatile.

§82. We can see things from many points of view.
§83. We have special ways to rapidly switch among these.
§84. We have developed special ways to learn very quickly. Move the
§85. We have efficient ways to recognize which knowledge is relevant.
§86. We can keep extending the range of our ways to think.
§87. We have many different ways to represent things.

At the start of this book, we noted that it is hard to conceive of ourselves as machines, because no machine that we've seen in the past seemed to understand the meanings of things, but could only react to the simple commands that we designed them to execute. Some philosophers argue that this must be because machines are merely material things, whereas meanings exist in the world of ideas, which lies outside the realm of physical things. However, Chapter §1 suggested that we, ourselves have constrained our machines by defining those meanings so narrowly that we fail to express their diversity:

If you 'understand' something in only one way then you scarcely understand it at all—because when something goes wrong, you'll have no place to go. But if you represent something in several ways, then when one method fails, you can switch to another. That way, you can turn things around in your mind to see them from different points of view —until you find one that works for you!

To show how this kind of diversity makes human thinking so versatile, we'll start with examples of the multiple ways we use to estimate our distance from things.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§8-2. Estimating Distances

Why has not man a microscopic eye?
For this plain reason, man is not a fly.
Say what the use, were finer optics giv'n,
T' inspect a mite, not comprehend the heav'n?

—Alexander Pope (in Essay on Man)

When you're thirsty, you look for something to drink—and if you notice a

nearby cup, you can simply reach out to pick it up—but if that cup lies further away, then you will have to move over to it. *But how do you know which things you can reach?* A naïve person sees this as no problem at all because, *"You just look at a thing and you see where it is."* But when Joan detected that oncoming car in §4-2 or grasped that book in §6-1, *how did she know its distance from her?*

In primeval times we had to guess how near our predators were to us; today we only need to judge if we have enough time to cross the street—but, still, our lives depend on this. Fortunately, we each have many different ways to estimate the distance to things.

For example, you know that a typical cup is about as large as your hand. So if a cup fills as much of the scene as does your outstretched hand

, then you can reach it from where you stand. Similarly, you can judge how far you are from a typical chair, because you know its approximate size.

However, even when you don't know an object's size, you still have ways to estimate its distance from you. For example, if you can assume that two things are of similar size, then the one that looks smaller is further away. Of course, that assumption may be wrong, if one of those objects is a small model or toy. And also, whenever two objects overlap, then the one in front must be closer to you, regardless of its apparent size.
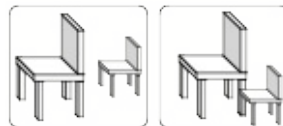
You can also get spatial information from how the parts of a surface are lighted or shaded, and from an object's perspective and context. Again, such clues are sometimes misleading; the images of the two blocks below are identical, but the context suggests that they have different sizes.
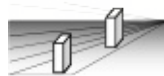
If you assume that two objects lie on the same level surface, then the one that looks higher lies further away. Also, finer-grained textures look further away, and so do things that look hazier.
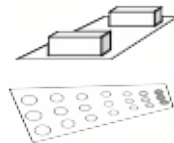
You can also judge a distance to an object by difference in its images in your two eyes—either by the angles between those two images or by the small 'stereoscopic' disparities between those slightly different images.
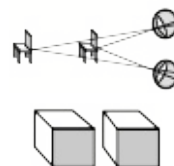
Also, if an object is moving, then the closer it is to you, the faster it will appear to move. You also can estimate its range by how you must change the focus of the lens of your eye.
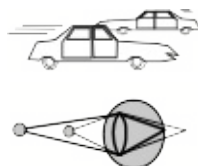
Finally, aside from all these perceptual schemes, one frequently knows

where some object is, without using any vision at all—because, if you've seen a thing in the recent past, its location is still in your memory!

Student: Why would we need so many different methods, when surely just two or three would suffice?

In almost every waking minute, we make hundreds of judgments of distance, and yet we scarcely ever fall down the stairs, or accidentally walk into doors. Yet each of our separate ways to estimate distance has many different ways to fail. Focusing works only on nearby things—and many persons can't focus at all. Binocular vision works over a longer range, but quite a few people are unable to compare the images in their two eyes. Some methods fail when the ground isn't level, and texture and haze are not often available. Knowledge only applies to objects you know, and an object might have an unusual size—yet we scarcely ever make fatal mistakes because we can use so many different techniques.

But if every method has virtues and faults, how do we know which ones to trust? The next few sections will discuss some ideas about how we manage to so quickly switch among so many different ways to think.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§8-3. Panalogy

The previous section emphasized how many different techniques we could use to accomplish the same objectives — mainly to know how far away some Object is. However, it would not help us very much to have so many methods available, unless we also had some way to switch among them almost effortlessly. This section will suggest a particular kind of machinery that, I suspect, our brains might use to do such switching almost instantly.

In Chapter 6 we mentioned that when you read the sentence, *"Charles gave Joan the Book,"* this can cause you to interpret 'book' in several different realms of thought: as an object, possession, or storehouse of knowledge.
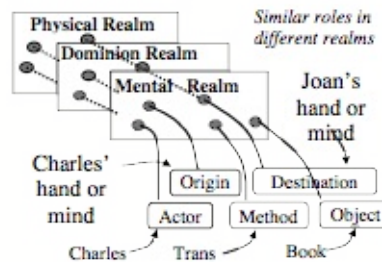
However, having multiple representations won't help you much unless you use the context to rapidly switch to the appropriate meaning.

Student: Why would you actually need to switch? Why can't you maintain all those meanings at once?

There are always limits to how many things a person can do simultaneously. You can touch, hear, and see things concurrently because those processes use different parts of the brain. But few of us can draw two different things with both hands, simultaneously—presumably, because these compete for resources that can do only one of those things at a time. This section will suggest how our brains could quickly switch between different meanings.[152]

Whenever you walk into a room, you expect to see the opposite walls, but you know that you will no longer see the door through which you entered that room.
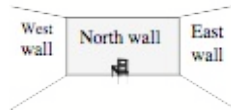
Now walk to the West wall that is now to your left, and turn yourself to face to the right; then you will be facing toward the East.
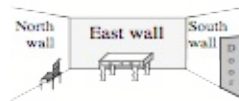
The South wall has now come into view, and the West wall now is in back of you. Yet although it now is out of sight, *you have no doubt that it still exists.* What keeps you from believing that the South wall just now began to exist, or that the West wall has actually vanished? This must be because you assumed all along that you are in a typical, box-like room. So, of course you knew just what to expect: all four sides of that room will still exist.

Now consider that each time you move to another place, every object you that you have seen may now project a different shape on the retinas in the back of your eyes—*and yet those objects do not seem have changed?* For example, although the visual shape of that North wall has changed

, you still see it as rectangular. What makes those meanings remain the same? [153] Similarly, you now see an image of that chair in which it appears to have turned around

—but you usually don't even notice this, because your brain knows that it is *you* who has moved and not the chair. Also, you now can see the door that you entered through—yet none of this surprises you!

What if you next turn right to face the South? Then the North wall and chair will disappear, and the West wall will re-enter the scene—just as anyone would expect.
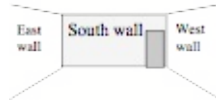
You are constantly making these kinds of predictions without any sense of how your brain keeps dealing with that flood of changing appearances: *How do you know which things still exist? Which of them have actually altered their shapes? Which of them have actually moved? How do you know you're still in the same room?*

Student: Perhaps those questions do not arise because we're seeing those objects continuously. If they suddenly changed we'd notice this.

In fact, our eyes are always darting around, so our vision is far from continuous.[154] All this evidence seems to suggest that, even before you entered that room, you have already, somehow, assumed a good deal of what you were likely to see.

"The secret is that sight is intertwined with memory. When face to face with someone you newly meet, you seem to react almost instantly—but not as much to what you see as to what that sight "reminds" you of. The moment you sense the presence of a person, a world of assumptions are aroused that are usually true about people in general. At the same time, certain superficial cues remind you of particular people you've already met. Unconsciously, then, you will assume that this stranger must also resemble them, not only in appearance but in other traits as well. No amount of self-discipline can keep those superficial similarities from provoking assumptions that may then affect your judgments and decisions."

　　—Section §241 of *SoM*.

What would happen if every time you moved, you had to re-recognize every object in sight? You would have to re-guess what each object is, and get evidence to support that conjecture. If so, then your vision would be so unbearably slow that you would be virtually paralyzed! But clearly, this is not the case, because:

"When we enter a room, we seem to see the entire scene almost instantly. But, really, it takes time to see—to apprehend all the details and see if they confirm our expectations and beliefs. Our first impressions often have to be revised. Still, how could so many visual cues so quickly lead to consistent views? What could explain the blinding speed of sight?"[155]

Answer: we don't need to constantly 'see' all those things because we build virtual worlds in our heads. Hear one of my favorite neurobiologists:

William H. Calvin: "The seemingly stable scene you normally "see" is really a mental model that you construct—the eyes are actually darting all around, producing a retinal image as jerky as an amateur video, and some of what you thought you saw was instead filled in from memory."[156]

We construct those mental models so fluently that we feel no need to ask ourselves how our brains make them and put them to use. However, here we need a theory about why, when we move, the objects around us seem to remain in place. When first you see the three walls of that room, you might have represented them with a network like this:
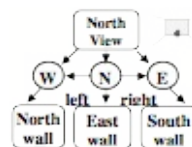
However, even before you entered that room, you expected it to have *four* walls—and already knew how to represent a *'typical box-like four-walled room.'* Consequently, you 'assumed by default' that its edges, corners, ceiling, and floor would be parts of *a larger, non-moving framework that doesn't depend on your present point of view.* In other words, the 'reality' that we perceive is based on mental models in which things don't usually change their shapes or disappear, despite their changing appearances. We mainly react to what we expect—and tend to represent what we see as 'things' that remain in their places.[157]
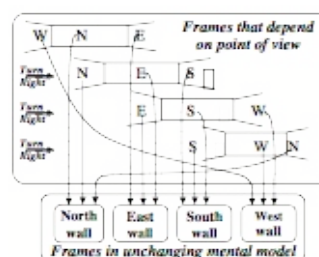
If you use this kind of larger-scale structure, then as you roam about that room, you can store each new observation in some part of that more stable framework. For example, if you represent that chair as *near* the North wall, and the door as *part* of the South wall, then these objects will have a fixed 'mental place'—regardless of where you were when you noticed them—and those locations will stay the same even when those objects are out of sight. (Of course this could lead to accidents, if an object was moved without your knowing it!)

For vision, this shows how the space that surrounds us would seem to stay the same when we see it from different views—by linking features in different realms to similar roles in a larger-scale frame. For language, in §6-1 we saw how this method could make *"Charles gave Joan the Book"* seem to keep a single meaning when we interpret it in different realms. We introduced the term "Panalogy" to describe such schemes for linking analogous features together.
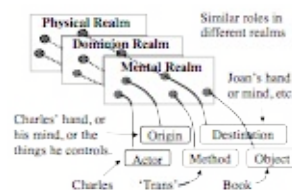
Student: How do we make new panalogies? How hard are they to build and maintain? Is the talent for making them innate or learned? Where do we place them in our brains? And what happens when one of them makes a mistake?

We rarely make an entirely new idea, because, before you make records of anything new, you are likely already to have recalled some similar object or incident—so then, we will usually copy and modify some structure that we already have. This is extremely useful because, otherwise, we would have no way to get access to that 'new idea,' or know which old skills to apply to it. Also, if that older concept already belongs to a panalogy, we can add the new idea as an additional leaf; then it will inherit the techniques by which that older is retrieved and applied.

For example, you can think of a chair as a physical structure whose parts consist of a back, seat and legs. In that physical view, the chair's legs support its seat, and both of these support the chair's back.
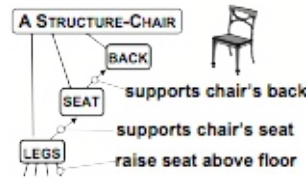
However, you can also think of a chair as a way to make people feel comfortable. Thus the chair's seat is designed to support one's weight, the chair's back serves to support one's back, and the chair's legs support one up to a height designed to help a person to relax.
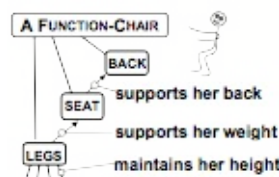
Similarly, you could also regard that very same chair as an item of personal property, or as a work of art or of carpentry—and each of those various contexts or realms could lead you to represent chairs in different ways. Then, when your present way to think fails, your critics could tell you to switch to another.
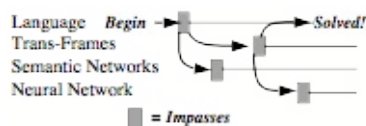
However, that switching might cost you a good deal of time—if you were forced to start over. But, if you have grouped those several representations into a panalogy, by linking their similar features together, then you'll be able more quickly to switch. Indeed, I suspect that much of what we know has been organized into panalogies—so that we can see an automobile as a vehicle, or as a complex mechanical object, or as a valuable possession. We can see a city as a place for people to live, as a network of social services, or as an object that

requires a lot of water, food, and energy. Chapter §9 will argue that, whenever you think about your Self, you are reflecting about a panalogy of mental models of yourself.

From where do our panalogies come, and how and when do we develop them? I suspect that the architecture of our brains has evolved so that we tend to link every fragment of knowledge we learn to similar ones that we already know, in analogous aspects of multiple realms—and that we do this so automatically that it seems to require no reasoning.[158]

Student: But wouldn't that make you mistake whatever you see for something else that it reminds you of? You would always be confusing things.

Yes, and we're constantly making those kinds of 'mistakes'—but although this may seem paradoxical, that actually helps to keep us from being confused! For if you saw each object as totally new, then it would have no meaning to you —and you would have no ideas about what to do. However, if each new item gets linked to some older ones—as when a new chair reminds you of previous ones—then you will know some things you could do with it.

If our memories mainly consist of panalogies, then most of our thinking will have to deal with ambiguous representations of things. However, this is a virtue and not a fault because much of our human resourcefulness comes from using analogies.

<center>∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞</center>

§8-4. How does Human Learning work?

The word 'learning' is useful in everyday life—but when we look closely we see that it includes many ways that our brains can change themselves. To understand how our minds grow, we would need ideas about how people learn such different skills as how to build a tower or tie a shoe, or to understand what a new word means, or how a person can learn to guess what their friends are thinking about. If we tried to describe all the ways in which we learn, we'd find ourselves starting a very long list that includes such methods as these:

Learning by adding new If-Do-Then rules, or by changing low-level connections,

Learning by making new subgoals for goals, or finding better search techniques,

Learning by changing or adding descriptions, statements and narrative stories.

Learning by changing existing processes.

Learning to prevent mistakes by making Suppressors and Censors.

Learning to make better generalizations.

Learning new Selectors and Critics that provide us with new Ways to Think,

Learning new links among our fragments of knowledge.

Learning to make new kinds of analogies.

Learning to make new models, virtual worlds, and other types of representations.

As our children develop, they not only learn particular things, but they also acquire new thinking techniques. However, there is no evidence that, by itself, an infant could ever invent enough such things. So, perhaps the most important human skill is to learn, not only from one's own experience, but also *to learn from being told things by others*. However, long ago, the philosopher Hume raised a yet more fundamental question, namely of why learning is possible at all:

David Hume: *"All inferences from experience suppose, as their foundation, that the future* will *resemble the past, and that similar powers will be conjoined with similar sensible qualities. If there be any suspicion that the course of nature may change, and that the past may be no rule for the future, all experience becomes useless, and can give rise to no inference or conclusion."*[159]

In other words, learning itself can only work in a suitably uniform universe.

But still we need to ask how learning works. In particular, the following section will ask how a person can learn so much from seeing a single example.[160]

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

How do we Learn so Rapidly?

No other creatures come close to us in being able to learn so much—and we do this with astonishing speed, as compared to other animals. Here's an example that illustrates this:

Jack saw a dog do a certain trick, so he tried to teach it to his own pet, but Jack's dog needed hundreds of lessons to learn it. Yet Jack learned that trick from seeing it only once. How did Jack so quickly learn so much—although he has only only seen one instance of it?

People sometimes need long sessions of practice, but we need to explain those occasions in which we learn so much from a single experience. However, here is a theory which suggests that Jack does indeed need many repetitions— but he does them by using an 'animal trainer' inside his head, which he uses to train other resources inside his brain, in much the same ways that he would use to teach his pet!

To do this, Jack could use a process like the Difference-Engine in §6-3. It begins with a description *D* of that trick, which is now in his 'short-term memory.' Then Jack's 'mental animal trainer' can work to produce a copy *C* of this in some other, more permanent place—by repeatedly altering *C* until there remains no significant difference between it and D. Of course, if *D* has many intricate parts, this cycle will need many repetitions.[161]
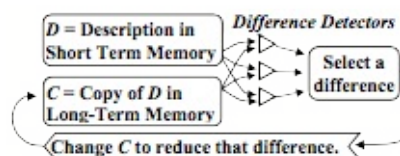
A "Mental Animal Trainer"

So this 'animal-trainer' theory suggests that when a person learns something new, they *do* need multiple repetitions. However, we are rarely aware of this, perhaps because that process goes on in parts of the brain that our reflective

thinking cannot perceive.

*Student: Why can't we simply remember **D** by making that short-term memory more permanent—in the same place where it now is stored? Why should we have to copy it to some other, different place?*

We are all familiar with the fact that our short-term memories are limited. For example, most persons can repeat a list of five or six items, but when there are ten or more items, then we reach for a writing pad. I suspect that this is because each of our fast-access 'memory boxes' is based on such a substantial amount of machinery that each brain includes only a few of them—and so, we cannot afford to use them up. This would answer the question our student asked: each time we made those connections more permanent we would lose a valuable short-term memory box!

This could account for the well-known fact that whatever we learn is first stored temporarily—and then it may take an hour or more to convert it into a more permanent form. For example, this would explain why a blow to the head can cause one to lose all memory of what happened before that accident. Indeed, that 'transfer to long-term memory' process sometimes can take a whole day or more, and often requires substantial intervals of sleep.[162]

What could be the different between our short-and long-term memory systems? One answer to that appears to be our short-term memory systems use certain short-lived chemicals, so that those memories will quickly fade unless unless those chemicals keep being refreshed; in contrast, we have good evidence that long-term memories depend on the production of longer-lived proteins that make more permanent connections between the cells of the brain.[163]

It probably is no coincidence that modern computers evolved in a somewhat similar pattern: at every stage of development, fast-acting memory boxes were much more costly than slower ones. Consequently, computer designers invented ways to confine those expensive but fast-acting units into devices called 'caches' that only store data that is likely soon to be needed again. Today, a modern computer has several such caches that work at various different speeds, and the faster each is, the smaller it is. It may be the same inside our brains.

Here are a few other reasons why our memory systems may have evolved to require so much time and processing.

Retrieval: When one makes a memory record, it would make no sense to store this away without providing some ways to retrieve it. This means that each record must also be made with links that will activate it when relevant (for example, by linking each new memory to some other existing panalogy).

Credit Assignment: A record of how one solved a problem would be

unlikely to have much future use if it applied only to that particular problem. We'll discuss this more in §8-5.

The 'Real-Estate' problem for Long-term memories. How could an 'animal-trainer' find room in the brain for the copy that it is trying to make? How could it find appropriate networks of brain cells to use, without disrupting connections and records that one would not want to erase? So, finding places for new memories may involve complex constraints and requirements, and this could be a reason why making permanent records takes so much time.

Copying complex descriptions. It is easy to imagine ways to record a simple list of symbols or properties, but I have never seen any plausible schemes for how a brain could quickly make copies of structures with more complex connections. That is why this section proposed to use a sequential, difference engine-like scheme.[164]

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

Learning by building "Knowledge-Lines"

Suppose that you've just had a 'good idea' or solved a certain hard problem P. What, then, should you learn from this experience? One thing you could do is to construct a new rule: **If** the problem you're facing is like problem P, **Then** *try the solution that worked on P.* That solution will probably help you with problems that closely resemble P—but it is unlikely to help with less similar problems.

However, instead of just recording your solution to P, you could make a record of the *Way To Think* that used were using when you found that solution! Generally, one would expect this to help with a much wider range of other problems. But how you could make copy of anything like your entire state of mind? Clearly, that would never be practical—but you might get a good approximation to it if you could, later, re-activate a substantial portion of the resources that were active when you solved problem P.
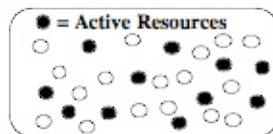
This suggests that to remember the method you used to find the solution to P, you could simply construct a new Selector that activates that set of resources. We call this kind of structure a "K-line,"
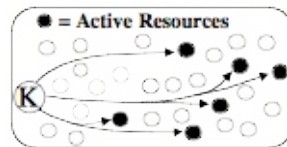
One can see such a K-line as a sort of 'snapshot' of a mental state because, when you later activate it, that will put you into a similar state, which should help to solve other problems like P. Here is an analogy that illustrates how K-lines work:

Kenneth Haase: "You want to repair a bicycle. Before you start, smear your hands with red paint. Then every 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.' If you use different colors for different jobs, some tools will end up marked with several colors. [...] Later, when you have some job to do, just activate the set of tools with the right color for that kind of job, and the resources that you've used for such jobs then become available." [See SoM 8.1]

This way, for each kind of problem or job, your K-lines can fill your mind with ideas that might be relevant—by getting you into a mental state that resembles one that, in the past, helped with a similar job.

Student: How could this help you do anything new, if it only re-uses a way to think that you already knew how to use—when you used it to solve that old problem?

Whenever you work on a novel problem, you may start up several K-lines at once, and this is likely to activate some resources that conflict with one another. Then, this will lead to a cascade in which some resources will get suppressed and some other resources will be aroused—and now you'll be using a somewhat new set of resources, and your state of mind won't be quite the same as any that you have been in before. Thus, every new situation is likely to lead to a somewhat novel Way to Think—and if you make a 'snapshot' of *that,* you will now have a K-line that differs from all of your previous ones.[165]

Student: I see how this new K-line could be used as a Selector for a new Way to Think. But how would you build a new Critic to recognize when to activate it?

If we want to use that new K-line for problems like P, then a simple such Critic would recognize some combination of features of P.
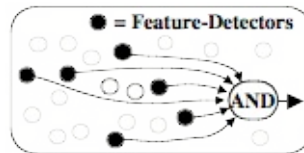
However, such a Critic will rarely be useful if it requires *all* the features of P, because then it will not recognize situations that are slightly different that P. Also, each new K-line should include only resources that actually helped.

Student: I see what you mean. Suppose that when you were fixing that bicycle, at some point you tried to use a tool that happened to make the problem worse. It wouldn't be good to paint that tool red because, then, later you would waste more time again.

This suggests that when we make new Selectors and Critics—or more generally, whenever we learn—we should try to make sure that *what* we learn will mainly consist of things that are likely to help. The following section discusses some processes that might help to ensure that what we learn will be relevant in future times.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§8-5. Credit-Assignment

When first we met Carol in chapter §2, she learned to use spoons for moving fluids. But then we asked about which aspects of her several attempts should get credit for her final success:

Should her learning include which shoes she wore, or the place in which those events occurred, or whether the weather was cloudy or clear? Which of the thoughts she was thinking then should be recorded in what she remembers? What if she smiled while using that fork, but happened to frown when using that spoon? What keeps her from learning irrelevant rules like, "To fill a cup, it

helps to frown"?[166]

Some early ideas about how animals learn were based on schemes in which each reward for a success will cause a small "reinforcement" of some connections in the animal's brain—while every disappointment will cause a corresponding weakening. In simple cases, this can lead a brain to select the right features to recognize. [See §§§Reinforcement] However, in more complex situations, such methods will not work so well to find which features are relevant —and then we'll need to think more reflectively.

Some other theories about how learning works assumed that this consisted of making and storing new *If-Do* reaction-rules. This could be one reason why Jack's dog in §84 needed so many repetitions: perhaps, each time that dog attempted that trick, it made a small change some *If* or some *Do*—but then, it only recorded that change in the case that it got a reward.

That might work well in a simple case. However, in more complex situations, this kind of learning is likely to fail if the *If* of any new *If-Do* rule turns out to describe too few details (because then that rule will be applied too recklessly)—or if that *If* includes too many details (because then it may never apply again, because no two situation are ever exactly the same.) The same applies to the *Do* of that rule—and so, each *If* and *Do* must be just abstract enough to make it apply to 'similar' situations—but not to too many dissimilar ones. Otherwise, Jack's dog might need a new rule for every posture that it might be in. All this means that those old 'reinforcement' schemes might work well enough for some animals, but it seems unlikely they could be help much to explain how humans learn more complicated things.

This brings us back to that question about how *a person can learn so quickly, without doing so many repetitions.* Earlier we suggested that we actually do many repetitions, but that these go on inside our minds, and cannot be seen by a person outside. But here we'll take another view in which we use higher-level processes to decide what to learn from each incident—because, to understand what you have done, you need to reflect on your recent thoughts. Here are a few of the processes that these 'credit-assignments' might involve.[167]

Choosing how to represent a situation will affect which future ones will seem similar.

Unless you select only the parts of your thinking that helped, you may learn too many irrelevant things.

It helps to do mental experiments to discover which features were relevant, by varying some of their elements.

Each new fragment of knowledge or skill must be connected so that you can access it when it is relevant.

The better those decisions are made, the more you will benefit from each experience. Indeed, the quality of those processes could be important aspects of the suitcase of traits that people call "intelligence." However, merely recording solutions to problems will help us only to solve somewhat similar problems, whereas if we can record *how we found* those solution*s*, that could further enable us to deal with much broader classes of situations.

For example, in playing a game like checkers or chess, if you should happen to win a game, you won't gain much by simply recording the moves that you made— because you're unlikely ever again to encounter those same situations. However, you can do better if you can learn which of your higher-level decisions helped to reach those winning positions. For, as Allen Newell observed fifty years ago,

Allen Newell: "It is extremely doubtful whether there is enough information in "win, lose or draw," when referred to the whole play of the game [so, for learning to be effective], each play of the game must yield much more information. ... If a goal is achieved its subgoals are reinforced: if not they are inhibited. ... Every tactic that is created provides information about the success or failure of tactic search rules; every opponent's action provides information about success or failure of likelihood inferences and so on."[168]

Thus, when you finally achieve a goal, you should assign some credit for this to the higher-level method you used to divide that goal into subgoals. Instead of just storing solutions to problems, you thus can refine the strategies you used to discover those solutions.

Student: But then you'd also want to remember the strategies that led to those methods—and you've started a process that never will end!

There is no clear limit to how long one could dwell on what might have led to a certain success. Indeed, such realizations are sometimes delayed for minutes, hours or even days; this suggests that some of our credit-assignments involve extensive searches that go on in other parts of our minds.

For example, we sometimes have 'revelations' like, *"Now I see the solution to this,"* or *"I suddenly see just why that worked!"* But as we saw in §7-7, we cannot assume that those problems were solved during those wonderful moments of time, because of being unaware of the unconscious work that preceded them. If so, then such a moment may merely celebrate the times at which some Critic

has said, *"This has taken so long that it's time to stop— and to adopt the tactic already considered which, at this moment would seem the best."*[169]

We usually make our Credit Assignments without much reflection, but sometimes one may say to oneself, after completing some difficult job, *"It was stupid of me to waste all that time, when I knew how to do it all along."* To keep from making that error again, we need to modify our way to retrieve that particular fragment of knowledge—or perhaps make some change in some Critic that failed to remind us of it.

Similarly, one sometimes may ask, *"How did I manage to solve that hard problem?"* or *"What were the critical steps in that process?* Of course, we cannot always find answer to those, because it may be harder to understand how one found the solution than it was to solve the problem. Nevertheless, such questions suggest that our credit-assignments sometimes depend on high-level reflections.

In any case, if we want to understand how people learn, we will need more research on such questions as what kinds of credit assignments infants can make, how children develop better techniques, how long such processes persist, and the extent to which we can learn to control them. In chapter §9 we will also discuss how our feelings of pleasure might relate to how we make our credit-assignments.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

Transfer of Learning to other realms. Every teacher knows the frustration that comes when a child learns something to pass a test, yet never applies that skill to anything else. What makes certain children excel at "transferring" the things they learn to other, different realms—whereas other children seem to need to relearn the same ideas in each domain?

It would be easy just to say that some children are 'more intelligent'—but that would not help us to explain how they use their experiences to make more helpful generalizations. This could partly be because they are better at making and using panalogies. But also, as we have just seen, the better our ways to describe each event, the more we can learn from each incident. Indeed, those 'smarter' children may have come to learn more efficiently because they have learned to reflect (perhaps unconsciously) about how their own learning processes work—and then found ways to improve those processes. For example, such reflections may lead to better ideas about which aspects of things they *ought* to learn.

It seems clear that the qualities of our thoughts must depend, to a large

extent, on how well we make our credit-assignments. For those must be among the processes we use to make our most significant generalizations. This means that persons who do not learn to make good credit-assignments would be likely to show deficiencies in their ability to apply what they learn to new situations. This is what psychologists call *'Transfer of Learning'.*[170]

This section has argued that, to gain more from each experience, it would not be wise for us to remember many details of each situation—but only those aspects that were relevant to our goals. Furthermore, what we learn can be yet more profound, if we assign the credit for our success, not only to the final act that led to our failure or success —or even to the strategy that led to it—but to whatever earlier choices we made that selected our winning strategy. Perhaps our unique abilities to make such high-level credit-assignments accounts for many of the ways in which we surpass our animal relatives.

<p style="text-align:center">∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞</p>

§8-6. Creativity and Genius

> *The best way to have a good idea is to have lots of ideas.*
> *—Linus Pauling*

We admire our Einsteins, Shakespeares, and Beethovens—and many people insist that their accomplishments are inspired by "gifts" that no one could ever explain. If so, then machines could never do such things because (at least, in the popular view) no machine could hold any mysteries.

However, when one has the fortune to meet one of those persons that we portray as "great," one finds no single, unusual trait that seems to account for their excellence. Instead (at least it seems to me) what we find are unusual combinations of otherwise common ingredients.[171]

They are highly proficient in their fields.　　　*(But by itself we just call this expertise.)*

They have more than usual self-confidence.　　　*(Hence better withstand the scorn of peers.)*

They often persist where others would quit.　　　*(But others may just call this stubbornness.)*

They accumulate more ways to think.　　　*(But then they'll need better*

ways to switch.)

They habitually think in novel ways   *(But so do others, albeit less frequently.)*

They often reflect on their goals and ideals.   *(Or are less reluctant to modify them.)*

They have better systems for self-control.   *(So they waste less time on irrelevant goals.)*

They reject many popular myths and beliefs.   *(Especially about what cannot be achieved.)*

They tend to keep thinking more of the time.   *(They spend less effort at wasting their minds.)*

They excel at explaining what they've done.   *(So their work is less likely to fade from neglect.)*

They tend to make better credit-assignments.   *(So they learn more from less experience.)*

Everyone has some share of each such trait, but few develop so many of them to such unusually great extents.

Citizen: Each of those traits might help to explain how regular people solve everyday problems. But surely there be something unique about such great thinkers as Feynman, Freud, and Asimov.

Here is a statistical argument against the belief that genius comes from singular gifts or characteristics:

Suppose that there were, say, 20 traits that might help to make someone exceptional, and assume that each person has an even chance to excel at each particular one. Then we'd expect only one in each million persons to excel at all of those 20 traits.

But statistics alone never tell us the reasons! For even if that argument were correct, it would shed no light at all upon either the nature of those variations or why just those particular people develop so many of those traits. For example, perhaps to acquire so many such qualities, *a person must first develop some unusually good ways to learn.* In any case, there is plenty of solid evidence that, to a significant extent, many of our mental traits are genetically inherited. However, I suspect that yet more important are the effects of fortunate mental accidents.

For example, most children discover various ways to arrange their toy blocks into columns and rows—and if observers praise what they've done, those

children may go on to refine those new skills. Then, some of those children may also go on to play at *discovering new ways to think.* However, no outside observer can see those mental events, so those children will have to learn by praising successes inside their own minds. This means that when such a child does remarkable things, outsiders may see no clear cause for this—and will tend to describe that child's new skills with terms like *talents, endowments, traits,* or *gifts.*

The psychologist Harold McCurdy suggested another 'fortunate accident' that could bring out exceptional traits in a child—namely to have been born with exceptional parents.

Harold G. McCurdy: *"The present survey of biographical information on a sample of twenty men of genius suggests that the typical development pattern includes these important aspects: (1) a high degree of attention focused upon the child by parents and other adults, expressed in intensive educational measures and usually, abundant love; (2) isolation from other children, especially outside the family; (3) a rich efflorescence of fantasy [i.e. creativity] as a reaction to the preceding conditions."*[172]

It would also appear that outstanding thinkers must have developed some effective techniques that help them to organize and apply what they learn. If so, then perhaps those skills of 'mental management' should get some credit for what we perceive as the products of genius. Perhaps, once we understand such things, we'll be less concerned with teaching particular skills and more teaching children how to develop those more generally powerful mental techniques.

Citizen: But can we really hope to understand such things? It still seems to me that there is something magical about the ways in which some people imagine completely new ideas and creations.

Many phenomena seem magical until we find out what causes them. In this case, we still know so little about how our everyday thinking works, that it would premature to assume that there is a real difference between "conventional" and "creative" thought. Then why would we cling to the popular myth that our heroes must have inexplicable 'gifts'? Perhaps we're attracted to that idea because, if those achievers were *born* with all their wonderful tricks, we would share no blame for our deficiencies—nor would those artist and thinkers deserve any credit for their accomplishments.

However, let's turn this discussion upside down, and change our question to ask instead, what could cause one person to become *less* resourceful than another one. Here is one condition that could have the effect of hindering further

mental growth:

The Investment Principle: *If you know of two ways to achieve some goal, you'll usually start with the one you know best—and then that method may gain so much additional strength that you'll tend to use the second one less—even if others have told you that that the second is the better one.*

Thus, sometimes the problem is simply that for one to develop a new way to think, one may have to endure the discomfort of many awkward or painful performances. So, one 'secret of creativity' may be to develop the knack of enjoying that sort of unpleasantness! We'll explore this more in chapter §9, when we talk about Adventurousness.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§8-7. Memories and Representations

"There is no property absolutely essential to one thing. The same property, which figures as the essence of a thing on one occasion, becomes a very inessential feature upon another."
—William James [Principles, Chap. XXII.]

Everyone can imagine things; we hear words and phrases inside our minds. We envision conditions that don't yet exist—and then exploit those images to predict the effects of possible actions. Much of our human resourcefulness must come from our abilities to manipulate mental models and representations of objects, events, and other conceptions.

But what do we mean by a *model* or a *representation?* As I suggested in §4-5, I am using those words to talk about any structure inside one's brain that one can use to answer some questions about some subject. Of course, those answers will only be useful when your 'model' or 'representation' behaves enough like the model's subject does—for the purposes that concern you now.

We sometimes use actual physical objects to represent things, as when we use a picture or map to help us find paths between parts of a city. However, to answer a question about a past event, we must use what we call our "memories."

But what do we mean by a *memory?* Each memory must be some kind of record or trace that you made at the time of that prior event—and, of course, you cannot record an event itself. Instead, your brain can only make some records about some of the objects, ideas, and relationships that were involved in that incident. (Indeed, you cannot record an idea itself—and so the best that you can do is to record some aspects of your mental state.)

For example, when you hear a statement like, *"Charles gave Joan the book,"* you might represent that incident with a script-like sequence of *If-Do-Then rules*:

Charles extends arm. Joan extends arm. Charles retracts arm. Joan retracts arm.

C. holds book. Both hold book. J. holds book.
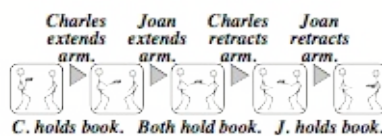
However, you also may have wondered about whether that book was a gift or a loan, or did Charles want to ingratiate Joan, or was merely disposed to help a friend. You might have envisioned how the actors were dressed, or some of the words they might have said. Then you might have made several representations for that incident, perhaps including:

A verbal description of that incident.
A visual simulus of the scene.
Some models of the persons involved.
Simulations of how those persons felt.
Analogies with similar incidents.
Predictions about what might happen next.

Why would your brain represent the same event in so many different ways? Perhaps each realm of thought that was engaged left an additional record or trace in some different network inside your brain. This will enable you, later, to use multiple ways to think about that same incident—for example, by using verbal reasoning, or by manipulating mental diagrams, or by envisioning the actors' gestures and facial expressions.

Today, we still know little about how our brains make those memory traces or how they later retrieve and 'replay' them. We do know a lot about how separate brain-cells behave, but we do not yet have good explanations of about how our larger-scale columns and networks of cells manage to represent past events. Nor do our self-reflections help; although as we saw in §8-4, this must involves complex processes, nevertheless it seems to us that we simply 'remember' what happens to us.

In any case, one cannot record an event itself, but one can only make some descriptions of how that event affected one's mental state. Some earlier sections of this book discussed some structures that could used to represent such information. The following section will review some of these, and then speculate about how such structures might be arranged in our brains.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

Multiple Ways to Represent Knowledge

This section reviews some structures that researchers have used to represent knowledge inside computers. I will have to leave out most smaller details (many of which are discussed in chapter 8, 19, and 24 of *The Society of Mind)*. Some non-technical readers might do well to skip this section.

Narrative Scripts: Perhaps our most familiar way to represent an incident is to recount it a story or script that depicts a sequence of events in time—that is, in the form of a story or a narrative. The previous section described such a script for the sentence, *"Charles gave Joan the book,"* and we saw a similar one in §5-3 for Carol's plan about how to build an arch:

A sequence-script of If-Do-Then Rules

Semantic Networks: However, when we need to describe more details, such as the relations between an object's parts, it may be better to use the kinds of 'semantic networks' we saw in §4-6 to represent a person's self-model, and in §5-8 to represent the structure of a physical book.
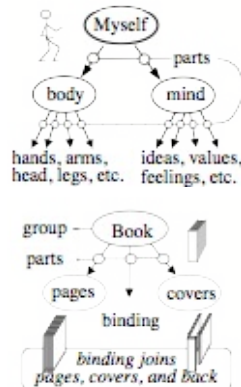
Semantic Networks for 'Person' and 'Book'

Trans-Frames: To represent the effects of an action, it is convenient to use pair of semantic networks to represent what was changed. This is what we did in §5-8 to imagine replacing the top of an arch. This way, one only needs to change the name of a single relationship—instead of altering thousands of points to change a visual picture-like image.
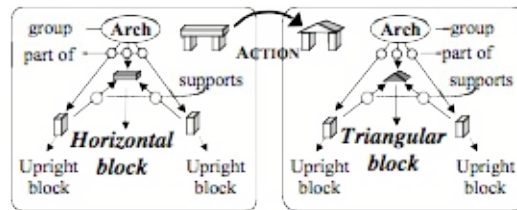
A Trans-Frame for changing the top of an arch

We use the term "Trans-Frame" to name such a pair that represents the conditions before and after some action was done. Then we also can represent the effect of a sequence of actions by linking together a chain of the Trans-Frames to form a story or narrative. Here is a sequence of trans-frames for giving a book:
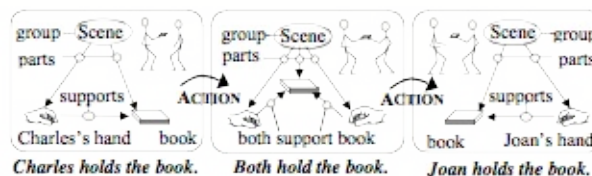
Such a sequences can describe a script that includes any further details that one might need.
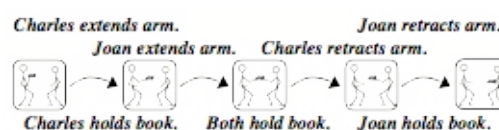
A Script for transferring a Book

Each of these types of representation can answer certain types of questions—but what could enable computers to produce answers so quickly as human brains do? When someone says 'apple,' you seem to almost instantly know that a typical apple grows on a tree, is round and red, is about the size of a human hand, and has a certain texture, flavor and taste—yet almost no time seems to elapse between hearing that word and then becoming aware of such things.

To explain how that information could so quickly appear, I conjecture that much of such knowledge is wrapped into structures called *Frames*. The simplest type of Frame consists of nothing more than a labeled list of some properties of particular object, and you can think of this kind of list as like a printed form that has 'blanks' or 'slots'—each of which can be filled-in with a link to some fragment of knowledge. Then, when you know which slot to inspect, you can quickly retrieve that fragment of knowledge, without need much time to search for it.
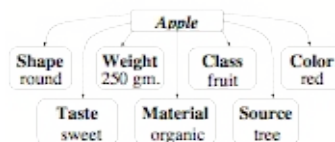
A Frame for an Apple's Properties

Default Assumptions: A valuable feature of a typical frame is that every slot comes already filled in with some 'default' or typical value for it. Then you can use such a value to make a good guess whenever you don't have a more definite answer. For example, you might assume 'by default' that an apple is red —but if your particular apple is green, then you will replace 'red' by 'green' in its color slot. In other words, a typical frame describes a stereotype whose 'default assumptions' are usually right—but which we can easily change or replace when we encounter exceptions to them.[173] This would seem to be an important aspect of how we do commonsense reasoning.

Picture-Frames: Every slot of a property list is directly connected to the name of that frame, as in the list above for describing an Apple. However, other more complex kinds of frames may have more connections between their various slots. For example, to represent some view of a room, we could use what we call a "picture frame" to represent each wall of that room by in terms of a few large spatial regions as shown below. Then each such region can have some

links that describe the objects close to that part of the wall, as well as some links to other nearby regions. This kind of structure would allow us to do a good deal of commonsense spatial reasoning. [See §§24 of SoM.]

Frames for Including Additional Slots: It makes sense to allow each Frame to include some additional slots for representing knowledge that is not already described by the networks contained inside that frame. For example, here is a Trans-frame for Joan's trip to New York:
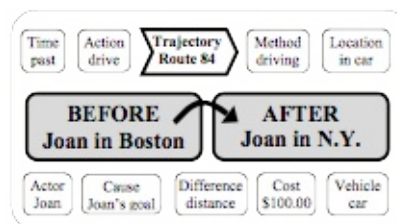
This frame includes two semantic networks that describe the situations *Before* and *After* that trip was taken. However, it also contains other slots that describe when, how and why Joan took that trip. Then the default assumptions included in those slots can supply additional knowledge for answering such questions as these.

Where did that action occur and when? Who or what caused it to happen?
Was it intentional or not? What purposes was it intended to serve?
What devices or tools were used? What were its other side effects?
Which resources did it engage? What was expected to happen next?

This suggests an explanation of how we quickly use our commonsense knowledge—without any sense that we're doing it: it is an example of the "Immanence Illusion" that we described in §4-3.1. As soon as you activate such

a frame, then many questions that you might otherwise ask will already be answered before you can ask them—because they are among the default values of that frame's slots. For example, if you heard that Charles was holding a book, you would not stop to ask why he was holding it; you would simply assume that he has the most usual goal for which *any person holds anything*—namely, to keep it from falling to the floor.[174]

Connectionist and Statistical Representations.

Student: This book suggests some ideas about how high-level knowledge-based systems could come to achieve things like human commonsense reasoning. But why were no such systems built in the past?

Work on such systems almost came to a stop in the 1980's because most researchers recognized that that this would need ways to acquire and to organize millions of fragments of commonsense knowledge. That prospect seemed so daunting that most researchers decided to search for simpler alternatives. This led to many attempts to design some single process that would somehow evolve whatever it needed—along with learning all the knowledge it would need by interacting with the external world. Some of these "baby machines" did learn to do some useful things (such as to recognize various kinds of patterns) but as we noted in Chapter §6, none of them went on to develop more higher-level reflective ways to think.

Why were none of those "Baby Machines" able to keep extending their abilities? It appears to me that this failure came mainly because most of their designers decided that *their systems should represent the knowledge they were to acquire mainly in numerical terms.* Consequently, most of those 'baby machines' were designed to use the techniques called *Neural Networks, Perceptrons, Fuzzy Logic systems,* and *"Statistical Learning Programs."* All such systems represent knowledge about the relations between things by assigning numerical 'weights' or 'strengths' to connections inside a network of nodes. Such a representation might look like this:
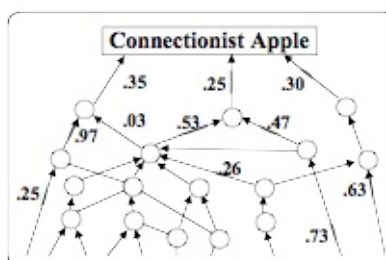
Here we see only one kind of link, which reduces every type of relationship to a single numerical value or 'strength.' The trouble with this is that a single numbers is 'opaque' in the sense that it has so little expressiveness. For, whenever one computes an average or a probability, this conceals the knowledge or evidence that led to it.[175] For, consider that if you only see the number 12, you cannot tell if that number represents 5 plus 7, or 9 plus 3, or 27 minus 15! Did it come from counting the eggs in a nest, or from counting the years of your grandchild's age? For example, if you represent the concept of 'apple' this way, your machine may be able to recognize an apple, but it won't be able to reason about it. In short, numerical representations become obstacles to using more reflective ways to think—because it is difficult for other, higher-level processes to think about the knowledge that such systems contain. [We'll discuss this more in §§§Opacity.]

Let's contrast this with representing a concept of "apple" by using a semantic network like this:
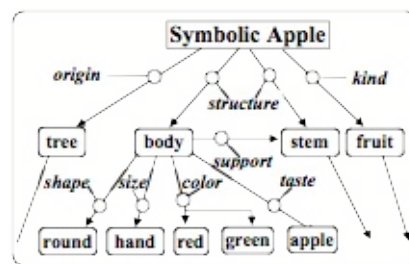
This kind of representation can help you answer many questions about an apple, such as where you can find one and what you can use it for—because a semantic network can express all sorts of different relationships, whereas numerical representations ultimately limit a system's mental growth, because they provides no parts that the rest of a mind can use to produce more elaborate explanations.

Micronemes for Contextual Knowledge. We always face ambiguities. The significance the things that you see depends on the rest of your mental context. This also applies to events in your mind, because what they mean depends on which mental resources are active then.[176] In other words, no symbol or object

has meaning by itself, because your interpretation of it will depend on the mental context you're in. For example, when you hear or read the word *block,* you might possible think that it means an obstacle to progress, a certain kind of rectangular object, a wooden board to chop things on, or a stand on which things in an auction are shown. Then which interpretation will you select?

Such choices will depend, of course, on the preferences that are active in your current mental context—which, somehow, this will dispose you to make selections from such sets of alternatives as these:

Conceptual or material.
Animal, mineral, or vegetable.
Well-established or speculative.
Common, rare, or irreplaceable.
Robust, fragile or reparability.
Indoors or outdoors.
Public or private.
Residence, office, theater, or car.
Urban, rural, forest, farm.
Color, texture, hardness, strength.
Irregular or symmetrical.
Hunting, gambling, entertainment.
Cooperation, conflict, etc.

Many contextual features like these have common names, but many others (such as aromas) have no such words. I have proposed to use the term *"micronemes"* for the myriad of nameless clues that color and shade how we think about things, and the diagram below suggests some machinery through which such contextual features could affect many mental processes. Imagine that the brain contains a bundle of thousands of wire-like fibers that pass through a great many of other structures inside that brain—so that the state of each of those 'micronemes' can influence many processes:
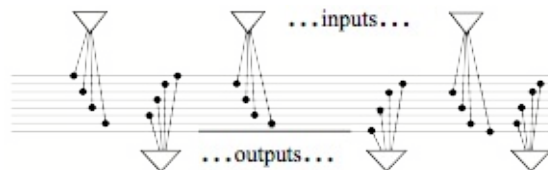
On the input side, we shall assume that many of your mental resources—such as K-lines, Frame-slots, or *If-Do-Then* rules— can alter the states of some micronemes. Then the present state of your micronemes could represent much of your current mental context—and the states of those fibers are changed, your far-reaching bundle of micronemes will broadcast that information to many other mental resources—so that this will change some of your attitudes, outlooks, and states of mind. In other words, this system could switch you into other, different ways to think. I think that this concept of micronemes could help to replace many old and vague ideas about 'association of ideas.' In §§Brain-Waves, we suggest more details about how such a system could play central roles in how our mental processes are organized.

A Hierarchy of Representations

The sections above have briefly described several kinds of structures that we could use to represent various types of knowledge. However, each of these representation types has its own virtues and deficiencies—so each of them may need some other connections through which they can exploit some of the other types of representations. This suggests that our brains need some larger-scale organization for interconnecting our multiple ways to represent knowledge. Perhaps the simplest such arrangement would be a hierarchical one like this:
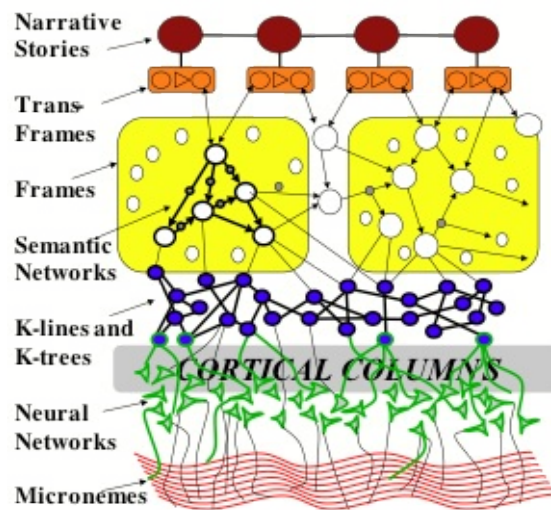
This diagram suggests how a brain might organize its ways to represent knowledge. However, we should not expect to find that actual brains are arranged in such an orderly way. Instead, we should not be surprised if anatomists find that different regions of the brain evolved somewhat different such organizations to support mental functions in different realms—such as for maintaining our bodily functions, manipulating physical objects, developing social relationships, and for reflective and linguistic processes.

This hierarchy of representation appear to roughly correspond to the various levels of thinking that were proposed in our previous chapters—in the sense that increasingly higher levels will tend to more depend on using story-and script-like representations. However, each of those levels itself may use several types of representations. In any case, even if this diagram turns out to be a good

description of the relations between those representations, it is unlikely to closely match the gross anatomy of the brain—because the structures shown in this diagram need be spatially close to each other. Indeed, a substantial volume of a human brain consists of bundles of nerves that interconnect regions that are quite far apart.[177]

How do we learn new Representations?

From where do we obtain our ways to represent knowledge, and why do we find it so easy to arrange our knowledge into panalogies? Are these abilities installed genetically into our infant of memory systems, or do we learn them individually from our experiences? These questions suggest a more basic one: how do we manage to learn at all? As Immanuel Kant pointed out long ago, *learning to learn* is one of the things that we cannot learn from experience!

Immanuel Kant: *"That all our knowledge begins with experience there can be no doubt. For how is it possible that the faculty of cognition should be awakened into exercise otherwise than by means of objects which affect our senses, and partly of themselves produce representations, partly rouse our powers of understanding into activity, to compare, to connect, or to separate these—and so to convert the raw material of our sensations into a knowledge of objects?"*

"But, though all our knowledge begins with experience, it by no means follows that all arises out of experience. For, on the contrary, it is quite possible that our empirical knowledge is a combination of that which we receive through impressions, and [additional knowledge] altogether independent of experience ... which the faculty of cognition supplies from itself, sensory impressions giving merely the occasion.[30]

So, although sensations give us *occasions* to learn, this cannot be what makes us *able* to learn, because we first must have the 'additional knowledge' that our brains would need, as Kant has said, to *"produce representations"* and then *"to connect"* them.[178] Such additional knowledge would also include inborn ways to recognize correlations and other relationships among sensations. I suspect that, in the case of physical objects, our brains are already innately endowed with machinery to help us *"to compare, to connect, or to separate"* objects so that we can represent them as existing in space.

All this leads me to suspect that we must be born with primitive forms of structures like K-lines, Frames, and Semantic Networks—so that that no infant needs to wholly invent the kinds of representations that we depicted above. However, I doubt that we're born with those stuctures complete, so it still

requires some effort and time for us to refine those primitive representations into their more adult forms. I hope that soon there will be some research on how that development process works.

Could any person ever invent an totally new kind of representation? Such an event must be quite rare because no type of representation would be useful without some effective skills for working with it— and a new set of such skills would take time to grow. Also, no fragment of knowledge could be of much use unless it is represented in a familiar way. For reasons like these, it makes sense to conjecture that most of our adult representations come either from refining our primitive ones, or by acquiring them from our culture. However, once a person has learned to use several different representations, then that person might be more able to invent new ones. Indeed, we see such skills in the work of those exceptional writers, artists, inventors, and scientist who repeatedly discover new and useful ways to represent things.

How should a brain proceed to select which representation to use? As we have emphasized several times, each particular kind of description has virtues and deficiencies. Therefore it makes more sense to ask, "Which methods might work well for the problem I'm facing—and which representations are likely to work well with those methods?"

Most computer programs still, today, can do only one particular kind of task, whereas our brains accumulate multiple ways to deal with each type of problem we face. However, once we better understand how to organize such resources, along with knowledge to help us decide which technique to use in each situation. To do this we need to develop a wide range of ways to represent those all those capabilities—so that, whenever the method we're using fails, we can switch to another alternative.[179]

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

Part IX. The Self

Could be
I only sang because the lonely road was long;
and now the road and I are gone
but not the song.
I only spoke the verse to pay for borrowed time:
and now the clock and I are broken
but not the rhyme.
Possibly,
the self not being fundamental,

What makes each human being unique? No other species of animal has such diverse individuals; each person presents a different set of appearances and abilities. Some of those traits are inherited, and some come that person's experiences—but in every case, each of us ends up with different characteristics. We sometimes use *'Self'* for the features and traits that distinguish each person from everyone else.

However, we also use *Self* in a sense which implies that all our activities are controlled by powerful creatures inside ourselves, who do our thinking and feeling for us. We call these our *Selves* or *Identities,* and sometimes we tend to think of them as like separate persons inside our heads.

Daniel Dennett: "A homunculus (from Latin, 'little man') is a miniature adult held to inhabit the brain ... who perceives all the inputs to the sense organs and initiates all the commands to the muscles. Any theory that posits such an internal agent risks an infinite regress ... since we can ask whether there is a little man in the little man's head, responsible for his perception and action, and so on."[180]

What attracts us to the queer idea that we can only think or feel with the help of those Selves inside our minds? Chapter §1 suggested some reasons for this:

Citizen: Whether or not you believe that Selves exist, I'm perfectly sure that there's one inside me, because I can feel it working to make my decisions."

Psychotherapist: The Single-Self legend helps makes life seem pleasant, by hiding from us how much we're controlled by goals that we'd rather not know

about.

Pragmatist: The Single-Self concept helps make us efficient by keeping our minds from trying to understand everything all the time.

The Single-Self view thus keeps us from asking difficult questions about our minds. If you wonder how your vision works, it answers that: *'Your Self simply peers out though your eyes.''* If you ask how your memory works, you get: *"Your Self knows how to recollect whatever might be relevant.''* And if you wonder what guides you through your life, it tells that your Self provides you with your desires and goals—and then solves all your problems for you. Thus, the Single-Self view diverts you from asking about how your mental processes work, and leads you to wonder, instead, about these kinds of questions about your Self:

Is an infant born with anything like what an adult would call a Self'? *Some would insist on answering with, "Yes, infants are persons just like us—except that they don't yet know so much.'' But others would take an opposite view: "An infant begins with almost no intellect, and developing one takes a sizeable time.''*

Does your Self have a special location in space? *Most 'western' thinkers might answer, "Yes"—and tend to locate it inside their heads, somewhere not far behind their eyes. However, I've heard that some other cultures situate Selves between the belly and chest.*

Which of your beliefs are your "genuine" ones? *The Single-Self view suggests that some of your many intents and views are your "sincere" and "authentic" ones—whereas the models of mind discussed in this book leave room for a person to hold conflicting values and attitudes.*

Does your Self stay the same throughout your life? *We each have a sense of remaining the same, no matter whatever may happen to us. Does this mean that some part of us is more permanent than our bodies and our memories?*

Does your Self survive the death of your brain? *Different answers to that might leave us pleased or distressed, but would not help us to understand ourselves.*

Each such question uses words like *self, we*, and *us* in a somewhat different sense—and this chapter will argue that this is good because, if we want to understand ourselves, we'll need to use several different such views of ourselves.

Whenever you think about your "Self," you are switching among a network of models,[181] each of which may help to answer questions about different aspects of what you are.

For example, some of our models are based on simplistic ideas like *"All our actions are based on the will to survive,"* or *"we always like pleasure more than pain,"* while some other self-models are far more complex. We develop these multiple theories because each of them helps to represent certain aspects of ourselves, but is likely to give some wrong answers about other questions about ourselves.

Citizen: *Why should a person want more than one model? Would it not be better to combine them into a single, more comprehensive one?*

In the past, there were many attempts to make 'unified' theories of psychology.[182] However, this chapter will suggest some reasons why none of those theories worked well by itself, and why we may need to keep switching among various different views of ourselves.

Jerry Fodor: "If there is a community of computers living in my head, there had also better be somebody who is in charge; and, by God, it had better be Me."[183]

Cosma Rohilla Shalizi: "I have been reading my old poems, and they were written by somebody else. Yet I am that selfsame person; or, if I am not, who is? If no one is, when did he die—when he finished this poem, or that one, or the next day, or the end of that month?"

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§9-1. How do we Represent Ourselves?

"O wad some Pow'r the giftie gie us
To see oursels as ithers see us!"

—*Robert Burns 1759-1796*

How do people construct their self-models? We'll start by asking simpler questions about how we describe our acquaintances. Thus, when Charles tries to think about his friend Joan, he might begin by describing some of her characteristics. These could include his ideas about:

The appearance of Joan's body and face,
The range and extents of her abilities,
Her motives, goals, aversions, and tastes,
The ways in which she is disposed to behave,
Her various roles in the social world,

However, when Charles thinks about Joan in different realms, his descriptions of her may not all agree. For example, his view of Joan as a person at work is that she is helpful and competent, but tends to undervalue herself; however, in social settings he sees her as selfish and overrating herself. What could lead Charles to make such different models? Perhaps his first representation of Joan served well to predict her social performance, but that model did not well describe her business self. Then, when he changed that description to also apply to that realm, it made new mistakes in the contexts where it formerly worked. Eventually, he found that he had to make separate models of Joan to predict her behaviors in other roles.

Physicist: Perhaps Charles should have tried harder to construct one single, more unified model of Joan.

This would not be feasible, because each of a person's mental realms may need different kinds of representations. Indeed, whenever a subject becomes important to us, we build new kinds of models for it—and this ever-increasing diversity must be a principal source of our human resourcefulness.

To more clearly see the need for this, we'll turn to a simpler situation: Suppose that you find that your car won't start. Then, to diagnosis what might be wrong, you may need to switch among several different views of what might be inside your car:

If the key is stuck, or the brake won't release, you must think in terms of mechanical parts.

If the starter won't turn, or if there is no spark, you must think in terms of electrical circuits.

If you've run out of gas, or the air intake's blocked, you need a model of fuel and combustion.

It is the same in every domain; to answer different types of questions, we often need different kinds of representations. For example, if you wish to study Psychology, your teachers will make you take courses in at least a dozen subjects, such as *Neuropsychology, Neuroanatomy, Personality, Perception, Physiology, Pharmacology, Social Psychology, Cognitive Psychology, Mental Health, Child Development, Learning Theories, Language and Speech, etc.* Each of those subjects uses different models to describe different aspects of the human mind.

Similarly, to learn Physics, you would need subjects with names like these: *Classical Mechanics; Thermodynamics; Vector, Matrix and Tensor Calculus; Electromagnetic Waves and Fields; Quantum Mechanics; Physical Optics; Solid State Physics; Fluid Mechanics; Theory of Groups, and Relativity.* Each of those subjects has its own ways to describe the events that occur in the physical world.

Student: I thought that physicists seek to find a single model or "grand unified theory" to explain all phenomena in terms of some very small number of general laws.

Those 'unified theories of physics' are grand indeed—but to apply them to any particular case, we usually need to use some specialized representation to deal with each particular aspect of what all the scientists of the past have discovered. Thus, whenever we deal with complex subjects like Physics or Psychology—we find ourselves forced to split such fields into 'specialties' that use different representations to answer different kinds of questions. Indeed, a major part of education is involved with learning when and how to switch among different representations.

Returning to Charles' ideas about Joan, these will also include some models of Joan's own views about herself. For example, Charles might suspect that Joan is displeased with her own appearance (because she is constantly trying to change this) and he also makes models of how Joan might think about herself in realms like these.

Joan's ideas about her own ideals.
Her ideas about her abilities,

Her beliefs about her own ambitions,
Her views about how she behaves,
How she envisions her social roles.

Joan would probably disagree with some of Charles views about her, but this may not make him change his opinion because he knows that the models that people make of their friends are frequently better than the models that people make of themselves. As programmer Kevin Solvay has said,[184]

"Others often better express myself."

We need Multiple Models of Ourselves

> *"... But even as these ordinary thoughts and perceptions flowed unimpeded, a new kind of question seemed to spin through the black space behind them all. Who is thinking this? Who is seeing these stars, and citizens? Who is wondering about these thoughts, and these sights? And the reply came back, not just in words, but in the answering hum of the one symbol among the thousands that reached out to claim all the rest: Not to mirror every thought, but to bind them. To hold them together, like skin. Who is thinking this? I am."*
>
> *—Greg Egan[185]*

We've discussed a few models that Charles might use when he thinks about his friend Joan. But what kinds of models might people use when they try to think about themselves? Perhaps our most common self-model begins by representing a person as having two parts—namely, a 'body' and a 'mind'.

That *body* division soon then grows into a structure that describes more of one's physical features and parts. Similarly, that model of *mind* will divide into a host of parts that try to depict one's various mental abilities.
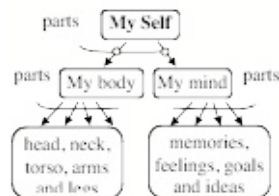
However, any model one makes of oneself will only serve well in certain situation, so we each end with different views of ourselves to use in different realms, in which each person has different goals and capabilities. This means when we think about ourselves, we'll needs ways to radiply switch among different models we've made of ourselves.
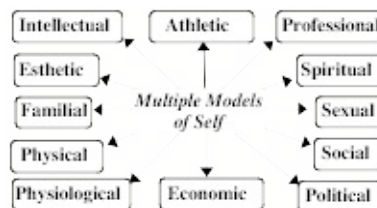
Any single model that tried to represent all this would soon become too complex to use. For, in each of those various realms of life, the models that we make for ourselves will portray us as having had somewhat different autobiographies, in which we pursued somewhat different aims, and ideals, while maintaining various different beliefs and making different interpretations of the same events. This suggests another idea about what one might mean when one tries to talk about one's self:

Daniel Dennett: "... we are all virtuoso novelists, who find ourselves engaged in all sorts of behaviour, and we always try to put the best "faces" on if we can. We try to make all of our material cohere into a single good story. And that story is our autobiography. The chief fictional character at the centre of that

autobiography is one's self."<super>[186]</super>

Multiple Sub-Personalities

> *"For there is not a single human being ... who is so conveniently simple that his being can be explained as the sum of two or three principal elements... Harry consists of a hundred or a thousand selves [but] it appears to be an inborn and imperative need of all men to regard the self as a unit. ... Even the best of us shares the delusion."*
>
> —Herman Hesse, in Steppenwolf.

When Joan is with a group of her friends, she regards herself as fairly sociable. But when surrounded by strangers she sees herself as anxious, reclusive, and insecure. For, we each make different self-models to use in different kinds of contexts and realms, just as we said in chapter §4: to think of herself.

Joan's mind abounds with varied self-models—Joans past, Joans present and future Joans; some represent remnants of previous Joans, while others describe what she hopes to become; there are sexual Joans and social Joans, athletic and mathematical Joans, musical and political Joans, and various kinds of professional Joans.

Thus each person has a variety of 'sub-personalities,' each of which may have some control over different sets of resources and goals, to be used when playing different—so that each has somewhat different ways to think. However, all of one's sub-personae will generally share most of one's skills and bodies of commonsense knowledge. However, conflicts among those sub-personalities can lead to problems because some of them might make rash decisions, while others might try to take over and maintain control.

For example, suppose that Joan is working at her professional job—but suddenly some social part of her mind reminds her of a time when she was trapped in an awkward relationship. She tries to shake off those memories, only to find herself thinking in childish ways about how her parents would view her present behavior. Thus Joan's thoughts might oscillate among such varying self-images as:

A partner in a business,
A person who likes to do research,
One with a certain collection of values,
A member of a family,
A person involved in a love affair,
A person having a pain in her knee.

In the course of such trains of everyday thinking, we frequently switch between different self-models, whose various outlooks may not be consistent, because we use them for different purposes. This means that when Joan needs to make a decision, the result will partly depend upon which sub-personalities are active then. A Business Self might be inclined to choose the option that seems more profitable; an Ethical Self might want to select the that suits her ideals 'best'; a Social Self might want to select the one that would most please her friends. For example, when we identify ourselves as members of a social group, then we can share their triumphs and failures with exuberance or remorse, and thus exhibit concern, compassion, and empathy. Thus, as we said in Chapter §1, each major change in emotional state may display a different sub-personality:

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 purposes. It's almost as though a switch had been thrown, and a different program has started to run.

Whenever we switch among sub-personalities, we are likely to change our ways to think—but because the context remains the same, we will still maintain some of the same of priorities, goals, and inhibitions, some contents of short-term memories, and some of our currently active Mental Critics.

However, some such changes may be larger, and we often hear sensational stories about persons who switch between totally different personalities. However, while such extremes are exceedingly rare, everyone undergoes changes of mood in which one exhibits somewhat different sets of intentions, behaviors and traits. Some of these shifts will be transiently brief, while others will be more persistent—but in each case, the sub-personality that is now in control will influence the views and goals that you use and pursue, and may claim that these are the views and goals of what it believes to be the 'genuine' You.

The Sense of Personal Identity

> *Augustine, Bishop of Hippo: "Of what nature am I? A life various, manifold, and vast. Behold in the numberless halls and caves, in the countless fields and dens and caverns of my memory, full without measure of numberless kinds of things— present there either through images as all bodies are; or present in the things themselves as are our thoughts; or by some notion or observation as our emotions are, which the memory retains though the mind feels them no longer..."*
> *—Confessions.*

In everyday contexts it often makes sense to think of one's Self as a permanent thing. But to what extent are you the same as you were ten minutes ago? Or are you like a carving knife that has had both its handle and blade replaced?

You're certainly not like the text of a book whose pages remain eternally unchanged; your 'contents' keep changing moment to moment. Nevertheless over months and years, enough of your knowledge remains the same—and different enough from anyone else's—that we recognize your stability. For one can argue that our 'identities' are mainly what's in our memories.

"A man is often willing to say that this is the same person who did something in the past, not on the basis of knowing that it is the same body but on a quite different basis—that the person recounts the past situation with great accuracy, exhibits similar personal reactions, and displays the same skills."— Encyclopedia Britannica.

A mind is an assembly of parts that each performs different activities. Each may have conflicting concerns and goals, which sometimes can only be resolved by turning off some participants. But then, when you change the way you think, what would it mean to say that you're still the same? Of course, that depends on how you describe yourself. To some of your models of yourself, significant parts of you will have changed—whereas, to some of your other models, you may remain unchanged. Consequently, it won't often make sense to ask what your Identity *is*. Instead, it usually would be better to ask, *which of your models of*

yourself would best serve your present purposes.

"Our brains appear to make us seek to represent dependencies. Whatever happens, where or when, we're prone to wonder who or what's responsible. This leads us to discover explanations that we might not otherwise imagine, and that helps us predict and control not only what happens in the world, but also what happens in our minds. But what if those same tendencies should lead us to imagine things and causes that do not exist? Then, perhaps that strange word 'I'—as used in 'I just had a good idea'—reflects the selfsame tendency. If you're compelled to find some cause that causes everything you do-why, then, that something needs a name. You call it 'me.' I call it 'you.'"

—SoM §22-7

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§9-2. Personality Traits

If you asked Joan to describe herself, she might say something like this:

Joan: "I think of myself as disciplined, honest, and idealistic. But because I am awkward at being sociable, I try to compensate by trying to be attentive and friendly, and when that fails, by being attractive."

Similarly, if you were to ask Charles to describe his friend Joan, he might declare that she is helpful, tidy, and competent, but somewhat lacking in self-confidence. Such descriptions are filled with everyday words that name what we call "character traits" or "characteristics"—such as *disciplined, honest, attentive,* and *friendly.* But what could make it possible for someone to describe a person at all? Why should minds so complex as ours show any clear-cut characteristics? Why, for example, should anyone tend to be usually neat or usually sloppy—rather than tidy about some things but not about others? *Why should personal traits exist at all?* Here are some possible causes for the appearance of such uniformities:

Inborn Characteristics. Two person could be born with the same resources, e.g., for becoming angry or afraid—but may differ in the conditions or priorities in which those functions become aroused—so that one of those individuals may tend to be more belligerent than the other one.

Investment Principle: Once we learn an effective way to do some job, we'll resist learning other ways to do it—even when we are told about some other technique that is better for this—because that new method will be more awkward to use until we become proficient at it. So, as our old procedures gain strength, it gets harder for new ones to compete with them.

Archetypes and Self-Ideals: Every culture comes with myths that describe the fates of beings that are endowed with larger-than-life traits. Few of us can prevent ourselves from becoming attached to those heroes and villains. Then some of us may adopt some goals of trying to guide how we change ourselves, to make those imagined traits become real.

Self-Control: It is hard to achieve any difficult goals unless you can make yourself persist at them. But one can't simply 'tell yourself' to persist, because different circumstance are likely to make one change ones goals and priorities.

So again, one may end up by training oneself.

Stereotyping: Whenever we encounter something, we assume 'by default' that it must be like other things that it reminds us of. This can yield such good results that we come to ignore other differences.

Of course, one could argue that personal traits don't really exist, but that trait-based descriptions are popular because, although they may wrong or incomplete, they seem so simple and understandable. For, perhaps the easiest way to describe a thing is by making a simple list of its properties. Thus we can describe a child's building block as being rectangular, heavy, and made of wood. Similarly, it is easy to say that a person is honest and tidy—as opposed to being deceitful and sloppy. However, of course, no person always tells the truth, or keeps everything perfectly neat. Nevertheless, it can save a great deal of effort and time to see people or things as stereotypes—by assuming their properties by default.

However, even when we know those assumptions may be wrong, they still may continue to influence us, so the concept of traits can be treacherous. Here is a common example of this: suppose that some stranger you've never met were to take your hand, look into your eyes, and then report this impression of you:

"Some of your aspirations tend to be unrealistic. At times you are extroverted, affable, sociable, while at other times you are introverted, wary and reserved. You have found it unwise to be too candid in revealing yourself to others. You are an independent thinker and do not accept others' opinions without good evidence. You prefer a certain amount of change and variety, and become dissatisfied when hemmed in by restrictions and limitations.

"At times you have serious doubts as to whether you made the right decision or did the right thing. Disciplined and controlled on the outside, you tend to be anxious and insecure inside. Your sexual adjustment has presented some problems for you. You have a great deal of unused capacity, which you have not turned to your advantage. You have a tendency to be critical of yourself, but have a strong need for other people to like and to admire you."[188]

Many people are amazed that a stranger could see so deeply inside of them—yet every one of those statements applies, to some extent, to everyone! Just look at the adjectives in that horoscope: *affable, anxious, controlled, disciplined, extroverted, frank, independent, insecure, introverted, proud, reserved, self-critical, self-revealing, sociable, unrealistic, wary.* Everyone has concerns in regard to each of those characteristics, so few of us can help but feel that each such prediction applies to us.

Thus, millions of people have been entranced by the prophecies of so-called

psychics, fortune-tellers, and astrologers—even when those forecasts are confirmed no more often than chance would predict.[189] Why do we give so much credence to them? One reason could be that we trust those "seers" more than we trust ourselves, because they appear to be 'reliable authorities.' Another possible cause could be that we tend to believe that we are already like what we wish to be—and fortune-tellers excel at guessing what most people most want to hear. But perhaps the simplest reason why such assessments may seem correct to us is that we each maintain so many self-models that almost any statement about ourselves will agree with at least some of those models.

Self-Control

It is hard to achieve any difficult goals unless, at least to some extent, you can make yourself persist at them. You would never complete any long-range plans if, whatever you tried, you kept "changing your mind." However, you cannot simply 'decide' to persist, because many kinds of events may later affect your goals and priorities. Consequently, we each must develop ways to impose less breakable self-constraints on ourselves.

For example, if you should ever need help from your friends, they will need to know what to expect from you—and to know when they can depend on you—and therefore, at least to some extent, you must make yourself predictable. Furthermore, it is important for you to be able to 'depend on yourself' to carry out at least some of your many plans, so again you need ways to restrict yourself. Our cultures help to us to acquire such skills by teaching us to admire various traits of consistency, commitment, and ambitiousness. Then if you come to admire those traits, you may make it your goal to train yourself to behave in those ways.

Citizen: Might not such restrictions cause you to pay the price of losing your spontaneity and creativity?

Artist: Creativity does not result from lack of constraints, but comes from discovering appropriate ones. Our best new ideas are the ones that lie just beyond the borders of realms that we wish to extend. An expression like "skdugbewlrkj" may be totally new, but would have no value unless it connects with other things that you already know.[190]

In any case, it is always hard to make yourself do things that do not interest you—because, unless you have enough self-control, the *Rest of Your Mind* will find more attractive alternatives. §4-7 showed how we sometimes control ourselves by offering bribes or threats to ourselves in the forms of self-incentives

like, *"If I fail, I'll be ashamed of myself,"* or *"I'll be proud if I can accomplish this."* To do this requires some knowledge about which such methods will work on ourselves—but generally, it seems to me, the techniques we use for self-control resemble those that we use to persuade our acquaintances— e.g., exploiting their various needs and fears by making promises of rewards or withdrawing their access to things that they want. Here are several other such tricks:[191]

> *Admonish yourself, "Don't give in to that."*
> *Try exercise, or take deep breaths.*
> *Ingest caffeine or other brain-affecting chemicals.*
> *Set your jaw, or furrow your brow.* [*Facial expressions work especially well because they affect you as much as they do your audience.*]

But why must you use such devious tricks to select and control your ways to think—instead of just choosing to do what you want to do? The answer is that you soon would be dead if any particular part of your mind could take over control of all the rest— and our species would quickly become extinct if we were able to simply ignore the demands of hunger or pain or sex. But fortunately, we evolved systems whereby, whenever we faced emergencies, our most urgent instinctive goals can supersede our fantasies.

Along with those built-in priorities, every human culture develops ways to help its members constrain themselves. For example, every game that our children play helps to train them to assume new roles and to swiftly switch among those mental states, while still obeying the rules of that game—which, in effect, is a virtual world. Therefore, and perhaps most important of all, every child should also recognize that *a game is only a game.* Alas, that's one lesson too few of us learn.

Self-Control is no simple skill, and many of us spend much of our lives seeking ways to make our minds 'behave.' Eventually, we each accumulate techniques whose workings are so opaque to us that we can only use vague suitcase-words for them; this leads to yet one more meaning for 'Self'—*our name for all the methods we use whenever we try to control ourselves.*

Dumbbell Ideas and Dispositions

There are two rules for success in life.
First, never tell anyone all that you know.

Why do we find it so easy to say that a certain person tends to be extroverted and sociable—as opposed to being shy and reclusive? More generally, why do we find it so easy to make such two-part distinctions for other aspects of our personalities? Thus we often group our tempers, emotions, moods and traits into pairs that we regard as opposites.

Solitary vs. Sociable Dominant vs. Submissive
Tranquil vs. Agitated Careless vs. Meticulous
Forthright vs. Devious Cheerful vs. Cranky
Audacious vs. Cowardly Joyous vs. Sorrowful

But what inclines us to describe our traits in terms of these kinds of two-part pairs when, for example, Agitation clearly is not the mere absence of Tranquility, nor is Joy just the absence of Sorrow? One explanation could be that this reflects a more general human tendency to see many other aspects of our minds as split into pairs of seemingly opposite qualities.

Left vs. Right Quantitative vs. Qualitative
Thought vs. Feeling Deliberate vs. Spontaneous
Rational vs. Intuitive Literal vs. Metaphorical
Logical vs. Analogical Reductionist vs. Holistic
Intellectual vs. Emotional Scientific vs. Artistic
Conscious vs. Unconscious Serial vs. Parallel

We see similar 'dumbbell' thinking at work when people try to describe the rest of the world in terms of opposing pairs of forces, spirits, or principles.

Goods vs Evils   Yangs vs. Yins   Angels vs. Devils

Perhaps the most amusing instance of this is the popular myth that each

person has two distinct kinds of mental activities that are embodied in opposite sides of the brain. In earlier times, the two halves of the brain seemed so alike that they were thought to be almost identical. But then, in the mid-20th century, when surgeons could cut the connections between those halves (in an adult), they were found to have some significant differences.[192] Then both press and public embraced the idea that every brain had two distinct and opposing sides—and this revived many views of the mind as a place for conflicts between such pairs of antagonists:

But how could so many such different distinctions be embodied in the same two halves of the very same brain? The answer is that this is largely a myth; a brain contains hundreds of different resources, and each of those activities involves many of these on both sides of the brain. However, there still may be some truth to that myth, but one can construct better explanations for this. For example, it long has been known that the 'dominant' side develops more machinery for language-based activities—and this could lead to more extensive development of self-reflective levels of thinking, while leaving more on the opposite side for spatial and visual activities. Here is what I think might be involved in this:

In early life, we start with mostly similar agencies on either side. Later, as we grow more complex, a combination of genetic and circumstantial effects leads one of each pair to take control of both. Otherwise, we might become paralyzed by conflicts, because many agents would have to serve two masters. Eventually, the adult managers for many skills would tend to develop on the side of the brain most concerned with language because those agencies connect to an unusually large number of other agencies. The less dominant side of the brain will continue to develop, but with more emphasis on lower-level skills, and with less involvement with plans and goals. Then if that brain-half were left to itself, it will seem more childish and less mature because it lacks those higher-level management skills.[193]

Here are some other reasons why we might like making two-part distinctions so much:

Many things seem to have Opposites. *It could be that, in early life, it is hard to distinguish what something 'is' without some idea of what it is not—and so we tend to think about things in relation to their possible opposites. For example, it often makes sense to classify physical objects as* large *or* small, *or as* heavy *or* light, *or as* cold *or* hot.

However, when you ask a young child about such things, you're likely to

hear that the opposite of *water* is *milk,* that the opposite of *dog* is *cat,* or that the opposite of a *spoon* is a *fork.* But that very same child may also insist that the opposite of *fork* is *knife.* Thus opposites depend on the contexts they're in, and so may overrule consistency.

Intensities and Magnitudes. Although it is hard to describe what feelings *are,* it seems easy to say how *intense* they are. This makes it seem quite natural to apply such adjectives as *slightly, largely,* or *extremely* to almost every emotion word—such as *sorry, pleasant, happy,* or *sad.*

One often justifies a choice, simply by declaring that one likes this option *more* than that one. However, *sorrow* is not the mere absence of *joy*—nor is *pleasure* merely the absence of *pain,* nor is *appetizing* an opposite to *disgusting.* It can be convenient to mis-represent such pairs as like the two ends of a single line, but doing this too frequently could lead to one-dimensional ways to think, no matter that such two-part distinctions may blur other dissimilarities between pairs of substantially different ideas, by leading us into supposing that both sides of each pair are almost the same—except for having 'plus' or 'minus' signs! Thus, representing feelings in terms of intensities can simplify how we make our decisions. However, I suspect that when we face more difficult choices, then we use more complex ways to settle conflicts among competing views or ways to think.[194]

Structural vs. Functional descriptions. Many of our distinctions are based on ways to make connections between what we learn about things and what we learn about using those things. Accordingly, it is often convenient to classify the parts of an object as playing 'principal' vs. 'supporting' roles—just as we did for 'a chair' in §8-3, where we identify the seat and back as its essential parts, and its legs and parts as merely serving to sustain them.[195]

Certainly, two-part distinctions can be useful when we need to choose between alternatives—but when that fails, we may have to resort to more complex distinctions. For example, when Carol is trying to build that Arch, it will sometimes suffice for her to first describe each block as being *short* or *tall,*

or *narrow* or *wide,* or *thin* or *thick;* then she may only need to decide which of those distinctions is relevant. However, on other occasions, Carol may need to find a block that satisfies some more elaborate combination of constraints that relate its height, width and depth; then she can no longer describe that block in terms of only a single dimension.

Inborn Brain-Machinery. Another reason why we tend to think in terms of pairs could be that our brains are innately equipped with special ways to detect differences between pairs of mental representations.

In §6-4 we mentioned that when you touch something very hot or cold, the sensation is intense at first, but then will rapidly fade away—because our external senses mainly react to how things change in time. (This also applies to our visual sensors, but we're normally unaware of this because our eyes are almost always in motion.) If this also applies to sensors *inside* a brain, this would make it easy to compare a pair of descriptions, simply by alternately presenting them. However, this 'temporal blinking' scheme would work less well for describing the relationships of more than two things—and that could be one reason why we are less proficient at making three-way comparisons. [See SoM *§23.3 Temporal Blinking*]

When is it appropriate to distinguish between only two alternatives? We often speak as though it is enough to classify a new thing or event in 'yes or no' terms like these:

Was this a failure or a success?
Should we see it as usual or exceptional?
Should we forget it or remember it?
Is it a cause for pleasure or for distress?

Such two-part distinctions can be useful when we have only two options to choose among. However, selecting what to remember or do will usually depend on making more complex decisions like these:

How should we describe this event?
What links should we connect it with?

Which other things is it similar to?
What other uses could we make of it?
Which of our friends should we tell about it?

More generally, it usually little sense to commit ourselves, for all future times, about which objects to *like or dislike*—or about which persons, places, goals, or beliefs we should *seek or avoid,* or *accept or reject*—because all such decisions should depend, upon the contexts that that we find ourselves in.

Accordingly, it seems to me that there is something wrong with most dumbbell distinctions: those divisions appear to be so simple and clear that they seem to be all that you need—and that satisfaction tempts you to stop. Yet most of the novel ideas in this book came from finding that two parts are rarely enough—and eventually my rule became: *when thinking about psychology, one never should start with less than three hypotheses!*

Why do people find it so hard to classify things into more than two kinds? Could this be because our languages don't come with verbs for speaking about *trividing* things? We all are good at 'comparing' pairs of things, and making lists of their differences—but few of us ever develop good ways to talk or to think about *trifferences*—that is, about relationships among triplets of things. Could this be because our brains don't come equipped with adequate, built-in techniques for this?

Perhaps this could be in part because a typical child's environment contains almost no significant 'triplets' of things. A typical two-year-old has a pair of hands, and is taught by a pair of parents to learn some way to put on a pair of shoes—and, soon, that typical two-year-old will learn to understand and to use word "two." perhaps from being familiar with such pairs as two hands or two shoes. But it usually takes yet another full year for that child to learn to use the word "three."

Robert Benchley: *"There are two kinds of people in the world: those who believe there are two kinds of people in the world and those who don't."*

<div align="center">∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞</div>

§9-3. Why do we like the idea of a Self?

Brian: You are all individuals!
Mob: We are all individuals!
Lone voice: I'm not.[196]
        —Monty Python: The Life of Brian

Most of the time we think of ourselves as having definite identities.

Introspectionist: I do not feel like a scattered cloud of separate parts and processes. Instead, I sense that there's some sort of Presence in me—an Identity, Spirit, or Feeling of Being—that governs and guides all the rest of me.

Other times we find ourselves feeling less decisive or less centralized.

Citizen: One part of me wants this, while another part of me wants that. I need to get more control of myself.

A few unusual persons claim never to feel any such sense of unity. Thus one philosopher went so far as to say,

Josiah Royce: "I can never find out what my will is by merely brooding over my natural desires, or by following my momentary caprices. For by nature I am a sort of meeting place of countless streams of ancestral tendency. ... I am a collection of impulses. There is no one desire that is always present to me."[197]

In any case, even when we feel that we're in control, we recognize conflicts among our goals. Then we may argue inside our minds, and try to find a compromise—but sometimes we still find ourselves subject to compulsions and urges we can't overcome. And even when we feel unified, others may see us as disorganized.

We solve easy problems in routine ways, scarcely thinking about how we accomplish these—but when our usual methods don't work, then we start to 'reflect' on what went wrong with what we were attempting to do—and this chapter maintains that we do such reflections by switching around in a great network of 'models'— where each purports to represent some facet or aspect of ourselves. Thus what we call 'Self' in everyday life is a loosely connected collection of images, models, and anecdotes.

In fact, if this view of the Self is correct, then there is nothing so special about it—*because that's how we represent everything else, namely, as a Panalogy.* Thus when you think about a telephone, you keep switching among different views of its appearance, its physical structure, the feelings you have when you use it, *etc.* It's the same when you think about your Self; you still may be using the same kinds of techniques that you use to think about everyday things; different parts of your mind are engaged with a variety of models and processes. But if so, then what impels us to believe that that we must be anything more than Josiah Royce's meetings of streams? What leads us to the strange idea that our thoughts cannot just proceed by themselves?

Jerry Fodor: "If there is a community of computers living in my head, there

had also better be somebody who is in charge; and, by God, it had better be Me."[198]

Citizen: Even if no central Self exists, you'd have to explain why we feel that one's there. When I think my thoughts and imagine things, must not there be someone who's doing those things!

After all, if we had those Single Selves to *want and feel and think for us,* then we would not have much need for Minds—and if our Minds could do those things by themselves, then it would be of no use to have those Selves? Then how could that curious "Self" idea ever help us to understand anything? *Aha!* Perhaps that is precisely the point: we use words like 'Me' and 'I' *to keep us from thinking about what we are!* For they all give the very same answer, "Myself," to every such question that we might ask. Here are some other ways in which that Single-Self concept is useful to us:

A Localized Body. You cannot walk through solid walls, or stay aloft without support. Where any part of your body goes, the rest of you must also go —and the Single-Self model includes the idea of being in only one place at a time.

A Private Mind. It is pleasant to think of your Self as like a strong, closed box, so that no one else can share your thoughts to learn the secrets you want to keep—for only you hold the keys to those locks.

Explaining our Minds. Perhaps it seems to make sense to say things like, *'I perceive the things that I see,'* because we know so very little about how our perceptions actually work. This way, that Single-Self view can help to keep us from wasting time on questions we don't know answers to.

Moral Responsibility. Each culture needs behavioral codes. For example, because our resources are limited, we sometimes have to censure Greed. Because we each depend on others, we have to chastise Treachery. And to justify our laws and decrees, we have to assume that some Single Self is 'responsible' for every willful, intentional deed.

Centralized Economy. We'd never accomplish anything if we kept asking questions like, "Have I considered every alternative?" We prevent this with Critics that interrupt us with, "That's enough thinking; I've made my decision!"

Causal Attribution. When we represent any thing or event, we like to attribute some Cause to it. So when we don't know what led to some thought, we assume that the Self was the cause of it. This way, we sometimes may use the word 'Self' the way we say *'it'* in *"it started to rain,"* because we don't know a more plausible cause.

Attention and Focus. We often think of our mental events as occurring in a

single 'stream of consciousness'—as though they all were emerging from some single, central kind of source, which can only attend to one thing at a time. We'll discuss this more in §§Attention.

Social Relations. Other people expect us to think of them as Single Selves, so unless we adopt a similar view, it will be hard to communicate with them.

These all are good reasons why the Single-Self view is convenient to use in our everyday lives. But when we want to understand how we think, no model so simple as that can portray enough useful details of how our minds work. For even if 'you' had some way to observe your entire mind simultaneously, that would be too complex to comprehend—so you still would be compelled to switch among simplified models of yourself.

Why should those models be simplifications? *That's because we would be overwhelmed by seeing too many unwanted details.* That's what makes a map more useful to us than seeing the entire landscape that it depicts; a good model helps us to focus on only those features of things that might be significant in some particular context. A good map or model may also include additional knowledge, as when the blueprint of a house shows the dimensions its parts were *intended* to have, as well as the name of its architect.)

The same applies to what we store in our minds. Consider how messy our minds would become if we filled them up with descriptions of things whose details had too little significance. So instead, we spend large parts of our lives at trying to tidy up our minds—selecting the portions we want to keep, suppressing others we'd like to forget, and refining the ones we're dissatisfied with.[199]

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§9-4. What is Pleasure, and why do we like it?

> *We may lay it down that Pleasure is a movement by which the soul as a whole is consciously brought into its normal state of being; and that Pain is the opposite. If this is what pleasure is, it is clear that the pleasant is what tends to produce this condition, while that which tends to destroy it, or to cause the soul to be brought into the opposite state, is painful.*
> *—Aristotle, Rhetoric, I, 10.*

We tend to feel pleased—or at least relieved—when we accomplish something we want. Thus, as we remarked in chapter 2,

"When Carol recognized that her goal was achieved, she felt satisfaction, fulfillment and pleasure—and those feelings then helped her to learn and remember."

Of course, we're delighted that Carol felt pleased—but how did those feelings help her to learn—and why do we like those feelings so much, and work so hard to find ways to attain them? Indeed, what does it mean to say that someone feels "pleased?" When people answer questions like these, we frequently hear examples of circular reasoning:

Citizen: I do the things that I like to do because I get pleasure from doing them. And naturally, I find them pleasant because those are the things that I like to do."

Certainly, we all can agree that *pleasure* is a feeling we 'get' when we're in a condition that we *like*—but that does not help much to explain what words like *pleasure* and *liking* mean. Our Dictionaries reflect the same problem:

Pleasure: a feeling of delight, happiness, or satisfaction.
Satisfaction: the pleasure that comes when a need is fulfilled.
Liking: a feeling of enjoying something or finding it pleasant.[200]

One reason why we get into such circles is that we don't have good ways to describe *any* feelings. To be sure, we find it rather easy to say how weak or strong a feeling feels—but when we are asked for more details, we usually cannot describe the feeling itself, but can only resort to analogies like, *"That*

pain was as piercing as a knife."

However, this should lead us next to ask what could make something so hard to describe. It seems to me that this is likely to happen when we fail to find a way to divide that thing—be it an object or mental a condition—into several separate parts (or layers, or phases, or processes). *This is because a thing that we cannot split into parts gives us nothing to use as pieces of explanation!* In particular, it is a popular view that pleasure is "elemental" in the sense that it cannot be explained in terms of anything else, and that the quest for pleasure or satisfaction is a "basic" human drive. Here is a parody of that idea:

Product Promoter: Happiness is the ultimate goal of all human beings, and all of us constantly aim toward this, whether through leisure, career, wealth, relationships or whatever. Our secure online ordering system offers a line of carefully chosen products to help you replace discomfort with pleasure.

However, this section will argue that what we call *pleasure* is indeed a suitcase-name for quite a few different processes that involve activities that we don't often recognize. For example, we usually see Pleasure as positive, but one can see it as negative—because of how it tends to suppress other competing activities. Indeed, to accomplish any major goal, one must suppress others that might compete with it, as in, *"I don't feel like doing anything else."* Discussing this is important because, it seems to me, the assumption that pleasure is simple or 'elemental' has been an obstacle to understanding our psychology. To see what is wrong with that idea, we'll catalog some of the feelings and activities that make this subject so difficult.

 Satisfaction. *A species of pleasure called 'satisfaction' comes when an ambition has been achieved.*
 Exploration. *We may also feel pleasure during a quest—and not only at the end of it. So it is not only a matter of being rewarded for achieving a goal.*
 Relief. *A species of pleasure called "relief" may come when a problem has been solved—if that goal was represented as an irritation or agitation.*
 Joy and Bliss. *Sometimes, when you enter a pleasant state you feel, if only transiently, as though all of your problems had been solved!*
 Critic-Suppression. *"I know this could be bad for me, but I like it so much that I'll do it anyway."*
 Credit Assignment and Learning. *Perhaps the most important aspects of pleasure are its connections with learning and memory.*

Success can also fill you with pleasure and pride—and may also motivate you to show other persons what you have done. But the pleasure of success soon

fades (at least for ambitious intellects) because, shortly after we put one problem to rest, another one quickly replaces it. For few of our problems stand by themselves, but are only parts of larger ones.

Also, after you've solved a difficult problem, you may feel relieved and satisfied, and sometimes may also feel a need to arrange for some sort of inner or outer celebration. Why might we have such rituals? Perhaps there's a special kind of relief that comes when one can dismiss a goal and release of resources that it engaged—along with the stresses that came with them. Clearing out one's mental house may help to make other things easier—just as the 'closure' of a funeral can help to assuage a person's grief.

But what if the problem you're facing persists? You can sometimes regard your present distress as a benefit, as in *"I'm certainly learning a lot from this,"* or *"Others may learn from my mistakes."* And everyone knows this magical trick for turning all failures into success: one can always tell oneself *"The true reward is the journey itself."*

So instead of trying to say what Pleasure *is,* we'll need to develop more ideas about what processes might be involved in what we often describe in simple terms such as "feeling good." In particular, it seems to me that we often use words like *pleasure* and *satisfaction* refer to an extensive network of processes that we do not yet understand—and when anything seems so complex that we can't grasp it all at once, then we tend to treat it as though it were single and indivisible.

> Pleasures are ever in our hands or eyes,
> And when in act they cease, in prospect, rise:
> Present to grasp, and future still to find,
> The whole employ of body and of mind.
>
> —Alexander Pope, in *Essay on Man*

The Pleasure of Exploration

> *"Pleasure pursues objects that are beautiful, melodious, fragrant, savory, soft. But curiosity, seeking new experiences, will even seek out the contrary of these, not to experience the discomfort that may come with them, but from a passion for experimenting and knowledge."*
>
> —St. Augustine, in Confessions, 35.55.

Understanding a new and difficult subject—or exploring an unfamiliar terrain—can lead to a lot of pain and stress. Then how can we keep this from holding us back from learning new ways to accomplish things? One antidote for this is *Adventurousness*.

> *"Why do children enjoy the rides in amusement parks, knowing that they will be scared, even sick? Why do explorers endure discomfort and pain—knowing that their very purpose will disperse once they arrive? And what makes people work for years at jobs they hate, so that someday they will be able to—they seem to have forgotten what! It is the same for solving difficult problems, or climbing freezing mountain peaks, or playing pipe organs with one's feet: some parts of the mind find these horrible, while other parts enjoy forcing those first parts to work for them."*
>
> —The Society of Mind, §9.4.

Most of our everyday learning involves only minor adjustments to skills that we already know how to use. One can do this by using 'trial and error; one makes a small change, and if that results in a pleasant reward (such as being pleased with an improved performance) then that change will become more permanent.[201] This fact has led many teachers to recommend that 'learning environments' should mainly consist of situations in which pupils get frequent rewards for success. To promote this, then, one should help the students to progress through a sequence of small, easy steps.

However, this strategy won't work well in unfamiliar realms because, when we learn a substantially new technique, this will involve more work with less frequent rewards, while enduring the additional stress of being confused and disoriented. It also may require us to abandon older techniques and representations that previously have served us well—and this might even arouse

a sense of loss that brings "negative" feelings akin to grief. Such periods of awkwardness and ineptitude would usually cause a person to quit.

This suggests that that "pleasant" or "positive" practice, alone, may not suffice for us to learn more radically different ways to think. This, in turn suggests that to become proficient at learning new things, a person must somehow acquire what Augustine called, in the extract above, *'a passion for experimenting and knowledge.'* Such persons must somehow have managed to train themselves to actually enjoy those discomforts.

Citizen: How can you speak of 'enjoying' discomfort? Isn't that a self-contradiction?

It is only a contradiction when you regard your Self as a single Thing. But when you see the mind as a society, then you no longer have to think of pleasure as an all-or-none thing. For now you can imagine that, while *some parts of your mind are uncomfortable, other parts of your mind may enjoy forcing those first parts to work for them.* For example, one part of your mind can still represent your state in a positive way by saying *"Good, this is a chance to experience awkwardness and to discover new kinds of mistakes!"*

Citizen: But wouldn't you still be feeling that pain?

Indeed, when struggling at their seemingly punishing tasks, athletes still feel physical pain, and artists and scientists feel mental pains—but, somehow, they seem to have trained themselves to keep those pains from spiraling into the awful cascades we call 'suffering.' But how could those persons have learned to suppress, ignore, or enjoy those pains, while preventing those disruption cascades? To answer that, we would need to know more about our mental machinery.

Scientist: Perhaps this does not really need any special explanation, because explorations can provide their own rewards. For me, few things bring more pleasure than making radical new hypotheses—and then showing that their predictions are correct, despite the objections of my competitors.

Artist: It seems almost the same to me, because nothing can surpass the thrill of conceiving a new kind of representation and then confirming that this will produce new effects in my audience.

Psychologist: It seems clear that many such achievers regard their ability to function in spite of pain, rejection, or adversity to be among their outstanding accomplishments!

In any case, all this suggests that 'exploration pleasure' (however it works)

may be indispensable to those who want to keep extending their development.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§9-5. What controls the mind as a whole?

How do human minds develop? We know that our infants are already equipped at birth with ways to react to certain kinds of sounds and smells, to certain patterns of darkness and light, and to various tactile and haptic sensations. Then over the following months and years the child learns many more perceptual and motor skills, and proceeds through many stages of intellectual development. Eventually, each normal child learns to recognize, represent, and reflect upon some its own internal states, and also comes to self-reflect on some its intentions and feelings—and eventually learns to identify these with aspects of other persons that it observes. This section will speculate about possible structures we might use to support those activities; the next section will suggest some ideas about how children might develop these.

This book has proposed several different views of how a human mind might be organized. We began by portraying the mind (or brain) as based on a scheme that deals with various situations by activating certain sets of resources—so that each such selection will function as a somewhat different "way to think."
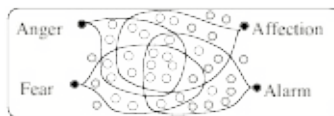
To determine which set of resources to select, such systems could begin with some simple sorts of *"If–>Do"* rules. Later these develop into more versatile *"Critic–>Selector"* systems.

Rule Based Reaction-Machine

Specific Situation	→	Specific Way to React
Ifs		DOs

"Critic-Selector" based Machine

Recognize a Problem-Type	→	Activate a Way to Think
Critics		Selectors
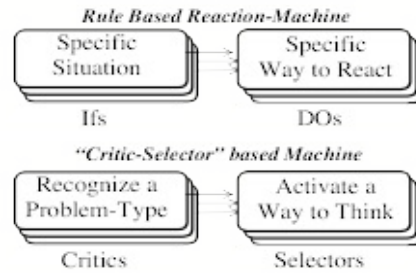
Chapter §5 conjectured that the adult mind comes to have multiple levels of organization. Each level has Critics to recognize situations and Selectors that can activate appropriate ways to think, by exploiting the resources at its own and at other levels. We also noted that these ideas could be seen as consistent with Sigmund Freud's early view of the mind as a system for resolving (or for ignoring) conflicts between our instinctive and acquired ideas.

Self-Conscious Emotions
Self-Reflective Thinking
Reflective Thinking
Deliberative Thinking
Reactive Thinking
Instinctive Reactions

Critics   Selectors

Values, Censors, Ideals and Taboos

Commonsense Thinking and Reasoning

Innate and Acquired Urges and Drives

Superego, Ego, and Id
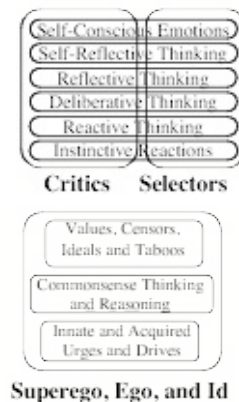
In chapter §8 we also reviewed various ways to represent knowledge and skills, and noted that these could be arranged in a stack with increasing degrees of expressiveness.
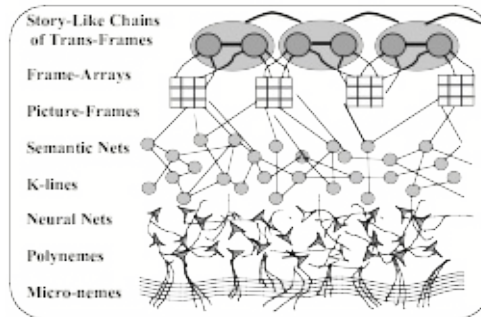
Each of those ways to envision a mind has different kinds of virtues and faults, so it would make little sense to ask which model is best. Instead, one needs to develop Critics that learn good ways to choose when and how to switch among them.

Is a Mind like a Human Community?

Perhaps a more popular model of mind portrays our mental processes as like a human community—such as a residential town or a company.[203] In a typical corporate organization, the human resources are (formally) organized in accord with some hierarchical plan.
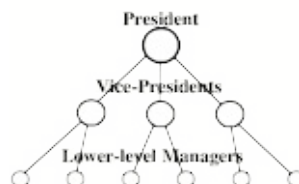
We tend to invent such 'management 'trees' whenever there's more than one person can do; then the work is divided into parts, which then are assigned to subordinates. This picture suggests that one might try to identify a person's Self with the Chief Executive of that company, who controls the rest through a chain of command in which instructions tend to branch down from the top.

However, this is not a good model for human brains because an employee of a company is a person who might be able to learn to perform virtually any new task—whereas, most parts of a brain are too specialized to do such things.[204] This difference may be important because, when a Company becomes wealthy

enough, it may be able to expand itself: when it wants to include new activities, it can hire additional employee-minds. In contrast, we humans don't (yet) have practical ways to expand our own individual brains. So, whenever you add a additional task (or break a large one into smaller parts)—and then try to do them all at once—then each sub-process will lose some of its competence, because it now has access to fewer resources. Perhaps we should state this as a general principle.

The Parallel Paradox: If you break a large job into several parts, and then try to work them all at once—then each process may lose some competence, from lacking access to resources it needs.

There is a popular belief that the brain gets much of its power and speed because it can do things many things in parallel. Indeed, it is clear that many of sensory, motor, and other systems do many things simultaneously. However, it also seems clear that as we tackle more complex and difficult goals, we increasingly need to divide those problems into subgoals—and then focus on these sequentially. This means that our higher, reflective levels of thought will tend to operate more serially.

This is less of a problem for a company, which can often divide a problem into parts, which it then can pass down to separate subordinates, who can deal with them all simultaneously. However, that leads to a different kind of cost:

The Pinnacle Paradox: As an organization grows more complex, its chief executive will understand it less, and will need to increasingly place more trust in decisions made by subordinates. (See §§Parallel Paradox.)

However, many human communities are less hierarchical than the companies that we just described, and make their decisions by using more cooperation, consensus, and compromise.[205] There is usually some 'leadership,' but in a working democracy, those leaders are somehow given authority by the membership to help, when needed, to assist in making decisions and settling arguments. Such negotiations can be more versatile than 'majority rule,' which gives to each participant a spurious sense of 'making a difference'—whereas that feeling ignores the fact that almost all differences get cancelled out. This raises questions about the extent to which our human 'sub-personalities' cooperate to help us accomplish larger jobs—but we don't know enough to say much about this.

Central and Peripheral Controls

Every higher animal has evolved many resources that can interrupt its

'higher level' processes, in reaction to certain states of affairs. These conditions include such signs of possible dangers as rapid motions and loud sounds, unexpected touches, and the sighting of insects, spiders, and snakes. We also react to such bodily signs of aches and pains, feelings of illness, and such needs as hunger and thirst. Similarly, we are subject to more pleasant kinds of interruptions, such as the sights and smells of foods to eat, and of signals of sexual interest.

Many such reactions work without interrupting most mental activities—as when your hand scratches an insect bite, or when you move into shade to avoid excessive light. A few of these mainly instinctive alarms are: *Itching, impending collision, hunger, thirst, bright light, excessive heat or cold, losing one's balance, loud noise, pain, hearing a growl or snarl, seeing a spider, insect, or snake.*

We are also subject to alarms that seem to come from 'inside the mind'— such as when we detect an unexpected pattern or opportunity, or a failure of some process to work, or a conflict between our goals and ideals. Here are a few of these mainly acquired, internal alarms, many of which could be represented by using Critics, Censors, and Suppressors: *phobias, obsessions, and sense of surprise; failure of a plan or goal; grief, guilt, shame, or disgust; and conflicts among one's goals and ideals.*

While most alarms could be handled by a Critic-Selector model of mind, one also could take a less centralized view, in which the processes that we call 'thinking' are affected by a host of other, partly autonomous processes. For example, one could think of a city or town as an entity whose processes are influenced by the activities of sub-departments concerned with transportation, water, power, fire, police, school, planning, housing, parks, and streets—as well as legal and social services, public works, and pest control, etc., each with its own sub-administrations. Can one think of a city as having a Self? Some observers might argue that each town has a certain 'ambience' or 'atmosphere,' and certain traits and characteristics. But few would insist that a city or town has a 'sentient' personality.

Reader: Perhaps that's because they don't have your idea that a "Self," is a network of models, each of which may help a system to answer questions about itself. But in fact, each of those departments for planning, power, parks, and streets—and each of those other agencies—have plenty of diagrams, charts and maps that represent aspects of the town they're in.

I cannot disagree with that—except to argue that, usually, few of those maps or models are accessible to the other departments. Perhaps if all those different

representations were assembled into efficient panalogies, the resulting system might indeed seem to have more of a personality.

Programmer: I like some of your theories about how minds work—except that all of your schemes seem far too complex for a system that functions reliably enough. What happens if some of its parts break down? A single error in a large computer program can cause the entire system to stop.

I suspect our human 'thinking processes' frequently 'crash'—perhaps as often as several times per second. However, when this happens you rarely notice that anything's wrong, because your systems so quickly (and imperceptibly) switch you to think in different ways, while the systems that failed are repaired or replaced. Here are a few of the kinds of failures that are likely to get somewhat more 'attention.'

You have trouble recalling past events.
You have trouble when solving an urgent problem.
You cannot decide which action to take.
You've lost track of what you were trying to do.
Something has happened that surprises you.

Nevertheless, in cases like these, usually you still can switch to other productive ways to think. For example, you might change the domain you are searching through, or select some other problem to solve, or switch to some different overall plan, or make a major switch in emotional state—without knowing or even being concerned with why your original project might have failed.

Furthermore, it seems possible that, whenever some of your systems fail, your brain may retain some earlier versions of it. Then in situations where you get confused, you may be able to ask yourself, *"How did I deal with such things in the past?"* Then this might cause some parts of your mind to 'regress' to an earlier version of yourself, from an age when such matters seemed simpler to you. This suggest another reason why we might like the idea of having a Self:

"One's present personality cannot share all the thoughts of all one's older personalities—and yet it has some sense that they exist. This is one reason why we feel that we possess an inner Self—a sort of ever-present person-friend, inside the mind, whom we can always ask for help."
    —§17.01 of "The Society of Mind."

However, we should not ignore the tragic fact that people also are subject to failures from which recovery may be difficult, or impossible. For example, if

something went wrong with the machinery that controls your Critic/Selector processes, then the rest of your mind may become reduced to a disorganized cloud of inactive resources, or get stuck with some single, unswitchable way to think.

"The most merciful thing in the world, I think, is the inability of the human mind to correlate all its contents. We live on a placid island of ignorance in the midst of black seas of infinity, and it was not meant that we should voyage far. The sciences, each straining in its own direction, have hitherto harmed us little; but some day the piecing together of dissociated knowledge will open up such terrifying vistas of reality, and of our frightful position therein, that we shall either go mad from the revelation or flee from the deadly light into the peace and safety of a new dark age."

—H.P. Lovecraft, "The Call of Cthulhu"

Mental Bugs and Parasites

If a mind could make changes in how it works, it would face the risk of destroying itself. This could be one reason why our brains evolved so many, partly separate systems, instead of a more unified and centralized one: there may have been substantial advantages to imposing limits on the extent to which our minds could examine themselves!

For example, no single Way to Think should be allowed to have too much control over the systems we use for credit assignment—because then it could make itself grow beyond bound. Similarly, it would be dangerous for any resource to be able to keep enforcing its goals, because then it could make the rest of the mind spend all its time at serving it. The same would apply to any resource that could control our systems for pleasure and pain—for any resource that found a way to completely suppress some instinctive drive might be able to force its person to never sleep, or to work to death, or to starve itself.

While such drastic calamities are rare, a great many common disorders do result from the growth of such 'mental parasites.' For many human minds do indeed get enslaved by the self-reproducing sets of ideas that Richard Dawkins entitled "memes. " Such a collection of concepts may include ways to grow and protect itself by displacing competing sets of ideas.[206] (I should note that many such sets of beliefs are so widespread that they are not regarded as pathological.)

However, many people are so clever that, when their minds get occupied by those 'mental parasites,' they may invent or find some social niche in which they can 'make a living' by recruiting yet other minds to adopt those same sets of strange ideas.

[Another class of bugs: thinking too much and getting in circles.] Nevertheless, the problem of meandering is certain to re-emerge once we learn how to make machines that examine themselves to formulate their own new problems. Questioning one's own "top-level" goals always reveals the paradox-oscillation of ultimate purpose. How could one decide that a goal is worthwhile —unless one already knew what it is that is worthwhile? How could one decide when a question is properly answered—unless one knows how to answer that question itself? Parents dread such problems and enjoin kids to not take them seriously. We learn to suppress those lines of thoughts, to "not even think about them" and to dismiss the most important of all as nonsensical, *viz.* the joke "Life is like a bridge." "In what way?" "How should I know?" Such questions lie beyond the shores of sense and in the end it is Evolution, not Reason, that decides who remains to ask them. (from "Jokes and the Logic of the Cognitive

Unconscious."

To protect themselves from such extremes, our brains evolved ways to balance between becoming too highly centralized, or too dispersed to have much use. We had to be able to concentrate, yet also respond to urgent alarms. We still needed some larger-scale organization because no much smaller part of us could know enough about the world to make good decisions in all situations.

Why don't we have more bugs than we do?

In the evolution of our brains, each seeming improvement must also have brought additional sorts of dangers and risks—because every time we extended our minds, we also exposed ourselves to making novel types of mistakes. Here are some bugs that everyone's subject to:

Making generalizations that are too broad.
Failing to deal with exceptions to rules.
Accumulating useless or incorrect information.
Believing things because our imprimers do.
Making superstitious credit assignments.
Confusing real with make-believe things.
Becoming obsessed with unachievable goals,

We cannot hope to ever escape from all bugs because, as every engineer knows, most every change in a large complex system will introduce yet other mistakes that won't show up till the system is moved a somewhat different environment. In any case, such failures are not uniquely human ones; my own dog also suffers from most of those bugs.

Each human brain is different because it is built by pairs of inherited genes (each chosen by chance from one of the parents). Also, many of its smaller details depend on other events that happened during its early development. So an engineer might wonder how such a machine could possibly work in spite of so many possible variations.

In fact, it was widely believed until recent years, that our brains must be based on some not-yet-understood principles, whereby every fragment of process or knowledge was (in some unknown manner) 'distributed' in some global way so that the system still could function in spite of the loss of any part of it. Today, however, we know that many functions do depend on highly localized parts of the brain. However, the arguments in this book suggests a different solution to this: we have so many different ways to accomplish most important jobs that we can tolerate the loss of some skills, because we may be

able to switch to another.

In any case, each human brain has many different kinds of parts, and although we don't yet know what all of them do, I suspect that many of them are involved with helping to suppress the effects of defects and bugs in other parts. Consequently, it will remain hard to guess why our brains evolved as they did, until we build more such systems ourselves—to learn which such bugs are most probable.

∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞

§9-6. Why makes feelings so hard to describe?

A color stands abroad

On solitary hills

That science cannot overtake

But human nature feels

—Emily Dickinson.[207]

We usually don't find it hard to compare two similar kinds of stimuli. For example, one can say that sunlight is brighter than candle light, or that Pink lies somewhere between Red and White, or that a touch on your upper lip is somewhere between your nose and your chin. However, this says nothing about those sensations themselves. Instead, it is more like talking about the distances between some nearby towns on a map—while saying nothing at all about those towns.

Similarly, when you try to describe the feelings that come with being in love, or from suffering fear, or when seeing a pasture or a sea, you'll soon find that you are mentioning other things that these remind you of, instead of what those feelings are. And then, perhaps, you will come to suspect that one can never really describe what anything *is;* one can only describe what that thing is *like*— or what that that thing *reminds* you of.

For example, if I were to ask about what *Red* means to you, you might say that this first makes you think of a rose, which reminds you, in turn, of being in love. Red might also remind you of blood, and make you feel some sense of dread or fear. Similarly Green might make one think about pastoral scenes and Blue might suggest the sky or the sea.

I have mentioned all this to emphasize the complexity of what can happen when a person attends to the sight of a *single* color or the sensation of a *single* touch. However, as many philosophers have complained, those remindings don't seem to describe or explain the *experience* of seeing that color or feeling that touch. Indeed, some present-day philosophers regard this to be one of the hardest problems they have tried to face: *Why do people experience events—instead of just simply processing them.* Listen to one such philosopher:

David Chalmers: "When we visually perceive the world, we do not just process information; we have a subjective experience of color, shape, and depth. We have experiences associated with other senses (think of auditory experiences of music, or the ineffable nature of smell experiences), with bodily sensations (e.g., pains, tickles, and orgasms), with mental imagery (e.g., the colored shapes that appear when one rubs ones eyes), with emotion (the sparkle of happiness, the intensity of anger, the weight of despair), and with the stream of conscious thought.

"[That we have a sense of experiencing] is the central fact about the mind, but it is also the most mysterious. Why should a physical system, no matter how complex and well-organized, give rise to experience at all? Why is it that all this processing does not go on "in the dark", without any subjective quality? Right now, nobody has good answers to these questions. This is the phenomenon that makes consciousness a real mystery."[208]

Let me summarize what I consider to be an explanation for that 'mystery,' and then I'll develop some further details.

When we see our friend Charles react to things, we cannot see the machinery that causes him to react in those ways—and so we have few alternatives to simply saying thinks like, *"he is reacting to what he is experiencing."* But then, we must be using that word *'experiencing'* as an abbreviation for what we would say if we knew what had happened inside some friend's head—such as, *"Charles must have detected some stimuli, and then made some representations of these, and then reacted to some of those by changing some of the plans he had made, etc."*

Furthermore, we ought to observe that, if your brain can begin to speak about some 'experience' it must already have access to some representations of some aspects of that event; otherwise, you would not remember it—or be able to claim that you have experienced it! So your very act of asserting that *you* have had that experience demonstrates that this 'experience' cannot be a simple or basic thing, but must be a complex process that is involved with the high-level networks of representations that you call your Self.

This means that the problem which Chalmers calls 'hard' is not really a single problem at all, because it condenses the complexity of all those many steps by squeezing them into the single word, 'experience' and then declares this to be a mystery.

What could make our sensations and feelings so difficult to talk about? You look at a color and see that it's Red. Something itches your ear and you know where to scratch. That's all there seems to be to it; you recognize that experience. No thinking seems to intervene. Indeed, quite a few philosophers have argued that the qualities of such sensations are so basic and irreducible that they will always remain inexplicable—because that is 'just the way those things are' and there is nothing else to an say about them.

However, here we shall will take the opposite view—that what we call *sensations* are extremely complex. They sometimes involve extensive cascades in which some parts of the brain are affected by signals whose origins they cannot detect—and therefore, would not be able to explain.

However, I do not mean at all to suggest that this complexity must be an obstacle to our ever being able to understand what a *sensation* 'really is.' Indeed, to recognize that a subject is complex is often the first step in the process of mastering it! This is can be seen as a principle that often applies to Psychology:

The "Easy is Hard Paradox": *The things that seem the simplest may actually be the ones that are the most complex.*

In other words, if you wrongly insist that something is simple, then it will remain a mystery—because, if you are actually facing an intricate problem, then you are unlikely to find a path toward solving it, until you recognize how complex it is.

In particular, the mystery of 'subjective experience' won't disappear until we recognize how it may engage many other aspects of how we think—including our highest forms of reflective thought.

<center>∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞</center>

§9-7. How do you know when you're feeling a pain?

Common sense might answer that you can't have a pain without knowing it. However, some thinkers disagree with that:

Gilbert Ryle: "A walker engaged in a heated dispute may be unconscious of the sensations in his blistered heel, and the reader of these words was, when he began this sentence, probably unconscious of the muscular and skin sensations in the back of his neck or his left knee. A person may also be unconscious or unaware that he is frowning, beating time to the music, or muttering."[209]

Similarly, Joan might first notice a change in her gait, and only later notice that she's been favoring her injured knee. Indeed, her friends may be more aware than she is, of how much that pain was affecting her. Thus, one's first awareness of being in pain may come only after detecting other signs of its effects, such as discomfort or ineffectiveness—perhaps by using the kind of machinery that we described in §4-3.
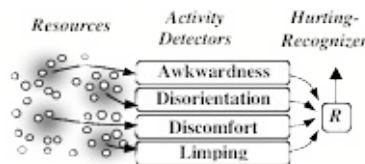
If you think you feel pain, could you be mistaken? Some would insist that this cannot be because *pain* is the same as *feeling pain*—but again our philosopher disagrees:

Gilbert Ryle: "The fact that a person takes heed of his organic sensations does not entail that he is exempt from error about them. He can make mistakes about their causes and he can make mistakes about their locations. Furthermore, he can make mistakes about whether they are real or fancied, as hypochondriacs do."

We can make such mistakes because what we 'perceive' does not come directly from physical sensors but from our higher-level processes. Thus, at first the source of your pain may seem vague because you have only noticed that something's disrupting your train of thought; then the best that you can say might be, *"I don't feel quite right, but I don't quite know why. It could be a*

headache just starting to hurt. Or maybe the start of a bellyache." And while such feelings indeed might result from a pain, they could also result from other conditions that your mental critics misrepresent as caused by a pain.

Similarly when you are falling asleep, the first things you notice might be that you've started to yawn, or keep nodding your head, or making a lot of grammatical errors; indeed, your friends might notice these before you do. One might even see this as evidence that people have no special ways to recognize their own mental states, but do this with the same methods they use to recognize how other persons feel.

Charles: Surely that view is too extreme. Like anyone else, I can observe my behavior 'objectively.' However, I also have an ability—which philosophers call 'privileged access'—with which I can inspect my own mind 'subjectively' in ways that no other person can.

We certainly each have some privileged access, but we should not overrate its significance. I suspect that our access to our own thoughts provides more *quantity* but does not seem to bring much more *quality*: our self-reflections reveal very little about the nature or causes of what we can see of our own mental activities. Indeed, our self-assessments are sometimes so inept that our friends may have better ideas about how we think. That's one reason why I suggested that we mainly represent ourselves in the same ways that we use to describe our friends.

Joan: Still, one thing is sure: none of my friends can feel my pain. I surely have privileged access to that.

It is true that the nerves from your knee to your brain convey signals that none of your friends can receive. But it's almost the same when you talk to a friend through a telephone. 'Privileged access' does not imply magic; it's merely a matter of privacy. No matter how private those lines may be, there still must be some processes that try to assign some significance to the signals that get to your brain from your knee. That's why Joan might find herself wondering, *"Is this the same pain that I felt last winter, when my ski boot did not release quickly enough?"*

Joan: I'm not even sure that it was the same knee. But isn't something missing here? If sensations are nothing but signals on nerves, then why are there such distinctive differences between the tastes of sour and sweet, or between the colors of red and blue?

<center>∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞∞</center>

Feelings are hard to describe because they are complex!

Our folk-psychology still maintains that our sensations have certain 'basic' or 'irreducible' qualities that, somehow, stand all by themselves and cannot be reduced to anything else. For example, each color like *Green* and each flavor like *Sweet* has its own 'ineffable' character, which is, therefore, unexplainable. For if such qualities do not have any smaller parts or properties, then there's no possible way to describe them.

Philosophers call this the problem of 'Qualia," and some them argue that to understand the nature of those qualities is a fundamental problem of philosophy, because they have no physical properties. To be sure, it is not hard to see (or measure) how much more Red is one spot than another or, at least to some degree, how much sweeter is one peach than another; however (those philosophers claim) such comparisons tell you nothing whatever about the nature of the *experience* of seeing *redness* or tasting *sweetness.*

I want to discuss this briefly here, because some readers might object that if we cannot explain such 'subjective' things, that would undermine the whole idea that we can explain the human mind entirely in terms of physical things (such as our brain's machinery). In other words, if the sensation of *sweetness* can never be measured or weighed, *or detected in any physical way,* then it must exist in a separate mental world, where it cannot possibly interact with any physical instruments.

Well, let's first observe that this claim must be wrong, because it is self-contradictory. For, if you can tell me that you have experienced *sweetness* then, somehow, *that sensation has caused your mouth to move!* So clearly, there must be some 'physical instrument' in your brain that recognized the mental activity that embodies your experience. In other words, we are simply facing, again, the same kind of problem that we solved in the previous section: we simply need another one of those internal "condition-detecting" diagrams!
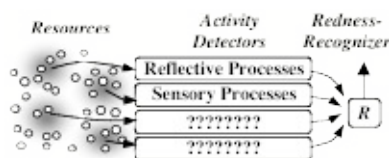
Of course, there is something missing here, because we do not yet quite

know how to connect those condition detectors. However, so far as I can see, this is merely another instance of what I called the "Easy is Hard Paradox." Indeed, it should be no more than a very few years before we can ask a brave philosopher to enter a suitable scanning device so that we can discover which brain-cells best distinguish the conditions that we wish to detect.

In other words, to understand how feelings work in more detail, we'll have to stop looking for simple answers, and start to explore more complex processes. When a ray of light strikes your retina, signals flow from that spot to your brain, where they affect other resources, which then transmit other kinds of reports that then influence yet other parts of your brain.[210]

At the same time signals from the sensors in your ears, nose, and skin will travel along quite different paths, and all these streams of information may come to affect, in various ways, the descriptions the rest of your mind is using. So, because those pathways are so complex and indirect, when you try to tell someone about what sensation you feel, or what you are experiencing, you'll be telling a story based on sixth-hand reports that use information that has gone through many kinds of transformations. So despite what some philosophers claim, we have no basis to insist that what we call our 'sense of 'experience' is uniquely direct.

The old idea that sensations are 'basic' may have been useful in its day, the way the four kinds of 'atoms' of antiquity were supposed to be elementary. But now we need to recognize that our perceptions are far less 'direct,' because they are affected by what our other resources may want or expect. This might relate to the fact that we sometimes clearly 'see' objects, which do not 'really' exist.

We like to attribute 'qualities' to stimuli—instead of to the more complex activities that they influence in the rest of our brains. Consider the two different feelings that you get when I first touch your right hand and then touch your left. How would you describe the difference between those two kinds of sensations? In some ways, they seem different, yet in other ways they seem similar. To compare them you would need to use descriptions that your resources have made, and that comparison will depend on the details of how those resources

were built (or have learned) to process those two kinds of evidence. Such a process must involve quite a few different levels and stages, perhaps extending to some self-reflection. This suggests that what we call 'qualities' might mainly reflect the ways we assess the relations *between* events in our brains.[211]

What would it mean to express what you're feeling?" Perhaps it is no accident that, literally, 'expression' means squeezing. But when you try to "express yourself", there is no single thing to squeeze; a typical human mental state is too complex to project through our tiny channels or language and gesture. However, your brain is equipped has many resources, each of which can make a description of some aspect of certain parts of your state of mind. Then, although some of these may disagree, you—that is, some of your language resources—can formulate some statement of "How I feel" by selecting from among those other descriptions—perhaps by simply choosing the one whose signals are the most intense.

But, of course you cannot describe your entire state, because this is something that's constantly changing. At one moment you're thinking about your foot; then your attention is drawn to the sky; next you notice a change in some sound, or turn to look at something that moves—*and then you notice that you are noticing these.* There's no way for you to be 'simply aware of yourself' because 'you' are a river of rivaling interests, always enmeshed in cascades of attempts to describe its rapidly changing currents.

Vitalist: Your theories are too mechanical, they're filled with parts, but have no wholes. What they need is some kind of lifelike cement—some coherent essence to hold them together.

I sympathize with that quest for cement, but I can't see how it could actually help because, whatever adhesive you might propose—such as a single central Self— you'd still be obliged to describe *its* parts, and what magical glue binds *them* together. So I don't see much use for words like 'spirit' or 'essence,' which only serve to make us re-ask the same questions again. As for terms like 'I,' 'Me,' or 'self-aware,' these appear to be useful ways to refer to the times when we use higher-level self-models.

§9-8. The Dignity Of Complexity

Citizen: I'm horrified by your idea—that a human being is just a machine—and worse, that it is programmed by a slipshod collection of sloppily organized parts and processes. No self-respecting

persons would want to think of themselves as any such mess of contraption.

[This section should begin by re-stating the "Easy is Hard Paradox": "If you wrongly insist that something is simple, then it will remain a mystery—because, if you are actually facing an intricate problem, then you are unlikely to find a path toward solving it, until you recognize how complex it is."]

Perhaps the most popular concept of what we are assumes that we each have a central core—some sort of invisible spirit or ghost that comes to us as an anonymous gift. But we ought to give credit where credit is due, by recognizing how we came to exist.

We came from an tremendous set of experiments—octillions of trials and errors to which sextillions of creatures devoted their lives, each living and dying in various ways that each may have contributed to giving us a slightly more powerful brain. This struggle proceeded for thirty million centuries and—so far as we know—no other such large and magnificent process has ever occurred in the universe. Accordingly each human mind has resulted from an unthinkably vast history in which animal on earth (including those not in our direct ancestry) spent its life adjusting, adapting, testing, and eventually dying—to see which ones had offspring who best could thrive in all those past environments, which began in oceans and seas, and then extended to the tidelands and shores, and then to the forests, deserts, plains—and, eventually, to our self-made villages and towns. Yet most traditional accounts of our origins make no mention whatever about this prodigious saga of sacrifice—and some institutions have even aimed to suppress any lessons about the eons through which we developed our bodies and brains.

It seems wrong to me to dismiss our minds as though they came as gratuitous gifts. For if, as many of us suspect, our intelligence is unique in this entire universe, then we also must be responsible to all those creatures who died for us: we need to ensure that the minds which we bear do not go to waste through some foolish mistake.

§9-9. Some Sources of Human Resourcefulness

We can credit our human resourcefulness to processes that developed over the vastly different time spans of our genetic endowments, cultural heritages, and personal experiences.

GENETIC ENDOWMENT: Inherited systems in our brains help us to survive

the most common kinds of hazards and threats. Those mental resources were selected from variations that occurred over some five million centuries.

If you find a big book about the brain, and examine the index at the end, you will see a list of names of hundreds of different parts of the brain— each of which is known to have appreciably different functions. Each those 'centers' or 'regions' contains a good many millions of brain-cells which also come in several varieties. Our scientists know a certain amount about what certain of those brain parts do but, generally, the details about how most of them actually work are still largely shrouded in mystery.

CULTURAL HERITAGE: The communal sets of beliefs called 'cultures' evolved over hundreds of centuries, during which intellectual processes selected ideas from many millions of individuals.

Our cultures must be the principal source of much of our personal knowledge and skills, because no single individual would have enough time to learn so much.

INDIVIDUAL EXPERIENCE: Each year, one learns millions of fragments of knowledge from one's own private experiences.

Almost everyone understands the extent to which each person's knowledge and beliefs have been passed down centuries from those of our cultural ancestors. No person, alone, could ever invent as many conceptions as we can observe in any typical four-year old. But fewer of us appreciate the extent of how much such knowledge hides in virtually every language word.[212]

Listen closely to anything anyone says, and soon you'll hear analogies. We speak of time in terms of space, as like a fluid that's 'running out'; we talk of our friendships in physical terms, as in "Carol and Joan are very close." All of our language is riddled and stitched with such ways of portraying things—and sometimes we call these "metaphors." Some metaphors seem utterly pedestrian, as when we speak of "taking steps" to cause or prevent some happening. Other metaphors seem more remarkable—as when a scientist solves a problem by conceiving of a fluid as made of tubes. When such analogies play surprisingly productive roles, we notice them, but we rarely notice how frequently we use the same techniques in commonsense thinking.

Some metaphors and analogies have very simple origins, as when they come from stripping away enough details to make two different objects seem the same. But other forms of metaphor are as complex as can be. In either case, metaphors are useful when they represent things in ways that help us to transport

knowledge into other realms, where we can still apply the same already - developed skills. And in many cases this results in our most productive, systematic, cross-realm correspondences. In this book these are called 'panalogies.'

How do we learn our most precious panalogies? I suspect that many of them are virtually born into our brains—because the regions, which represent various realms, have basically such similar wiring that we can discover those metaphors by ourselves. However, we also learn many of our metaphors from the patterns of usage of words by other members of our communities.

On rare occasion, some individual discover a new representation or formulation that is both so fruitful and so easy to explain that it becomes part of the general culture. Naturally, we'd like to know how those greatest discoveries were made. But because this is buried in the past, most of those great rare events may never be explained at all—because, like our evolutionary genes, these need happen by accident only once, and then can spread from brain to brain.

All this has empowered us to deal with huge classes of new situations. The previous chapters discussed many aspects of what gives us so much resourcefulness:

We have multiple ways to describe many things—and can quickly switch among those different perspectives.

We make memory-records of what we've done—so that liter we can reflect on them.

We learn multiple ways to think so that, when one of them fails, we can switch to another.

We split hard problems into smaller parts, and use goal-trees, plans, and context stacks to help us keep making progress.

We develop ways to control our minds with all sorts of incentives, threats, and bribes.

We have many different ways to learn, and also can learn new ways to learn.

We can often postpone a dangerous action and imagine, instead, what its outcome might be in some Virtual World.

Our language and culture accumulates vast stores of ideas that were discovered by our ancestors. We represent these in multiple realms, with metaphors interconnecting them.

Most every process in the brain is linked to some other processes. So, while any particular process may have some deficiencies, there frequently will be other parts that can intervene to compensate.

Nevertheless, our brains still have bugs. Similarly, in the coming decades of

research toward Artificial Intelligence, every system that we build will keep showing unexpected flaws. In some cases, we'll be able to diagnose specific errors in those designs, and hence be able to correct them. But when we can find no such simple fix, then we will be forced instead to evolve increasingly complex systems in which each process needs to be supervised by various Critics. [edit] And through all this, we can never expect to find any foolproof strategy to balance the advantage of immediate action against the benefit of more careful, reflective thought. Whatever we do, we can be sure that the road toward designing 'post-human minds' will be rough.

Bibliography

Acerra 1999: Francesca Acerra, Yves Burnod and Scania de Schonen, European Symposium on Artificial Neural Networks, Bruges (Belgium), 1999, ISBN 26000499X, pp. 129-134. Text at http://www.dice.ucl.ac.be/Proceedings/esann/esannpdf/es1999-22.pdf Aristotle a: *Nicomachean Ethics*, Book VIII. Text at http://etext.library.adelaide.edu.au/a/aristotle/nicomachean/

Aristotle b: *Rhetoric*. Text at http://etext.library.adelaide.edu.au/a/a8rh/

Arnold 1865: Matthew Arnold, *Essays in Criticism*. Ed. S. R. Littlewood. London: Macmillan. 1958.

Augustine 397: The Confessions, Book 10. Text at http://www.ourladyswarriors.org/saints/augcon10.htm#chap10.

Baars 1996: Bernard J. Baars, *"Understanding Subjectivity: Global Workspace Theory and the Resurrection of the Observing Self,"* Journal of Consciousness Studies, 3, No. 3, 1996, pp. 211-16. Also at http://www.imprint.co.uk/online/baars.html

Bacon 1620: Francis Bacon, *Novum Organum,* at http://etext.library.adelaide.edu.au/b/bacon/francis/organon/

Battro 2000: Antonio M. Battro, *Half a Brain is Enough,* Cambridge University Press, Nov. 2000, ISBN 0521783070. Also see *http://www.nobel.se/medicine/laureates/1981/sperry-lecture.html*.

Blakemore 1999: Susan Blackmore, *The Meme Machine,* Oxford (1999), ISBN 019286212X.

Bowlby 1973. John Bowlby [1907-1990]*, Attachment, Basic Books, N.Y. 1973,* ISBN 0465005438.

Bowlby 1973b: John Bowlby, *Separation p26 and p59. Basic Books, N.Y. 1973 ISBN 0465-076912*

Calvin 1966: William H. Calvin, *How Brains Think,* Basic Books, 1966.

Calvin 1994: William H. Calvin and George A. Ojemann, *Conversations with Neil's Brain, The Neural Nature Of Thought And Language,* Basic Books, ISBN 0201483378. Also at *http://williamcalvin.com/bk7/bk7.htm*.

Carlson 1985: Shawn Carlson, "A Double-Blind Test of Astrology," Nature, vol. 318, p.419 (5 Dec 1985)

Carroll 2003: Adapted from the entry on *Cold Reading* by Bertram Forer, in Robert Todd Carroll, *The Skeptic's Dictionary: A Collection of Strange Beliefs, Amusing Deceptions, and Dangerous Delusions,* Wiley 2003 ISBN:

0471272426. Text at http://skepdic.com/coldread.html

Chalmers 1995: David J. Chalmers, "Facing Up to the Problem of Consciousness," *Journal of Consciousness Studies* 2(3), 200-19, 1995. At http://www.u.arizona.edu/~chalmers/papers/facing.html.

Chalmers 1995b: David J. Chalmers, "The Puzzle of Conscious Experience" *Scientific American,* December 1995 pp. 62-68.

Chandler 2004: Keith Chandler, *Australian Journal of Parapsychology,* June 2004, Vol. 4, No. 1. Also at http://www.keithchandler.com/Essays/Savant_Syndrome.html

Charniak 1972: Eugene Charniak, *"Toward a Model Of Children's Story Comprehension,"* PhD thesis 1972, MIT, MIT Artificial Intelligence Laboratory Technical Report TR-266. Also at ftp://publications.ai.mit.edu/ai-publications/pdf/AITR-266.pdf

Clynes 1978: Manfred Clynes, *Sentics*, New York: Doubleday, 1978

Damasio 1995: Antonio R. Damasio, *Descartes' Error*, Avon Books, Nov 1995, ISBN: 0380726475

Darwin 1871: Charles Darwin, *The Descent of Man, 1871,* Simon&Schuster, 1986. Text at *http://www.infidels.org/library/historical/charles_darwin/descent_of_man*

Darwin 1872: Charles Darwin, *Expression of The Emotions In Man And Animals,* 1872, Paul Ekman (ed), *Oxford, 1998, ISBN 0195112717*

Davies 1992: Robertson Davies, *Tempest-Tost,* Penguin*, 1992, ISBN: 0140167927*.

Dawkins 1986: Dawkins, Richard, "*The Blind Watchmaker: Why the Evidence of Evolution Reveals a Universe Without Design* (W.W. Norton and Company, New York, 1986

Dawkins 1989: Richard Dawkins, *The Selfish Gene,* Oxford University Press, 1989, ISBN 0192860925.

[Ref. To Memes?]

Dennett 1978: Daniel C. Dennett, "Why you can't build a machine that feels pain," *Brainstorms*, MIT Press, Bradford Books, 1978, 190-229, ISBN 0262540371.

Dennett 1984: Daniel C. Dennett, *Elbow Room: The Varieties of Free Will Worth Wanting,* Bradford Books, ISBN 0262540428.

Dennett 1988: Daniel Dennett, "Quining Qualia," in A. Marcel and E. Bisiach (eds.), *Consciousness in Modern Science*, Oxford University Press 1988. Reprinted in A. Goldman, ed. *Readings in Philosophy and Cognitive Science*, MIT Press, 1993. Text at *http://cogprints.org/254/00/quinqual.htm*.

Dennett 1991: Daniel C. Dennett, *Consciousness Explained*, Little Brown,

1991, ISBN 0-713-99037-6

Dennett 1992: Daniel C. Dennett, and Marcel Kinsbourne, "Time and the Observer," *Behavioral and Brain Sciences* 15(2): pp183-247, 1992. Also at http://cogprints.ecs.soton.ac.uk/archive/00000264/

Dennett 1992b, Daniel C. Dennett "The Self as a Center of Narrative Gravity," in F. Kessel, P. Cole and D. Johnson (eds.) *Self and Consciousness: Multiple Perspectives*. Hillsdale, NJ: Erlbaum. Also at http://cogprints.org/266/00/selfctr.htm.

Dennett 1995: Daniel C. Dennett, *Darwin's Dangerous Idea*, Simon & Schuster, 1995, ISBN 068482471

Descartes 1637: Rene Descartes, in Discours de la méthode. *"Et le second est que, bien qu'elles fissent plusieurs choses aussi bien, ou peut-être mieux qu'aucun de nous, elles manqueraient infailliblement en quelques autres, par lesquelles on découvrirait qu'elles n'agiraient pas par connaissance, mais seulement par la disposition de leurs organes. Car, au lieu que la raison est un instrument universel, qui peut servir en toutes sortes de rencontres, ces organes ont besoin de quelque particulière disposition pour chaque action particulière; d'où vient qu'il est moralement impossible qu'il y en ait assez de divers en une machine pour la faire agir en toutes les occurrences de la vie, de même façon que notre raison nous fait agir."*

Drescher 1991: Gary Drescher, *Made-Up Minds*, MIT Press 1991, ISBN: 0262041200

Egan 1998: Greg Egan *Diaspora,* Millennium Press, 1998, ISBN 0752809253

Einstein 1950: Albert Einstein, *Out of My Later Years*, Philosophical Library, New York, 1950, pp. 15 - 20.

Evans 1963: Thomas G. Evans, *A Heuristic Program to Solve Geometric-Analogy Problems*, MIT PhD Thesis, 1963, abridged version in Minsky 1968, pp. 271-353.

Feigenbaum1963: *Computers and Thought*, Edward. A. Feigenbaum and Julian Feldman (eds.), McGraw-Hill, New York, 1963.

Feynman 1965: Richard Feynman*, The Character of Physical Law,* MIT Press, Cambridge, MA, 1965. ISBN 0262560038, *p168*

Fodor 1992: Fodor, J. A., "The big idea: Can there be a science of mind?" *Times Literary Supplement*. pp. 5-7, July, 1992

Fodor 1998: Jerry Fodor, "The Trouble with Psychological Darwinism," *London Review of Books,* Vol. 20 No. 2, 22 January 1998 Also at http://www.lrb.co.uk/v20/n02/contents.html

Franklin 1772: Benjamin Franklin, Letter to Joseph Priestly, *19 Sept. 1772.* Text at http://www.historycarper.com/resources/twobf3/letter11.htm.

Freud 1920: Sigmund Freud, *A General Introduction to Psychoanalysis,* 1920, p259, Liveright 1943 ASIN: B0007HEA6Q.

Friedrick-Cofer 1986: Friedrick-Cofer, L., & Huston, A. C., (1986), Television Violence and Aggression: The Debate Continues, *Psychological Bulletin,* 100, pp364-371.

Gardner 2000: Howard Gardner, *Intelligence Reframed: Multiple Intelligences for the 21st Century,* New York: Basic Books, 2000.

Goodall 1968: Jane van Lawick-Goodall, "The behavior of Free-living Chimpanzees in the Gombe Stream Reserve," Anim. Behav. Monogr. I: 161-311, 1968

Gregory 1998: Richard Gregory, "Brainy Mind,*" Brit. Med. Journal 1998 317:1693.* Also at www.richardgregory.org/papers/brainy_mind/brainy-mind.htm.

Gunkel 2006: Patrick Gunkel's ideas about ideas can be seen at http://ideonomy.mit.edu.

Haase 1986: Kenneth W. Haase, PhD thesis, "Exploration and Invention in Discovery," MIT 1986, at http://web.media.mit.edu/~haase/thesis.

Haase 1986a: Kenneth W. Haase, "Discovery Systems," In *Advances in Artificial Intelligence*, European Conference on Artificial Intelligence, North-Holland, 1986.

Haase 1987: Kenneth W. Haase. Typical: *A knowledge representation system for automated discovery and inference*, Technical Report 922, MIT Artificial Intelligence Laboratory, 1987.

Haber 1979: R.N. Haber in *Behavioral and Brain Sciences*, 2, pp583-629, 1979.

Hadamard (1945). Jacques Hadamard. *The Psychology of Invention in the Mathematical Field*, Dover, 1945, ISBN: 0486201074

Harlow 1958: Harry Harlow, *American Psychologist*, 13, 573-685, 1958, http://psychclassics.yorku.ca/Harlow/love.htm.

Hayes 1997: PSYCHE Discussion Forum, *29 Sep 1997, The onset of consciousness in speech,"* http://listserv.uh.edu/cgi-bin/wa?A2=ind9709&L=psyche-b&T=0&F=&S=&P=5262

Hinde 1971: Hinde, R. A. & Spencer-Booth, Y. (Towards understanding individual differences in rhesus mother-infant interaction, *Animal Behaviour,* 19, 165-173.

Hoffman 1994. Howard Hoffman, *Amorous Turkeys and Addicted Ducklings*, Authors Cooperative, ISBN 0-9623311-7-1, 1994

Horner 1998: John R. Horner and James Gorman, *Digging Dinosaurs*, Harper and Row, 1998, ISBN 060973145. See Chapter 4.

Hume 1748: David Hume, *An Enquiry Concerning Human Understanding*.

Hume 1757: David Hume, *The Natural History Of Religion.* http://www.soci.niu.edu/~phildept/Dye/NaturalHistory.html

James 1890: William James, *The Principles of Psychology*, Simon & Schuster, 1997, ISBN: 06844842971. Text at http://psychclassics.yorku.ca/James/Principles/preface.htm

James 1902: William James, *The Varieties of Religious Experience,* Random House 1994, ISBN 067960075–2.

Jamison 1994: Kay Redfield Jamison, *"Touched With Fire: Manic-Depressive Illness and the Artistic Temperament," pp 47-48,* The Free Press, Simon & Schuster, New York, 1994, ISBN 068483183–X.

Jamison 1995: Kay Redfield Jamison, "Manic-Depressive Illness and Creativity," Sci. Amer., Feb. 1995 V. 272 No. 2 Pp. 62-67

Johnston 1997: Elizabeth Johnston "Infantile Amnesia" at http://pages.slc.edu/~ebj/IM_97/Lecture6/L6.html

Kaiser 2006: Peter Kaiser, *The Joy of Visual Perception*, at http://www.yorku.ca/eye/toc.htm has a demonstration at www.yorku.ca/eye/disapear.htm

Kant 1787: Immanuel Kant, Introduction to *Critique Of Pure Reason.*

Koestler 1964: Arthur Koestler, *The Act of Creation,* MacMillan 1964.

Korzybski 1933: Alfred Korzybski, *Science and Sanity*, ISBN 0937298018. Text at http://www.esgs.org/uk/art/sands.htm

Laird 1987: John Laird, Allen Newell, and Paul S. Rosenbloom. Soar: An architecture for general intelligence, *Artificial Intelligence*, 33(1), 1987. See also http://ai.eecs.umich.edu/soar/sitemaker/docs/misc/GentleIntroduction-2006.pdf

Lakoff 1980: George Lakoff and Mark Johnson, *Metaphors We Live By*. Chicago: University of Chicago Press, 1980.

Lakoff 1992: George Lakoff, "The Contemporary Theory of Metaphor," Ortony, Andrew (ed.) Metaphor and Thought (2nd edition), Cambridge University Press, 1993, ISBN 0521405610. Also at http://www.ac.wwu.edu/~market/semiotic/lkof_met.html.

Landauer 1986: Thomas K. Landauer, "How much do people remember?" *Cognitive Science,* 10, 477-493, 1986.

Langley 1987: Pat Langley, Herbert A. Simon, Gary L. Bradshaw, and Jan M. Zytkow. *Scientific Discovery: Computational Explorations of the Creative Processes*. MIT Press, 1987.

Lawler 1985: Robert W. Lawler, *Computer Experience and Cognitive*

Development: A Child's Learning in a Computer Culture. John Wiley & Sons, 1985, *ISBN 0470201940.*

Lenat 1976: Douglas B. Lenat. *AM: An Artificial Intelligence Approach to Discovery in Mathematics as Heuristic Search*, PhD thesis, Stanford University, 1976.

Lenat 1983: Douglas B. Lenat and Jon S. Brown. Why AM and Eurisko Appear to Work, *Artificial Intelligence*, 23, 1983.

Lenat 1983b: Douglas B. Lenat. Eurisko: A program which learns new heuristics and domain concepts, *Artificial Intelligence*, 21, 1983. Also at http://web.media.mit.edu/~haase/thesis/node52.html

Lenat 1990: Douglas B. Lenat and Mary Shepard. *CYC: Representing Encyclopedic Knowledge*, Digital Press, 1990.

Lenat 1997: Douglas Lenat, *Artificial Intelligence as Common Sense Knowledge,*' http://www.leaderu.com/truth/2truth07.html

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

Lewis 1982: F.M. Lewis, "Experienced personal control and quality of life in late stage cancer patients. *Nursing Research,* 31(2) 113-119, 1982

Lewis 1995: Michael Lewis, *Shame, The Exposed Self, 1991, 1995,* The Free Press, Simon & Schuster New York, ISBN 068482311.

Lewis 1995b: Michael Lewis, *"Self-conscious Emotions,"* American Scientist vol. 83, Jan 1995

Lorenz 1970: Konrad Lorenz, *Studies in Animal and Human Behaviour,* Vol I, p132, Harvard Univ. Press, 1970, ISBN: 0674846303.

Lovecraft 1926: H.P. Lovecraft, *The Call of Cthulhu and other weird stories*, S.T. Joshi (ed), Penguin Books, 1999, ISBN 017118234-2

Luria 1968: Alexander R. Luria, *The Mind of a Mnemonist*: Harvard University Press, 1968. ISBN: 0809280078

McCarthy 1959: John McCarthy "Programs with Common Sense," in *Proc. Symposium on Mechanization of Thought Processes*, Vol 1, pp5-27. D.V. Blake and A.M. Uttley (eds.), Natl. Physical Lab., Teddington, England, HMSO, London, 1959. Also at http://www-formal.stanford.edu/jmc/mcc59.html

McCurdy 1960: Harold G. McCurdy, *The Childhood Pattern of Genius.* Horizon Magazine, May 1960, pp. 32-38.

McDermott 1992: Drew McDermott. In *comp.ai.philosophy, 7 Feb 1992.*

Meltzoff 1997: Andrew N. Meltzoff and M. Keith Moore, Explaining Facial Imitation: A Theoretical Model, *Early Development and Parenting,* Vol. 6, 179-192 (1997). Also at http://ilabs.washington.edu/meltzoff/pdf/97Meltzoff_Moore_FacialImit.pdf

Melzack 1965: Ronald Melzack and Patrick Wall, *in "Pain Mechanisms: A New Theory", Science, 150 p.975, 1965. Also, see Tania Singer et al* at http://www.fil.ion.ucl.ac.uk/~tsinger/publications/singer_science_2004.pdf.

Melzack 1993: "Pain: Past, Present and Future," Ronald Melzack, *Canadian Journal of Experimental Psychology* 1993, 47:4, 615-629.

Merkle 1988: See Ralph Merkle's description at http://www.merkle.com/humanMemory.html.

Miller 1960: George A. Miller, Eugene Galanter, and Karl Pribram, *Plans and the Structure of Behavior,* nternational Thomson Publishing; 1960, ASIN: 0030100755

Minsky 1953: Neural-Analog Networks and the Brain-Model Problem. PhD thesis, Mathematics Dept., Princeton University.

Minsky 1956: M. L. Minsky, "Heuristic Aspects of the Artificial Intelligence Problem," Lincoln Lab., M.I.T., Lexington, Mass., Group Rept. 34-55, ASTIA Doc. No. 236885, December 1956. (M.I.T. Hayden Library No. H-58.) Minsky 1968: Marvin Minsky (ed) *Semantic Information Processing, MIT Press,* 1968, ISBN: 0262130440. This anthology is currently out of print, but my chapter, "Matter, Mind, and Models," is at http://web.media.mit.edu/~minsky/papers/MatterMindModels.html.

Minsky 1969: Marvin Minsky and Seymour Papert, *Perceptrons,* MIT Press, 1969

Minsky 1971: Marvin Minsky and Seymour Papert, *Progress Report on Artificial Intelligence,* at http://web.media.mit.edu/~minsky/papers/PR1971.html. Also in *Artificial Intelligence*, Univ. of Oregon Press, 1974, out of print. A poorly scanned image is at ftp://publications.ai.mit.edu/ai-publications/pdf/AIM-252.pdf

Minsky 1974: Marvin Minsky, *A Framework for Representing Knowledge,* MIT, 1974. Also at http://web.media.mit.edu/~minsky/papers/Frames/frames.html Minsky 1977: Marvin Minsky, "Plain Talk About Neurodevelopmental Epistemology," in Proceedings of the Fifth International Joint Conference on Artificial Intelligence, Cambridge, Mass., 1977

Minsky 1980: Marvin Minsky, "Jokes and their Relation to the Cognitive Unconscious, " In *Cognitive Constraints on Communication*, Vaina and Hintikka (eds.) Reidel, 1981, ISBN: 9027714568. Also at web.media.mit.edu/~minsky/papers/jokes.cognitive.txt.

Minsky 1980b: Minsky M: 'K-lines, a theory of memory', Cognitive Science, 4, 1980, pp 117-133.

Minsky 1981: Marvin Minsky, "Music, Mind, and Meaning," *Computer*

Music Journal, Fall 1981, Vol. 5, Number 3. Also at web.media.mit.edu/~minsky/papers/MusicMindMeaning.html Minsky 1986: Marvin Minsky, *The Society of Mind*, Simon & Schuster, 1986, ISBN: 0671657135.

Minsky 1988: Marvin Minsky and Seymour Papert, *Perceptrons,* (2nd edition) MIT Press, 1988.

Minsky 1991: Marvin Minsky, "Logical vs. Analogical or Symbolic vs. Connectionist or Neat vs. Scruffy", in *Artificial Intelligence at MIT, Expanding Frontiers,* Patrick H. Winston (Ed.), Vol. 1, MIT Press, 1990. Reprinted in AI Magazine, 1991. Also at http://web.media.mit.edu/~minsky/papers/SymbolicVs.Connectionist.html Minsky 1992: Marvin Minsky, F*uture of AI Technology,* Toshiba Review, Vol.47, No.7, July 1992. Also at http://web.media.mit.edu/~minsky/papers/CausalDiversity.html Minsky 2000: Marvin Minsky, "Attachments and Goals," in Hatano, G.,Tanabe, H., and Okado, N., eds.. *Affective Minds.* Amsterdam: Elsevier 2000.

Minsky 2000b: Marvin Minsky, Future Models for Mind-Machines," in Sloman et al., (eds*), Proceedings Symposium on How to Design a Functioning Mind*, AISB00 Convention, pp124--129, 2000.

Minsky 2004: Marvin Minsky, Push Singh, and Aaron Sloman (2004). The St. Thomas common sense symposium: designing architectures for human-level intelligence. *AI Magazine*, Summer 2004, 25(2):113-124. http://web.media.mit.edu/~push/StThomas-AIMag.pdf Minsky 2005: Marvin Minsky. "Interior Grounding, Reflection, and Self-Consciousness," Proceedings of an International Conference on Brain, Mind and Society, Tohoku Univ., Japan, Sept 2005. http://www.ic.is.tohoku.ac.jp/~GSIS/. Also at http://web.media.mit.edu/~minsky/papers/Internal%20Grounding.html

Mooers 1956: Calvin N. Mooers, "Information retrieval on structured content," in *Information theory* (C. Cherry, editor, Butterworths, 1956).

Mueller 2006: Erik T. Mueller, *Commonsense Reasoning*, Morgan Kaufmann 2006, ISBN 0-12-369388-8,

Mueller 2006" Erik T. Mueller, *Commonsense Reasoning*, Morgan Kaufmann 2006, ISBN 0-12-369388-8,

Nelson 2001: Charles A. Nelson, "The Development and Neural Bases of Face Recognition," Infant and Child Development, 10: 3–18, 2001. Also at http://www.biac.duke.edu/education/courses/spring03/cogdev/readings/C.A.%20Nelson%20(2001).pdf

Nelson 2001: Charles A. Nelson, "The Development and Neural Bases of Face Recognition Infant and Child Development," *Infant and Child*

Development, 10: 3–18 (2001) Newell 1955: Allen Newell, "*The chess machine,"* in Proc. Western Joint Computer Conf. March 1955.

Newell 1960: Newell, A., J. C. Shaw, and H. A. Simon, "Report on a general problem solving program," in *Proceedings of the International Conference on Information Processing.* UNESCO, Paris, 1960 pp. 256-64.

Newell 1960b: 'Allen Newell. J. Clifford Shaw, and Herbert A. Simon, "A variety of intelligent learning in a general problem solver," in *Self-Organizing Systems*, M. T. Yovitts and S. Cameron (eds.), Pergamon Press, New York, 1960. I have seen no references to this profound idea since its original publication!

Newell 1963: Allen Newell and Herbert A. Simon. "GPS, a program that simulates human thought," *Computers and Thought*, E. A. Feigenbaum and J. Feldman (eds.), McGraw-Hill, New York, 1963.

Newell 1972: Allen Newell and Herbert A. Simon (1972), *Human Problem Solving,* Prentice Hall, June 1972, ASIN: 0134454030. Also, see http://sitemaker.umich.edu/soar for a description of the problem-solving architecture called *SOAR*.

O'Regan 2006: See Kevin O'Regan's articles on *"Change-Blindness"* at http://nivea.psycho.univ-paris5.fr/ECS/ECS-CB.html and *"Change blindness as a result of mudsplashes."* in Nature Aug. 2, 1998

Ortony 1988: Andrew Ortony, Gerald L. Clore, and Allan Collins, *The Cognitive Structure of the Emotions*, New York, Cambridge University Press, 1988, ISBN 0521386640.

Osterweis 1987: Marian Osterweis, Arthur Kleinman, and David Mechanic, *"Pain and disability: Clinical, Behavioral, and Public Policy Perspectives."* National Academy Press, 1987

Pagels 1988: Heinz Pagels, *The Dreams of Reason,* Simon &Schuster 1988 ISBN: 0-671-62708-2

Panksepp 1988. *Emotions and Psychopathology*, Clynes & Panksepp (eds.), 1988, Plenum Pubs ISBN: 0306429160

Panksepp 1998: Jaak Panksepp. *Affective Neuroscience*, Oxford, 1998, ISBN 0195096738.

Pepperberg 1998: See reports on Irene Pepperberg's research on parrots at http://www.alexfoundation.org/irene.htm. Also at http://pubpages.unh.edu/~jel/video/alex.html Perkins 1981): David N. Perkins, DN () *The Mind's Best Work.* Cambridge, MA: Harvard University Press, 1981

Phillips 2004: Melissa Lee Phillips, "Seeing with New Sight," at http://faculty.washington.edu/chudler/visblind.html

Piaget 1924: Jean Piaget, *The Language and Thought of the Child*, Rutledge (2001) ISBN 0415267501

Pivar 1964: Pivar, M. and Finkelstein, M. in The Programming Language LISP: Its Operation and Applications, MIT Press, Cambridge, Mass. 1944. Text at

http://community.computerhistory.org/scc/projects/LISP/book/III_LispBook_Apr

Plsek 1996: Paul E. Plsek, Models for the Creative Process, at www.directedcreativity.com/pages/WPModels.html

Pohl, 1970: This situation is described in the title story of Frederik Pohl's anthology, *Day Million,* Ballantine Books 1970 ISBN 0330236067.

Poincare 1913: Henri Poincaré. *The Foundations of Science, Science and Hypothesis, The Value of Science, Science and Method*, The Science Press, 1929, ISBN: 0819123188

Polya 1954: George Polya*, Induction and Analogy in Mathematics,* Princeton Univ. Press, ISBN: 0691025096.

Polya 1962: G. Polya. *Mathematical Discovery*. John Wiley and Sons, 1962.

Proust 1927: Proust, Marcel. Remembrance of Things Past, New York: Random House 1927-32

Pylyshyn 1998: Zenon Pylyshyn, *Is Vision Continuous With Cognition, discusses many issues related to the structure of the visual system, at* http://ruccs.rutgers.edu/faculty/ZPbbs98.html.

Quillian 1966: Ross Quillian, *Semantic Memory,* Ph.D. thesis, Carnegie Institute of Technology, Pittsburgh, Pennsylvania, February 1966. Reprinted in Minsky 1968.

Ramachandran 2004: V.S. Ramachandran, *Science, Vol 305 no 5685, 6 August 2004*.

Rosenfeld 1996: Ronald Rosenfeld, "A maximum entropy approach to adaptive statistical language modeling," *Computer, Speech and Language, 10, 1996*. Also at http://www.cs.cmu.edu/afs/cs/user/roni/WWW/me-csl-revised.ps.

Royce 1908: Josiah Royce, *The Philosophy of Loyalty, Vanderbilt Univ. Press, 1995,* ISBN 0826512674

Ryle 1949: Gilbert Ryle, *The Concept of Mind*, The University of Chicago Press, 1949

Samuel 1959: Arthur L. Samuel, "Some studies in machine learning using the game of checkers," IBM J. Res. Dev., vol. 3, pp. 211-219, July 1959.

Schaffer 1964: H.R. Schaffer and P. E. Emerson in 'The development of social attachments in infancy,' Monogr. Soc. Res. Child Dev., 29, 3, 1964.

Schank 1975: Roger C. Schank, *Conceptual Information Processing*, Elsevier Science Publishers 1975. ISBN: 0444107738.

Schank 1977: Roger Schank and Robert Abelson, Scripts, Goals, Plans and Understanding," Erlbaum Associates, 1977.

Schank 1990: *Tell Me a Story,* Charles Scribner's Sons, New York, 1990. Reissued, Northwestern University Press, 1995, ISBN 0810113139.

Seay 1964: B. Seay, B. R. Alexander, & H. F. Harlow, *Maternal behavior of socially deprived rhesus monkeys,* J. Abnormal and Social Psychology, 69, 345-354.

Seckel 2004: Masters of Deception, Sterling Publishing, New York, ISBN 402705778.

Shannon 1948: Claude E. Shannon, "A Mathematical Theory of Communication," Bell System Technical Journal, vol. 27, pp. 379-423 and 623-656, July and October 1948. Also at http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf

Sindbad 1918: Anonymous, : *The Arabian Nights Entertainments*, Longmans, Green and Co, 1918 (1898). http://www.sacred-texts.com/neu/lang1k1/tale21.htm .

Singh 2002: Push Singh, Thomas Lin, Erik T. Mueller, Grace Lim, Travell Perkins and Wan Li Zhu, "Open Mind Common Sense: Knowledge acquisition from the general public," *Proceedings of the First International Conference on Ontologies, Databases, and Applications of Semantics for Large Scale Information Systems*. Irvine, CA.

Singh 2003: Push Singh, Examining the Society of Mind. *Computing and Informatics*, 22(5):521-543. 2003. This article briefly describes the history of the Society of Mind theory, explains some of its essential components, and relates it to recent developments in Artificial Intelligence. Also at http://web.media.mit.edu/~push/ExaminingSOM.pdf .

Singh 2003a: Push Singh and Marvin Minsky, An architecture for combining ways to think. *Proceedings of the International Conference on Knowledge Intensive Multi-Agent Systems*. Cambridge, MA. (2003).

Singh 2003b: A preliminary collection of reflective critics for layered agent architectures. *Proceedings of the Safe Agents Workshop (AAMAS 2003)*. Also at http://web.media.mit.edu/~push/ReflectiveCritics.pdf

Singh 2003c: Push Singh and William Williams (2003). **LifeNet: a propositional model of ordinary human activity**. *Proceedings of the Workshop on Distributed and*

Collaborative Knowledge Capture (DC-KCAP) at K-CAP 2003. Sanibel Island, FL. At http://web.media.mit.edu/~push/LifeNet.pdf

Singh 2004: Push Singh, Marvin Minsky and Ian Eslick, *Computing commonsense*, BT Technology Journal, Vol 22 No 4, October 2004. Also at http://web.media.mit.edu/~push/Computing-Commonsense-BTTJ.pdf   Singh 2005: Push Singh, *EM-ONE: An Architecture for Reflective Commonsense Thinking*,   PhD   thesis,   MIT,   June   2005.   Also   at http://web.media.mit.edu/~push/push-thesis.pdf

Singh 2005b: Push Singh and Marvin Minsky. An architecture for cognitive diversity, *Visions of Mind*, Darryl Davis (ed.), London: Idea Group Inc. (2005). http://web.media.mit.edu/~push/CognitiveDiversity.html .

Sloman 1992: Aaron Sloman, "Developing concepts of consciousness." *Behavioral and Brain Sciences 14, 1992.*

Sloman 1994:Aaron Sloman, in newsgroup *comp.ai.philosophy,* 14 Dec. 1994

Sloman 1996*:* Aaron Sloman*,* in newsgroup *sci.psychology.consciousness,* 19 Jun 1996. See also *A Systems Approach to Consciousness*, at http://www.cs.bham.ac.uk/~axs/misc/consciousness.rsa.text, with lecture slides at http://www.cs.bh.am.ac.uk/~axs/misc/consciousness.lecture.ps.

Solomonoff 1957: Raymond J. Solomonoff, "An Inductive Inference Machine," *IRE Convention Record, Section on Information Theory, Part 2*, pp. 56-62, 1957.

Solomonoff 1964: Solomonoff, R. J. "A formal theory of inductive inference," *Information and Control*, 7 (1964), pp.1-22;

Solomonoff 1997: Solomonoff, R. J. "The Discovery of Algorithmic Probability," *Journal of Computer and System Sciences*, Vol. 55, No. 1, 1997, at http://world.std.com/~rjs/barc97.html.

Spencer-Booth 1971: Y. Spencer-Booth and R. A. Hinde, *Animal Behavior,* 19, 174-191 and 595-605, 1971

Spencer-Brown 1972: G. Spencer-Brown, *Laws of Form,* Crown Pub. 1972, ISBN: 0517527766

Sri   Chinmoy   2003: http://www.yogaofsrichinmoy.com/the_higher_worlds/consciousnesss/

Stickgold 2000: Robert Stickgold, April Malia, Denise Maguire, David Roddenberry, Margaret O'Connor, "Replaying the Game: Hypnagogic Images in Normals and Amnesics," *Science,* Volume 290, Number 5490, 13 Oct 2000, pp. 350-353.

Thagard 2001: Paul Thagard, "How to make decisions: Coherence, emotion,

and practical inference," In E. Millgram (Ed.), *Varieties of practical inference,* MIT Press. 355-371. text at http://cogsci.uwaterloo.ca/Articles/Pages/how-to-decide.html Thorndike 1911: Edward L. Thorndike. *Animal Intelligence*. New York: Macmillan 1911, p. 244.

Tinbergen 1951: Nikolaas Tinbergen, *The Study of Instinct,* Oxford University Press, London 1951.

Turing 1936: Alan Turing, ""On Computable Numbers," http://www.abelard.org/turpap2/tp2-ie.asp#section-1

Turing 1950: Alan Turing, "Computing Machinery and Intelligence." *Mind 49: pp*433-460, 1950. Also at http://cogprints.org/499/00/turing.html and at www.cs.swarthmore.edu/~dylan/Turing.html

Viezzer 2000: Manuela Viezzer, Ontologies and Problem-solving methods, ECAI 2000, 14th European Conference on Artificial Intelligence, August 2000, Humboldt University, Berlin. Also at www.cs.bham.ac.uk/~mxv/publications/onto_engineering.

Vinacke 1952: W.E. Vinacke, *The Psychology of Thinking,* 1952, McGraw Hill

Waltz 1985: David L. Waltz and Jordan Pollack in "Massively Parallel Parsing," Cognitive Science, 9(1), 1985.

Watts 1960: Alan Watts, *This is It,* Random House, 1960, ISBN 0394719042 , pp. 32-33

Wertheimer 1945: Max Wertheimer, *Productive Thinking*. New York: Harper 1945.

West 1928: Rebecca West, *The Strange Necessity,* Doubleday, 1928, ISBN 0781270626.

Wilde 1905: Oscar Wilde, *De Profundis*, Methuen and Co., 1905.

Winston 1970: Patrick Winston, *Learning structural descriptions from examples,* AITR -231, PhD thesis, Cambridge, Mass.: MIT AI Lab, 1970. ftp://publications.ai.mit.edu/ai-publications/pdf/AITR-231.pdf Winston 1975: Patrick H. Winston (ed.), *The Psychology of Computer Vision*. New York: McGraw-Hill, 1975.

Winston 1984: Patrick H. Winston. *Artificial Intelligence*. Addison-Wesley, Third Edition, 1992 1984. (1984 first edition is easier for beginners.) Wundt 1897: Wilhelm Wundt *Outlines of Psychology*. Translated by C.H. Judd. Text at http://psychclassics.yorku.ca/Wundt/Outlines/

Zealand 2001: *Around the clock*, Statistics, New Zealand Time Use 1998-99, at http://www.stats.govt.nz/analytical-reports/time-use-survey.htm

Notes

1

See http://www.english.uiuc.edu/maps/poets/m_r/parker/lightverse.htm

2

Barry Took and Marty Feldman, *Round the Horne,* BBC Radio, 1966

3

This list is adapted from a note from Aaron Sloman in comp.ai.philosophy, 16/5/1995.

4

Nikolaas Tinbergen, *The Study of Instinct,* Oxford University Press, London, 1951. [See *§§Tinbergen's Theory*.]

5

In *The Strange Necessity, 1928.* **ISBN:** 0781270626.

6

See Glossary: *Cross-Exclusion*.

7

However, I recommend Aaron Sloman's discussion of this in http://www.cs.bham.ac.uk/~axs/misc/talks/gatsby.slides.pdf.

8

In *Girl, Interrupted,* Vintage Books, 1994, pp. 137-143.

9

Such an 'all-or-none' view of what 'understand' means can be seen at http://home.hanmir.com/~prolog/ai/mind.html or in *"Minds, Brains, and Computers"* by John Searle, in *Philosophy: The Quest for Truth*. Oxford Univ. Press; 5th edition, **ISBN:** 0195156242

10

R. Feynman, *The Character of Physical Law*, Modern Library, 1994, **ISBN:** 0679601279. Note that when scientists say that two are representations are 'equivalent,' they do not mean to suggest that both are equally practical.

11

Michael Lewis, "Self-conscious Emotions," American Scientist vol. 83, Jan 1995.

12

http://etext.library.adelaide.edu.au/a/a8rh/

13

This could relate to some psychoanalytic theories, which argue that such objects might help to make transition from early attachments to other kinds of relationships. See, for example, *www.mythosandlogos.com/Klein.htm*.

14

See §Memes: Dawkins, Henson, Blackmore.

15

[John Bowlby, Attachment, Basic Books, N.Y. 1973, p. 217]

16

ibid. Bowlby bases this on some research of H.R. Schaffer and P. E. Emerson, '*The development of social attachments in infancy,*' Monogr. Soc. Res. Child Dev., 29, 3, 1-77, 1964.

17

John Bowlby, Separation p26. Basic Books, N.Y. 1973 ISBN 465-07691-2

18

Harry Harlow, *American Psychologist*, 13, 573-685, 1958, http://psychclassics.yorku.ca/Harlow/love.htm.

19

Jane van Lawick-Goodall, 'The behavior of Free-living Chimpanzees in the Gombe Stream Reserve,' Anim. Behav. Monogr. I: 161-311, 1968

20

In 1973, Konrad Lorenz and Nikolaas Tinbergen shared a Nobel Prize for these and other discoveries.

21

There also is some evidence that imprinting resembles addiction. For example, Jaak Panksepp's [1988] experiments suggest that separation-distress may be similar to pain, because it is relieved by opiods. Howard Hoffman [1994] speculates that an object can become an Imprimer when certain aspects of its motion or shape arouse an innate mechanism that releases endorphins in the imprintee's brain, and he conjectures that the resulting feelings of pleasure or comfort then somehow cause the object to be classified as 'familiar' enough to overcome other fearful reactions. In §9-x-Pleasure I'll suggest that such feelings may play a somewhat less direct role.

22

Y. Spencer-Booth and R. A. Hinde, *Animal Behavior,* 19, 174-191 and 595-605, 1971

23

S. Seay, 1964

24

[See Chapter 4 of *Digging Dinosaurs*, John R. Horner and James Gorman,

Harper and Row, 1998, ISBN -06-097314-5.]

25

For example, see Charles A. Nelson's article at
http://www.biac.duke.edu/education/courses/spring03/cogdev/readings/C.A.%20

26

Francesca Acerra, Yves Burnod and Scania de Schonen,
http://www.dice.ucl.ac.be/Proceedings/esann/esannpdf/es1999-22.pdf

27

Meltzoff and Moore (1977) appear to have shown that infants can imitate lip protrusion, mouth opening, tongue protrusion, and finger movement. See http://ilabs.washington.edu/meltzoff/pdf/97Meltzoff_Moore_FacialImit.pdf

28

"Studies in Animal and Human Behaviour," Vol I, p. 132, Harvard Univ. Press, 1970

29

Multiple attachments are reported in Schaffer, H.R. and Emerson P.E. (1964) *The development of social attachments in infancy,* Monographs of Social Research in Child Development 29: no. 94. However, I could not find any studies of the long-term effects of having several Imprimers.

30

From a 1961 letter to Mrs. H. L. Austin

31

In *Expression of The Emotions In Man And Animals*

32

In §6-3 we'll say more about what sometimes makes a goal feel like a force.

33

Ronald Melzack and Patrick Wall, *in "Pain Mechanisms: A New Theory",* Science, 150, p. 975, 1965.

34

For example, see www.umass.edu/preferen/mpapers/SingerEmpathy.pdf

35

in "Why you can't build a machine that feels pain," *Brainstorms*, Bradford Books, 1978. This is an ironic title for the deeper idea that 'pain' is a suitcase word that comprises so many ideas and processes that it does not make much technical sense to speak of it as definite kind of entity.

36

See "Pain: Past, Present and Future, " Ronald Melzack, Canadian Journal of Experimental Psychology 1993, 47:4, 615-629.

37

Marian Osterweis, Arthur Kleinman, and David Mechanic, *"Pain and disability: Clinical, Behavioral, and Public Policy Perspectives."* National Academy Press, 1987

38

F.M. Lewis, "Experienced personal control and quality of life in late stage cancer patients. *Nursing Research,* 31(2) 113-119, 1982

39

—From a letter to Lord Alfred Douglas, written during Wilde's imprisonment in Reading.

40

See http://www.counselingforloss.com/article8.htm .

41

"Touched With Fire: Manic-Depressive Illness and the Artistic Temperament," pp 47-48, The Free Press, Macmillan, New York, 1993.

42

Kay Redfield Jamison, "Manic-Depressive Illness and Creativity," Sci. Amer., Feb. 1995 V. 272 No. 2 Pp. 62-67

43

Most animals simply do not have the high-level resources that people have, and this makes it risky to apply to ourselves what we learn from laboratory animals.

44

"Duplication describes a remedy for this."

45

Thus, to ascend from the top of Kilimanjaro to the summit of, say, Mt. Everest, you would have to climb down and then up again.

46

See my essay on Jokes, at web.media.mit.edu/~minsky/papers/jokes.cognitive.txt

47

See the extensive discussion in William James' text, *"The Varieties of Religious Experience."*

48

Sigmund Freud, in *A General Introduction to Psychoanalysis*, 1920, p. 259.

49

For more details of this episode, see §4.5 of SoM.

50

www.srichinmoy.org/html/library/questions_answers/consciousness_qa.htm

51

Fodor, J. A., Can there be a science of mind? *Times Literary Supplement.* July 3, 1992, pp5-7.

52

Chapter 2 of *Conversations with Neil's Brain, REF*

53

In comp.ai.philosophy, 14 Dec. 1994

54

An Enquiry Concerning Human Understanding, 1748

55

In *sci.psychology.consciousness, 15 Jun 96.*

56

There are important exceptions to this. It would seem that experts like J.S. Bach developed ways to accomplish more multiple, yet still similar goals in parallel. However, as their skills improve, most such experts become less and less able to tell the rest of us how they do them.

57

William James discussed this extensively. See: http://psychclassics.yorku.ca/James/jimmy11.htm. Several other more modern ideas about this are developed in Daniel Dennett's 1991 book, *Consciousness Explained.*

58

So, despite a popular intuition, research on parallel processing has shown that such systems are frequently prone to end up accomplishing less for the same

amount of computational power Nevertheless, if that cost can be borne, then the final result may come sooner!

59

See *Parallel Distributed Processing,* Rumelhart, D., J. McClelland et al., MIT Press, Cambridge, MA: 1986. See also my discussion of 'opacity' in http:/*web.media.mit.edu*~minsky/papers/SymbolicVs.Connectionist.html. For some limitations of the most popular forms of neural networks, see [Perceptrons.][Ref.]

60

See Jeffrey Siskind, publication about…

61

Chapter §8 will propose more details about how our memory structures are organized to so swiftly deliver such information. Basically, when a problem arises, some processes may start to solve it before other processes formulate questions about it.

62

See §25.4 of The Society of Mind, p257.

63

" —psyche-b@listserv.uh.edu, 29 Sep 1997.

64

In *Outlines of Psychology*, 1897.

65

This idea is explained in more detail at http:/*web.media.mit.edu*~minsky/papers/MatterMindModels.html.

66

In a discussion on the newsgroup *comp.ai.philosophy, 7 Feb 1992.*

67

Daniel Dennett, *Consciousness Explained. [ref.]*

68

[Ref to Metaphor, Lakoff, etc.]

69

I don't think modern programming, on the whole, has reached this stage. Indeed, I did once suggest, very long ago, that a Cartesian Theater concept be a good model of programming. Old design paper]

70

http://www.imprint.co.uk/online/new1.html

71

Dennett, Daniel C and Kinsbourne, Marcel, (1992) Time and the Observer. *Behavioral and Brain Sciences* 15(2): pp183-247. http://cogprints.ecs.soton.ac.uk/archive/00000264/

72

SoM 25.04 Continuity.

73

http://www.u.arizona.edu/~chalmers/papers/facing.html.

74

Ref. to Penrose's book.

75

This example is from Frederik Pohl's prescient short story Day Million in his anthology, Day Million, Ballantine Books 1970 ISBN: **0330236067**

76

More details about construction planning were developed by Scott Fahlman in his 1973 paper at ftp://publications.ai.mit.edu/ai-publications/pdf/AITR-283.pdf

77

In *Principles of Psychology*, p359

78

According to Tinbergen, when an animal can't make a decision, this often results in dropping both alternatives and doing something that seem to be quite irrelevant. However, these "displacement activities" seem to be fixed, so they do not suggest that those animals have thoughtful ways to deal with such conflicts.

79

In *The Natural History Of Religion, 1757.* http://www.soci.niu.edu/~phildept/Dye/NaturalHistory.html

80

Some early steps in that project are described in ftp://publications.ai.mit.edu/ai-publications/pdf/AIM-200.pdf.

81

See http:*/web.media.mit.edu*~minsky/papers/PR1971.html

82

In fact, that darker horizontal streak is *not* the lower edge, but is part of the surface next to that edge, slightly shadowed because that edge is worn-down.

83

V.S. Ramachandran, *Science, v305 no.5685, 6 August 2004.*

84

in www.richardgregory.org/papers/brainy_mind/brainy-mind.htm. See also, www.physiol.m.u-

85

This program was based on ideas of Yoshiaki Shirai (and Manuel Blum). See ftp://publications.ai.mit.edu/ai-publications/pdf/AIM-263.pdf. However, I should add that *Builder* had almost no competence for any but neat geometrical scenes—and, so far as I know, there still are no 'general-purpose vision machines" that can, for example, look around a room and recognize everyday objects therein. I suspect that this is mainly because they lack enough knowledge about real-world objects; we'll discuss this more in Chapter 6.

86

See papers by Adolfo Guzman and David Waltz at ftp://publications.ai.mit.edu/ai-publications/pdf/AIM-139.pdf and

ftp://publications.ai.mit.edu/ai-publications/pdf/AITR-271.pdf

87

See Zenon Pylyshyn, *http://ruccs.rutgers.edu/faculty/ZPbbs98.html*. *[Broken Link]* The octagon example is from Kanizsa, G. (1985). Seeing and Thinking. *Acta Psychologica, 59*, 23-33.

88

In this kind of diagram, each object is represented by a network that describes relationships between its parts. Then each part, in turn, is further described in terms of relationships between *its* parts, etc.,—until those sub-descriptions descend to a level at which each one because a simple list of properties, such as an object's color, size, and shape. For more details, see §§Frames, Quillian's thesis in *Semantic Information Processing*, and Patrick Winston's book, *The Psychology of Computer Vision.*

89

Some persons claim to imagine scenes as though looking at a photograph, whereas other persons report no such vivid experiences. However, some studies appear to show that both are equally good at recalling details of remembered scenes.

90

See, for example, http://www.usd.edu/psyc301/Rensink.htm and http://nivea.psycho.univ-paris5.fr/Mudsplash/Nature_Supp_Inf/Movies/Movie_List.html.

91

This prediction scheme appears in section §6-7 of my 1953 PhD thesis, *"Neural-Analog Networks and the Brain-Model Problem,* Mathematics Dept., Princeton University, Dec. 1953. At that time, I had heard that there were 'suppressor bands' like the one in my diagram, at the margins of some cortical areas. These seem to have vanished from more recent texts; perhaps some brain researchers could find them again.

92

In Push Singh's PhD thesis, [ref] two robots actually consider such questions. Also refer to 2004 BT paper.

93

The idea of a panalogy first appeared in *Bib: Frames,* and more details about this were proposed in chapter 25 of SoM. A seeming alternative might be to have almost-separate sub-brains for each realm—but that would lead to similar questions at some higher cognitive level.

94

I got some of these ideas about 'trans' from the early theories of Roger C. Schank, described in *Conceptual information processing,* Amsterdam: North-Holland, 1975.

95

Tempest-Tost, 1992, ISBN: 0140167927.

96

As suggested in §3-12 we often learn more from failure than from success—because success means you already possessed that skill, whereas failure instructs

us to learn something new.

97

See Douglas Lenat, *The Dimensions of Context Space,* at http://www.ai.mit.edu/people/phw/6xxx/lenat2.pdf

98

This discussion is adapted from my introduction to *Semantic Information Processing,* MIT Press, 1969.

99

From: Alexander R.Luria, *The Mind of a Mnemonist*: Cambridge: Harvard University Press, 1968.

100

Landauer, Thomas K. (1986). "How much do people remember? Some estimates of the quantity of learned information in long-term memory." Cognitive Science, 10, 477-493. See also Ralph Merkle's description of this in http://www.merkle.com/humanMemory.html. Furthermore, according to Ronald Rosenfeld, the information in typical text is close to about 6 bits per word. See Rosenfeld, Ronald, "A maximum entropy approach to adaptive statistical language modeling," *Computer, Speech and Language, 10, 1996*, also at http://www.cs.cmu.edu/afs/cs/user/roni/WWW/me-csl-revised.ps. In these studies, the term 'bit' of information is meant in the technical sense of C.E. Shannon in http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html.

101

My impression that this also applies to the results reported by R.N. Haber in *Behavioral and Brain Sciences*, 2, 583-629,1979.

102

A. M. Turing, *Computing Machinery and Intelligence,* at www.cs.swarthmore.edu/~dylan/Turing.html

103

See several essays about self-organizing learning systems at: Gary Drescher, *Made-Up Mind*s, MIT Press 1991, ISBN: 0262041200; Lenat's 1983 "AM" system at http:/*web.media.mit.edu*~haase/thesis/node52.html; Kenneth Haase's thesis at http:/*web.media.mit.edu*~haase/thesis/; Pivar, M. and Finkelstein, M. (1964) in *The Programming Language LISP, MIT* Press 1966; Solomonoff, R. J. "A formal theory of inductive inference," *Information and Control*, 7 (1964), pp.1-22; Solomonoff, R. J. "An Inductive Inference Machine," *IRE Convention Record, Section on Information Theory, Part 2*, pp. 56-62, 1957. Also, see his essay at http://world.std.com/~rjs/barc97.html. In recent years this has led to a field of research with the name of 'Genetic Programming.'

104

Technically, if a system has already been optimized, then any change is likely to make it worse until one find a higher peak, some distance away in the "fitness space."

105

See §2.6 of *Frames, §27.1 of SoM, and* Charniak, E. C., *Toward a Model of Children's Story Comprehension.* ftp://publications.ai.mit.edu/ai-publications/pdf/AITR-266.pdf

106

There has been some recent progress toward extracting such kinds of knowledge from large number of users of the Web. See Push Singh's 'OpenMind Commonsense' project at http://commonsense.media.mit.edu/.

107

John McCarthy, "Programs with Common Sense," in *Proc. Symposium on Mechanization of Thought Processes*, 1959. Reprinted in *Semantic Information Processing,* p404.

108

People sometimes use 'abstract' to mean 'complex' or 'highly intellectual'— but here I mean almost the opposite: a more abstract description ignores more

details—which makes it more useful because it depends less on the features of particular instances.

109

See Elizabeth Johnston's notes on "Infantile Amnesia" at http://pages.slc.edu/~ebj/IM_97/Lecture6/L6.html

110

In each cycle of operation, the program finds some differences between the current state and the desired one. Then it uses a separate method to guess which of those differences is most significant, and makes a new subgoal to reduce that difference. If this results in a smaller difference, the process goes on; otherwise it works on some other difference. For more details of how this worked, see Newell, A., J. C. Shaw, and H. A. Simon, "Report on a general problem solving program," in *Proceedings of the International Conference on Information Processing*. UNESCO, Paris, pp. 256-64. A more accessible description is in Newell, A., and Simon, H. A., "GPS, a program that simulates human thought," *Computers and Thought*, E. A. Feigenbaum and J. Feldman (Eds.), McGraw-Hill, New York, 1963.

111

See Allen Newell and Herbert Simon (1972), *Human Problem Solving,* Prentice Hall; (June 1972), ASIN: 0134454030. Also, see a problem-solving architecture called "SOAR.". [Ref.]

112

See A. Newell. J. C. Shaw, and H. A. Simon, "A variety of intelligent learning in a general problem solver," in *Self-Organizing Systems*, M. T. Yovitts and S. Cameron, Eds., Pergamon Press, New York, 1960.

113

In Nicomachean Ethics (Book III. 3, 1112b). This appears to be a description of what today we call 'top-down search."

114

This was written before 'security' began to be imposed on trains.

115

See Peter Kaiser's ^{www.yorku.ca/eye/disapear.htm}. [Also, see §§Change-Blindness] However, there are some signals that do not 'fade away.' Because we also have some additional sensors that evolved to keep responding to certain particular harmful conditions. [See §§Alarms.]

116

Roger Schank has conjectured that this may be one of our principal ways to learn and remember—in *"Tell Me a Story "* Charles Scribner's Sons, New York, 1990.

117

There are more details about this in my essay at *web.media.mit.edu~*minsky/papers/MusicMindMeaning.html

118

In *Sentics*, New York: Doubleday, 1978, the pianist-physiologist Manfred Clynes describes certain temporal patterns, each of which might serve as a 'command' to induce a certain emotional state.

119

G. Spencer-Brown, *Laws of Form,* Crown Pub. 1972, ISBN**:** 0517527766

120

One could ask the same questions about gossip, sports, and games. See *How people spend their time*, http://www2.stats.govt.nz/domino/external/pasfull/pasfull.nsf/0/4c2567ef00247c the-clock.pdf

121

In his 1970 PhD thesis, Patrick H. Winston called this a "similarity network." See [AIM xxx].

122

http://www.gutenberg.net/etext94/arabn11.txt

123

Letter to Joseph Priestly, *19 Sept. 1772.*

124

See http://cogsci.uwaterloo.ca/Articles/Pages/how-to-decide.html

125

Section 30.6 of SoM discusses why the idea of free will seems so powerful. There are many more ideas about this in Daniel Dennett's 1984 book, Elbow *Room: The Varieties of Free Will Worth Wanting,* ISBN 0262540428.

126

Section 30.6 of SoM discusses why the idea of free will seems so powerful. There are many more ideas about this in Daniel Dennett's 1984 book, Elbow *Room: The Varieties of Free Will Worth Wanting,* ISBN 0262540428.

127

See the book, Computers and Thought for some of the accomplishments of that period.

128

Evans, Thomas G. (1963) *A Heuristic Program to Solve Geometric-Analogy Problems,* abridged version in Minsky (ed) *Semantic Information Processing, MIT Press* 1968, pp. 271-353.

129

Aristotle, On the Soul, Book I, Part 1.

130

Richard P. Feynman*, The Character of Physical Law,* MIT Press,

Cambridge, MA 1965. ISBN 0262560038, *p168.*

131

Antonio R. Damasio, *Descartes' Error,* Avon Books, Nov 1995, ISBN: 0380726475

132

[A system would also face similar problem if several Selectors were turned on at once? Refer to §Currencies.]

133

At the lowest levels, the Critics and Selectors become the same as the *Ifs* and *Thens* of simple reactions. At the reflective and higher levels, the Critics will tend to engage so many resources that they can't be distinguished from *Ways to Think.* In his essay, *"Reflective Critics,"* Push Singh discusses Critics with such abilities. See http:/*web.media.mit.edu~*push/ReflectiveCritics.pdf

134

Logic can be useful after a problem is solved, for making credit assignments [§8.5] and for solving simplified versions of problems. See §§Logic.

135

There is an excellent survey of attempts to classify Problem-Types on Manuela Viezzer's webpage at *www.cs.bham.ac.uk/~mxv/publications/onto_engineering.* One such attempt was made in a rule-based theory of thinking called SOAR. There, obstacles were called 'impasses' and were classified into just four types: (1) no rules apply to the situation, (2) several rules match, but there is no higher-level rules to choose among them, (3) there are several such rules but they conflict, and, (4) all such rules have met with failure. For more about Soar, See http://tip.psychology.org/newell.html

136

Reference to Push Singh's paper on "Reflective Critics."

137

Principles of Psychology. Chap. 25 p452

138

See Ortony, A., Clore, G.L., Collins, A., *The Cognitive Structure of the Emotions*, New York, Cambridge University Press (1988).

139

These quotations are from Poincare 1908. *The Foundations of Science,* 1982, ISBN: 0819123188.

140

In *comp.ai.philosophy,* Nov 20 1995.

141

Some theorists question the existence of —this sort of unconscious processing. Paul Plsek discusses this issue at length: *"Some experts dismiss the notion that creativity can be described as a sequence of steps in a model. For example, Vinacke (1953) is adamant that creative thinking in the arts does not follow a model [and] Gestalt philosophers like Wertheimer assert that the process of creative thinking ... does not lend itself to the segmentation implied by the steps of a model. But while such views are strongly held, they are in the minority. ... In contrast to the prominent role that some models give to subconscious processes, Perkins (1981) argues that subconscious mental processes are behind all thinking and, therefore, play no extraordinary role in creative thinking."*—Paul E. Plsek in www.directedcreativity.com/pages/WPModels.html Ask him at paulplsek@directedcreativity.com. See also Perkins, DN (1981) *The Mind's Best Work.* Cambridge, MA: Harvard University Press; Vinacke, WE (1953) *The Psychology of Thinking.* New York: McGraw Hill; and Wertheimer, M (1945) *Productive Thinking.* New York: Harper.

142

http://nobelprize.org/medicine/laureates/1973/tinbergen-lecture.html

143

http://nobelprize.org/medicine/laureates/1973/lorenz-lecture.html

144

This could be related to why some brain waves become irregular when our thinking encounters obstacles.

145

This figure includes the names of some current ideas about how such records are represented. One can see descriptions of some of these schemes by searching the Web with keywords like *working memory, short-term memory,* and *global workspace.* The ideas of Bernard Baars (see http://www.imprint.co.uk/online/baars.html) seem especially relevant to me.

146

The construction of long-term memories appears to involve special kinds of sleep, in ways that are not yet understood. It also appears that different kinds of memories (e.g., about autobiographical events, about other kinds of episodes, about what are called 'declarative' facts, and about perceptual and motor events) are each stored in somewhat different ways and in different locations in the brain.

147

Section 19.10 of *The Society of Mind* described a scheme called *"Closing the Ring"* that could help to re-connect some of the parts that were not at first retrieved.

148

This is a version of a scene described in chapter §1.0 of "The Society of Mind."

149

"Et le second est que, bien qu'elles fissent plusieurs choses aussi bien, ou peut-être mieux qu'aucun de nous, elles manqueraient infailliblement en quelques autres, par lesquelles on découvrirait qu'elles n'agiraient pas par connaissance, mais seulement par la disposition de leurs organes. Car, au lieu

que la raison est un instrument universel, qui peut servir en toutes sortes de rencontres, ces organes ont besoin de quelque particulière disposition pour chaque action particulière; d'où vient qu'il est moralement impossible qu'il y en ait assez de divers en une machine pour la faire agir en toutes les occurrences de la vie, de même façon que notre raison nous fait agir." Rene Descartes, in Discours de la méthode (1637)

150

Chapter III of *The Descent of Man*

151

Turing described these "universal" machines before any modern computers were built. For more details about how these work, see http://mathworld.wolfram.com/UniversalTuringMachine.html.

152

This switching usually happens so quickly that we don't notice it; this is a typical instance of the Immanence Illusion [See §4-3.1.]

153

There is a detailed theory of how this works in §24.6 Direction-Nemes of *The Society of Mind.*

154

It was recently discovered only recently that people often do not perceive some very large changes in a scene. See [give reference] for astonishing demonstrations of this.

155

From http:/*web.media.mit.edu*~minsky/papers/Frames/frames.html

156

See Chapter 3 of William Calvin, *How Brains Think,* Basic Books, 1966.

157

For more details about the relations among different nearby things, see chapter 24 of SoM, which also tries to explain why the shapes of things don't seem to change when we look at them from different directions—as well as why things don't seem to change their locations when you move your eyes.

158

I wonder if Hume had some such idea when he said: *"All belief of matter of fact or real existence is derived merely from some object, present to the memory or senses, and a customary conjunction between that and some other object. ... [This results from] a species of natural instincts, which no reasoning or process of the thought and understanding is able either to produce or to prevent."*—David Hume, *An Enquiry Concerning Human Understanding,* 1748.

159

David Hume, ibid. Part II.

160

Hume was especially concerned with this question of how evidence can lead to conclusions: *"It is only after a long course of uniform experiments in any kind, that we attain a firm reliance and security with regard to a particular event. Now where is that process of reasoning which, from one instance, draws a conclusion, so different from that which it infers from a hundred instances that are nowise different from that single one? I cannot find, I cannot imagine any such reasoning."*

161

Note that this is a difference-engine 'in reverse,' because it changes the internal description, instead of changing the actual situation. See *"Verbal expression"* in SoM §22.10).

162

Robert Stickgold et al., *Cognitive Neuroscience* (Vol. 12, No. 2) in March 2000, also *Nature Neuroscience* (Vol. 3, No. 12) in December 2000.

163

Another answer might be that some information is stored 'dynamically'—for example by being repeatedly echoed between two or more different clusters of brain cells.

164

I described a similar system for verbal communication in §22.10 of SoM.

165

The K-line idea was first developed in [Ref: Plain Talk] and [Ref: K-lines]. Chapter 8 of SoM describes more ideas about what might happen when K-lines conflict.

166

Perhaps she used that facial expression to help her maintain her concentration. If this became part of her subsequent skill, it could later be hard to eliminate.

167

In the field of Artificial Intelligence, the importance of credit-assignment was first recognized in Arthur Samuel's research on machine learning. [Ref.]

168

A. Newell, *"The chess machine,"* in Proc. Western Joint Computer Conf. March 1955.

169

People often describe such moments as the times at which they make their decisions—and then regard these as 'acts of free will." However, one might instead regard those moments as merely the times at which one's 'deciding' comes to a stop.

170

Presumably, these capacities also may vary among different parts of the

same mind.

171

Some of this section is adapted from §7.10 of SoM.

172

Harold G. McCurdy, *The Childhood Pattern of Genius.* Horizon Magazine, May 1960, pp. 32-38. McCurdy concluded that mass education in public schools has "the effect of reducing all three of the above factors to minimum values."

173

Where do we get those default assumptions? Answer: we usually make a new frame by making changes in some older one, and values that were not changed at that time will be inherited from those older ones.

174

I should add that a frame can include some additional slots that activate other processes or sets of resources. This way, a frame could transiently activate ways to think—so that one almost instantly knows how to deal with some familiar object or situation.

175

I should add that numerical representations have many useful applications. However, even when those numbers have some practical use, one can only alter them by increasing or decreasing them, but cannot add other nuances. It is much the same 'logical' systems; each 'proposition' must be true or false, so the system still uses something like numbers, except that their values can only be 0 or 1. Also, see see SOM, section 5.3.

176

§§20.1 of SoM argues that even our thoughts can be ambiguous.

177

Also, several such functions could be superimposed in the very same spatial

regions, by using by genetically distinct lines of cells that interact mainly among themselves.

178

Later Kant claims that our minds must start with some rules like *"Every change must have a cause."* Today, one might interpret this as suggesting that we're born with frames that are equipped with slots that we tend to link to the causes of changes. In the simplest case, of course, that need could be satisfied by a link to whatever preceded the change that occurred; in later years we could learn to refine those links.

179

There is more discussion of this in web.media.mit.edu/~minsky/papers/SymbolicVs.Connectionist.html.

180

Daniel Dennett, in The Cambridge Dictionary of Philosophy. A similar premise was prevalent before the dawn of modern genetics: that every sperm already contained a perfectly formed little personage. However, *in Brainstorms*, 1978, Daniel Dennett goes on to point out that, *"Homunculi are bogeymen only if they duplicate entire the talents they are rung in to explain. If one can get a team or committee of relatively ignorant, narrow-minded, blind homunculi to produce the intelligent behavior of the whole, this is progress."*

181

Here we use *"Model of X"* as in §4-3 to mean any structure or process that one can use to answer some questions about X.

182

We'll review some traditional 'unified theories of psychology in §§Models Of Mind.

183

"The Trouble With Psychological Darwinism" at http://www.lrb.co.uk/v20/n02/fodo01_.html

184

See http://www.theabsolute.net/minefield/witforwisdom.html

185

Greg Egan *Diaspora, Millennium,* 1998, ISBN 0-75280-925-3

186

Daniel Dennett in 1988, *Times Literary Supplement,* 16-22 ix.

187

Alfred Korzybski, *Science and Sanity,* (1933).

188

Ref: Adapted from an essay by Bertram Forer in http://skepdic.com/coldread.html

189

See Shawn Carlson's double-blind study of this in Nature, Dec. 5, 1985.

190

Thus Einstein's $E=Mc^2$ was only a small variation of Newton's $E=Mv^2$, but led to major changes in the ways that we then could understand the world.

191

There is a longer list in SoM §Self-Control.

192

See *http://www.nobel.se/medicine/laureates/1981/sperry-lecture.html*

193

Paraphrased from §11.8 of SoM.

194

Nevertheless, many feelings seem to come with varied degrees of both 'positive' and 'negative' intensities, and this has led many psychologists to maintain that this dimension of intensity is what distinguishes emotions from other types of mental states. See SoM, 28.2 and 28.3, and Ortony, A., Clore, G.L., Collins, A., *The Cognitive Structure of the Emotions*, New York, Cambridge University Press (1988) ISBN 0521386640

195

In §13.1 of SoM, we discussed how we frequently make similar such distinctions among our various sorts of goals and subgoals.

196

See http://www.intriguing.com/mp/lifeofbrian/

197

The American philosopher Josiah Royce (1855-1916), in *The Philosophy of Loyalty, 1908, Vanderbilt Univ. Press, 1995,* **ISBN** 0-8265-1267-4

198

"The Trouble With Psychological Darwinism" at http://www.lrb.co.uk/v20/n02/fodo01_.html

199

Roger Schank has suggested that we mainly remember things that 'make sense' because our memory systems have ways to store representations that have the form of coherent stories. See Roger Schank, *"Tell me a Story,"*Northwestern University Press, 1995, **ISBN** 0-8101-1313-9

200

Encarta World English Dictionary © 1999 Microsoft Corporation.

201

Generally, if a certain reaction leads to a reward, then that response will

become more likely later. Psychologists attribute this" Law of Effect' to the American psychologist Edward Thorndike (1874-1949).

202

Piaget, Jean. (1923). *The Language and Thought of the Child,* Rutledge (2001) ISBN 0415267501

203

Many readers interpreted my earlier book, *The Society of Mind,* as proposing this kind of community view. However, that was not my intention, and this section may help to correct that impression.

204

Another difference between human employees and parts of a brain is that each member of a company has personal conflicts of interest. For example, each employee is hired to increase the company's profit, but this goes against each employee's ambition to earn more salary—because each paycheck reduces the firm's total earnings. It's the same when an army orders its fighters to try to kill those on the opposite side; each soldier still wishes to stay alive.

205

In a publicly held corporation, the officers are not autonomous (at least in principle, if not in actual fact) but actually are employees appointed by directors who are elected by stockholders.

206

Richard Dawkins, *The Meme Machine,* Oxford University Press March 1999, ISBN 019-850365-2. Also see Susan Blakemore, *The Meme-machine...* Oxford University Press, 2001, ISBN **ISBN** 0-19-850365-2

207

In *A Light Exists in Spring, at http://www.repeatafterus.com/title.php? i=6707*

208

See http://eksl-www.cs.umass.edu/~atkin/791T/chalmers.html and http://consc.net/papers/puzzle.html. For more details, see *Journal of Consciousness Studies* 2(3): 200-19, 1995 or http://consc.net/papers/facing.html.

209

Ryle, Gilbert. *The Concept of Mind,* The University of Chicago Press, 1949

210

In fact, a single spot of red may not be sensed as being red; in general the colors we see depend, to a large extent, on which other colors are in its neighborhood. Also, some readers might be surpised to hear that the visual system in a human brain includes dozens of different processing centers.

211

Those touches seem different in ones low-level descriptions, because each relates to a different *'hand'*. But those touches seem more similar at levels that can refer to *'your hand'*.

212

This discussion of verbal metaphor is paraphrased from §29.8 of SoM.