# DRAFT
# Future Models for Mind-Machines

Marvin L Minsky
Massachusetts Institute of Technology
http://www.media.mit.edu/people/minsky/

**Abstract**

***"Seek not to follow in the footsteps of men of old; seek what they sought."* – Matsuo Basho**

In recent years some political leaders of several countries have expressed concern that in future years their countries will not have enough young people to support the large proportion of old ones. So, incredibly, they propose to take steps to increase their reproduction rates. Instead we embark on programs to develop intelligent robots that could increase productivity in the fields where shortages may appear. Then, each working human could easily support many more other ones – with less damage to our environment. However, there has not been much progress in recent years toward making machines that are able to do most mundane jobs that people do. I think this is because most AI researchers have not used adequate large-scale models for designing systems that could have enough "common sense" or "resourcefulness."

## 1 Introduction

Humanity has always faced new technological frontiers – but rarely did it appreciate those wonderful opportunities. However, the past three centuries has been different, I think – and over the past fifty years, we've seen the most immense progress in history. For example Physics, Astronomy, and Cosmology have progressed perhaps more in the past half-century than they did since Galileo's time. Biology has moved even more quickly; the field of molecular Biology was virtually born just fifty years ago. Today, I think, we are entering a similar phase of Psychology.

To build reliable, humanlike robots, we'll need ways to make them understand the problems that we want them to solve. One way to do this would to enable them to think in ways like ours. However, we don't yet know how to do this – because we still know too little about our own minds. Our minds are working all the time, but we rarely think about what minds are. What are minds made of and how they work? How do minds build new ideas? Why could our scientists discover so much about atoms and oceans and planets and stars – yet so little about what our feelings are? Our minds are working all the time – yet we know almost nothing about them. We rarely discuss these subjects in schools, or think about them in our daily lives. It is almost as though we've imposed a taboo against trying to think about such things.

How does Imagination work? How do minds learn from experience? How do we recognize what we see? How do we choose which words to say? How do we understand what they mean? How does commonsense reasoning work? Each of these common abilities is based on huge networks of processes. So, to answer those questions, we'll need to accumulate more good ideas about what *are* those networks, how they evolved, and how their resources have managed to merge – to form the constructions we call our minds. In this essay I will start by reviewing some ideas about minds – each of which has just enough parts to answer certain kinds of questions. Then I will suggest how these simple models can be expanded and combined to make better theories about our psychology. (Each brief section below will be further discussed in my forthcoming book, *The Emotion Machine*.)

## 2 One-Part Models of Mind

The most popular concept of a human mind envisions each person as having a 'Self" – which embodies all those features and traits that distinguish you from everyone else. But when we ask what Selves actually do, we're likely to hear this vacuous view:

> *Your Self views the world by using your senses, and chooses all your desires and goals. Then it solves all your problems for you, by exploiting your 'intelligence'. It formulates plans for what next you should do – and then makes the pertinent muscles contract so that your body performs your acts.*

Isn't this a strange idea? It says that you make no decisions yourself but just delegate them to something else – to that mythical person you call 'your Self'? Clearly this 'theory' can't answer our questions – so why would our minds concoct such a fiction?

> *Therapist: "That simplistic legend makes life seem more pleasant. It keeps us from seeing how much of our soul are controlled by unconscious, detestable*

*goals."*

*Pragmatist: "It also helps to make us efficient! More complex ideas might just slow us down. It would take too long for our hardworking minds to understood everything all the time."*

The trouble with that "Self" idea is that it does not explain what's inside a mind. It's a theory that doesn't have enough parts we can use to build explanations. If you ask about how your mind makes decisions, the Central-Self model just avoids that question, by ascribing all your abilities to another mind inside your mind. (Before the dawn of modern genetics, a similar theory was prevalent: it proclaimed that every sperm already contained a perfectly formed little personage.) The notion of a Central Self can't help us to understand ourselves.

Many other popular theories try to derive all the virtues of minds from one single source or principle:

*Survival instinct: All our goals stem from the instinct to survive.*

*Pleasure Principle: All our drives are based on seeking pleasure*

*Aversion Principle: We're driven by needs to escape from pain.*

*Conflict Resolution: All our actions are directed at resolving conflicts.*

*Urge to control: Our resources evolved to control our environment.*

*Reinforcement and Association: The mind grows by accumulating various kinds of correlations.*

Each of these 'unified theories of mind' has virtues and deficiencies. For example, the Survival-Instinct hypothesis helps top describe a wide range of behaviors – but it's based on a wonderfully wrong idea. Over the course of our evolution, our brains assembled a great host of systems – each of which served in a separate way to protect us from certain kinds of harm. The result of the process was that a brain is a 'suitcase' of systems with similar functions; however, those systems have no common structure – so to understand how those systems work, we'd have to examine them one by one. That 'survival instinct' is just an illusion. When you look at mind as a single thing – instead of a grand architectural scheme – you'll see little more than a featureless blur, instead of the marvelous structure you are.

# 3 Two-Part "Dumb-Bell" Models of Mind

Many popular mental models are based on "dumb-bell" distinctions that try to divide the entire mind into just two complementary portions, such as *Left-Brain vs. Right-Brain, Rational vs. Intuitive; Intellectual vs. Emotional, or Conscious vs. Unconscious*. These can be better than Single-Self models. However, they too often support old superstitions that make it hard to develop more useful ideas. For example, when neurologists discovered some differences between the brain's two hemispheres, this revived many views of our minds which were, in our more ancient times, expressed in terms of opposites like Devils vs. Angels, Sinners vs. Saints, and *Yin*s vs. *Yang*s. So this pseudoscientific scheme revived nearly every dead idea of how to see the mental world as a battleground for two equal and opposite powers.

Why are dumbbell theories so popular? I suspect that this is because – just like those old myths – they provide just enough parts to tell stories of conflicts. Instead of believing such story-like myths, we should try to make theories of why they enchant us.

# 4 Three-part Models of Mind

Three-part theories, although still too simplistic, are rich enough to suggest better ideas. Here are a few of my favorite such frameworks:

Paul MacLean's "Triune brain" hypothesis [*The Triune Brain in Evolution*] tries to explain how minds behave in terms of machinery that evolved in three stages – namely, when our ancestors became Reptiles, then Mammal, and finally, Primates. He identifies those hypothetical 'layers' with stages of our evolutionary history – as well as with different aspects of thinking. However the evolution of our 'lower' brain systems did not suddenly cease when those 'higher' ones came. They all continued to co-evolve, so that each of our behavioral functions is based on components from every stage.

Eric Berne's "Transactional Analysis" hypothesis is based on the idea that every person evolves sub-personalities based on models of the child, adult, and parent. [Eric Berne, *Transactional Analysis in Psychotherapy*] This is quite different from MacLean's scheme, and more suitable for describing the development of social behaviors.

Sigmund Freud's "Psychoanalysis" theory was based on a psychological triad of interactions between a "Id" or collection of Instinctive urges, a "Superego" that embodies our high-level socialized goals and prohibitions, and an "Ego" that resolves or suppresses the conflicts between them. I especially like Freud's 'sandwich-like' architecture, first because it is non-hierarchical, and second because it emphasizes 'negative knowledge' – that is, knowing which things one should not do. Competence requires both positive and negative knowledge – and I suspect that as much as half of our commonsense knowledge may have of this negative character. [See Marvin Minsky, "Negative Expertise"]

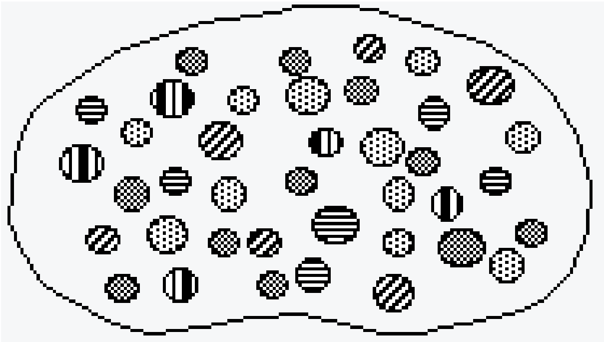

**Figure 1**

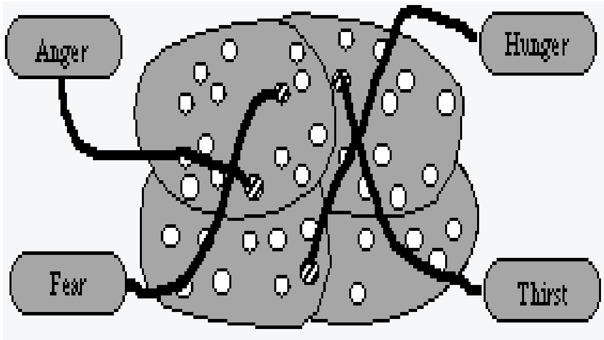


**Figure 2**

## 5 Viewing the Mind as a "Cloud of Resources"

The human brain has hundreds of parts that have different functions – so any comprehensive model of mind must include descriptions of all those resources. By "resources" I mean to include both bodies of knowledge and program-like processes – such as perceptual schemes for making descriptions, for forming goals and for making decisions, or methods for solving difficult problems. Especially, the brain needs resources to assess what other resources do – e.g., to decide which ones are making good progress or wasting our time, or to recognize conflicts and try to resolve them. This suggests that we think of the brain as a cloud of varied resources, where each can use others in certain ways. [Figure 1]

*Holistic Philosopher: That whole idea seems wrong to me. By dividing the mind into smaller parts, aren't you likely to miss the whole point? Unless you look at a thing as a whole, you'll miss its most vital aspects. Surely you need a more holistic view.*

Every representation we use is bound to miss some important aspects, for which we must switch to another view or a different type of representation. So to understand anything well, we'll usually need to use several such views, and some ways to interconnect them. Certainly, this must include some "high level" views that try to describe the entire thing. However, 'holistic thinkers' don't always recognize that vague summaries have their limits, too. Like cartoons, they give us illusions of "seeing the whole thing at once." However, these tend to be oversimplified views that cannot explain anything in detail – just as maps display only a few striking features, while suppressing details of the actual regions.

This idea of a mind as a cloud of resources might seem too vague to have much use, but it helps us to escape from those dumb two-part models. For consider the following type of phenomenon: 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 just a few moments the air fills with screams. The Single-Self model has no way to explain what could possibly cause such changes – but this is easier to explain if we assume that an animal's brain contains several almost-separate sets of resources – where each set evolved to serve some vital need like procreation, nutrition, or defense. This model, developed by Nikolaas Tinbergen and Konrad Lorenz, is described in Tinbergen 's book "The Study of Instinct" (Oxford University Press, 1951). It does not explain much about human thought but has turned out to be surprisingly good at accounting for much of what animals do.

One form of a system with such a description might resemble a human community, where different people do different jobs – as in Howard Gardner's theories about Multiple Intelligences, which lead to good models for representing a person's largest scale behavior. [See, for example, Howard Gardner, *Frames of Mind: The Theory of Multiple Intelligences*.] However, each member of a human family, village, or corporation is already a competent and autonomous person – whereas inside a single person's brain, each resource is far more specialized; it can do only a certain few things, and depends on the rest for everything else. So, when we envision an individual human mind, it may be better to think of a large network of smaller machines. Of course the resource-cloud view is not quite what one would call 'a theory' – because while it says that the system has parts, it does not specify what those parts are. It says they're connected, but doesn't say how. It suggests no particular architecture. However, the very vagueness of the Resource-Cloud idea is what makes it a powerful tool for thought, just because it reminds us of those deficiencies.

In particular, it suggests that the brain must contain enough "managers" to monitor, supervise, appraise, and control the activities in particular sets of other resources. A typical resource is connected to several others and can use those connections in various ways, e.g., to exchange some information with them, to exploit them for various purposes. In particular, some resources will be especially equipped to turn some other resources on or off. Thus, from every moment to the next, only certain resources will be active – and these will determine what your mind does at any particular moment of time. This suggests a theory of emotions in which each emotion or 'disposition' results from some more or less persistent arrangement in which certain resources are highly active, while others are more

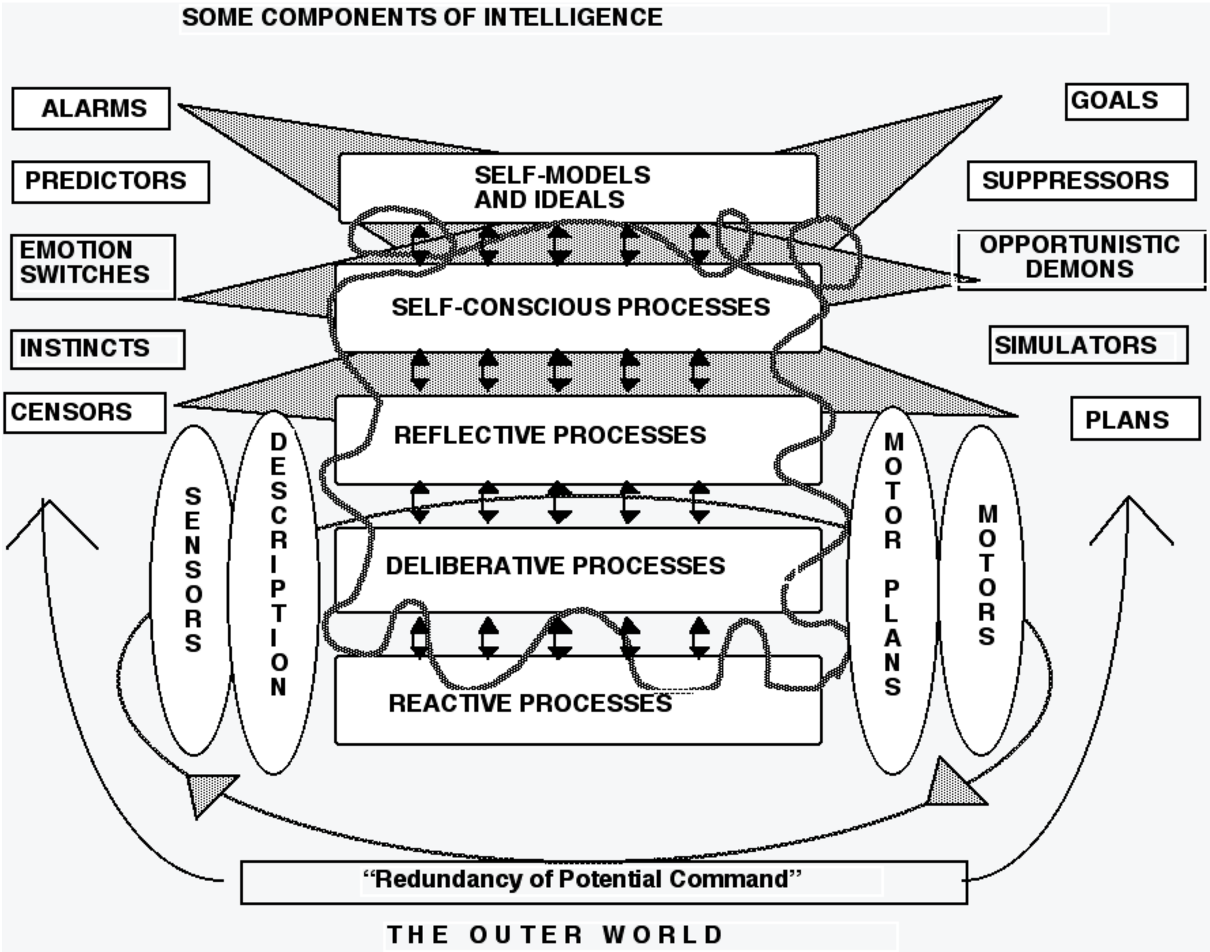


**Figure 3**

quiescent:

*An emotional state is what happens when we 'turn on' a certain large set of resources.* [Figure 2]

## 6 A Large-Scale Model of Conscious Thought

How does a brain employ its resources? One way to start would be to assume, as has been suggested by Aaron Sloman, that our resources are arranged in three or more levels [Figure 3]:
– A "reactive" collection of resources "A" that includes systems for memory, perception, and other procedures, etc.
– A "deliberative" collection of resources "B" that observe and react to the activities in A.
– A "self -reflective collection of resources that observe and react to what happens in "B," etc.

No such a complex system could work without more machinery to control it. To see what that management might involve, let's look at one fragment of human behavior.

*Joan is part way across the street on the way to present her finished report, and she's thinking about what to say at the meeting. She hears a sound and turns her head to see a quickly oncoming car. Uncertain whether to cross or to retreat, but uneasy about arriving late, she elects to sprint across the road. Later she reflects about her rather reckless decision. "I could have been killed if I'd missed my step – and then what would my friends have thought of me?"*

Every minute of every day, we experience streams of events like these. To some of them, we react without thinking. To others we act more deliberately. Let's try to imagine what goes on in Joan's mind as she makes her way to that meeting.

**Reactive Awareness:** *She hears a sound and turns her head in that direction.*

When Joan turned her head to look around, was she conscious of that sound, or was that a 'mindless' reaction? Was she aware of which muscles she used to make herself walk across that road? Not likely, because most of us don't even know which muscles we own. Other resources inside Joan's brain must be more involved with such affairs – but because no pathways communicate this, Joan is not 'aware' of this. What is awareness, anyway? What determines its focus and range? What machinery does it use in the brain? How many things can you do at one time – and how many can you be aware of? Presumably, that will depend on the extent to which they each use different resources for them. But when Joan perceives that approaching car, this quickly takes the center stage and takes hold of her full attention.

**Deliberative thinking:** *She is thinking about what to say at the meeting.*

To do this she must first consider how several alternatives might be received – and then compare those imagined reactions. This may require so many resources that she has to do this sequentially.

**Reflective reasoning:** *Joan reflects about what she has done, and concludes that she made a poor decision.*

To what extent was she aware of what determined her risky decision? Reflection involves thinking about what one's brain's has recently done. That kind of reflection requires resources to examine the records that other resources have been keeping.

**Internal "Meta-Management":** *but uneasy about arriving late*

Another family of resources is monitoring Joan's temporal progress, and decides that whatever the merits of what she is thinking, she cannot afford to delay her decision.

**Self-conscious Reflection:** *"What would my friends have thought of me?"*

Joan thinks about how her friends might change their mental representations of her. Reflections like this have as their subject, a person's private self-representations – the models or self-images that we all construct to describe ourselves.

So the architecture of our minds must include at least these five kinds of layers. This idea is further developed in my forthcoming book *The Emotion Machine*. Of course, a real brain is far more complex, and each of those layers and arrows eventually must be replaced by hundreds of smaller components, interconnected by thousands of pathways. (This scheme is partly inspired by the research of Aaron Sloman.)

# 7 Psychology Needs a Network of Large Scale Models

To understand the human mind, we'll need to use several kinds of models. Some will need only a few parts – enough to answer just certain questions – but others will have to be much more complex, to explain such 'higher mental functions' as reasoning, imagination, decision-making, and consciousness. And, since no one such vision can explain everything that we want to explain, we'll have to keep switching between different models.

*Critic: That sounds very disorderly. Why can't you simply combine them all, like the physicists try to do, into a single one that combines the virtues of all those theories?*

That would result in such a mess that no one could hold it in mind all at once. We have to be able to use different views to highlight different aspects of things, and that's why we still tend to speak about Physics, Chemistry, and Biology – as though these were more or less separate subjects. Some of the contents of each of those fields can be deduced 'in principle' from more basic physical principles. The trouble is that we can't do this "in practice" because no one can actually solve those equations. (And in Psychology, we can't expect to have any such set of equations.)

The 'large-scale models' that we've described are not 'hypotheses' to prove false or true. Instead, they are more like 'points of view' – particular ways to think about things, or to focus attention on various problems. So it's not a question of which one is 'right', but where and when to use each view. Each is a rough architectural plan that will help us to understand certain things. However, because each of them has limitations, we'll have to keep changing our points of view, by shifting between different Large-Scale Model. Our own human brains are too complex for us to envision all at once – so we'll have to keep changing our representations. This shifting around might at first seem disturbing, but later we'll see that it's worthwhile – because it will also enable us to describe the process that actually happens *inside* our minds!

> *Using multiple models is not just a way to state theories about psychology. It is part of psychology itself – because we can only understand complex things by switching between different representations. This is the basis of our most powerful way to think: to keep interweaving different views so fluently that we never suspect that we're doing it.*

No system as complex as a human mind can be well described by a few simple rules – because each rule would have many exceptions. This is because each part of such a system is likely to reflect the particular ways that it once worked in the environment in which it evolved (both out in the world and inside the brain). Then whenever some subsystem fails to work, those brains will tend to evolve a 'patch' – an 'ad hoc' way to help it to work. The result is the accumulation of multiple layers of patches, over hundreds of megayears of evolution.

What does it mean when you say to yourself, "That was a stupid thing to do," or "I didn't expect to succeed at that!" You're always praising or blaming yourself, and holding yourself responsible. But whenever you change your emotional states, you're using some different processes and memories – so you are no longer the very same 'you'. What gives us the sense that we remain the same while shuttling among those states? Partly this must be because we use the terms for describing ourselves. Terms like 'me', 'myself' and 'I' help us to envision ourselves as like the 'eye' of a cyclone that stays in one place while everything circles around it. In *The Emotion Machine* I'll argue that the mind has no single well-defined thing that remains the same while controlling the rest. Instead we each have a rich collection of personal, large-scale models of ourselves.

Our 'commonsense' ideas about ourselves have so many bad misconceptions. We all have grown up with certain traditions that tacitly assume, for example, that we each 'hold' a single set of beliefs. Thus, when someone asks what you "really" believe – or what your 'true' intentions are – or what you 'really' meant to say – those phrases make sense in the Single-Self realm. But a realistic view of your mind would show how it uses at various times, different arrangements of its resources – each of which can make you exhibit different opinions, ideas, and convictions. And despite what each of us likes to think, no particular one of those cliques deserves to be called "what I truly believe".

# 8 Advice to Students

How should student select a career in these future burgeoning technical fields? One approach is to ask what is the most popular field now. Another approach is the opposite: to choose an underpopulated area. Now, the popular fields offer great current opportunities. (For example, in genetics, each of our hundred thousand genes may take a few lifetimes to understand – for evolution has used all the tricks that the physical world permits.)

However, a young, ambitious student who wishes to make a great and fundamental contribution should consider the idea of deliberately avoiding the most popular fields! For, consider the arithmetic. Imagine that in the next ten years there will be ten major discoveries in a certain field where already ten thousand researchers are working. (This is the case at present in such areas, for example, as Neural Networks, Genetic Programming, Simple Mechanical Robots, Statistical Linguistics, and Statistical Information Retrieval.) Then in each decade, each of those researchers will have perhaps one chance in 1,000 to make a major discovery. Contrast this with the situation in an equally important field that currently employs only the order of a dozen good researchers – as in the areas of *Representing Commonsense Knowledge* or *Large-Scale Cognitive Architectures*. Then you'll have a thousand times better chance to make an important discovery! Many students have complained to me that it's easier to get a job in a currently popular field. However, if one looks for less faddish alternatives, one may find that the competition is accordingly less.

# References


Eric Berne, *Transactional Analysis in Psychotherapy*, Grove Press, New York, 1961

Howard Gardner, *Frames of Mind: The Theory of Multiple Intelligences*, Basic Books, 1993, ISBN: 0465025102.

Paul MacLean *The Triune Brain in Evolution*, Plenum Pub Corp., ISBN: 0306431688, New York, 1990

Marvin Minsky, Negative Expertise, *Int. J. Expert Systems*, 7,1, pp. 13–19, 1994 or www.media.mit.edu/people/minsky/papers/NegExp.mss.txt

Nikolaas Tinbergen The Study of Instinct, Oxford University Press, 1951