MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

A.I. Memo No. 516 June, 1979

K-LINES: A THEORY OF MEMORY

Marvin Minsky

ABSTRACT. Most theories of memory suggest that when we learn or memorize something, some "representation" of that something is constructed, stored and later retrieved. This raises questions like:

How is information represented?
How is it stored?
How is it retrieved?
Then, how is it used?

This paper tries to deal with all these at once. When you get an idea and want to "remember" it, you create a "K-line" for it. When later activated, the K-line induces a partial mental state resembling the one that created it. A "partial mental state" is a subset of those mental agencies operating at one moment. This view leads to many ideas about the development, structure and physiology of Memory, and about how to implement frame-like representations in a distributed processor.

Acknowledgements: This report describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the laboratory's artificial intelligence research is provided in part by the Office of Naval Research under Office of Naval Research contract N000014-79-C-0260.

K-LINES: A THEORY OF MEMORY

Marvin Minsky

M. I. T.

Most theories of memory suggest that when you learn or memorize something, a *representation* of that something is constructed, stored and later retrieved. This leads to questions like:

How is the information represented?
How is it stored?
How is it retrieved?
How is it used?

New situations are never exactly the same as old. So if the information in an old "memory" is to be useful, it must somehow be generalized or abstracted. This leads us also to ask:

How are the abstractions made?
When -- before or after storage?
How are they later instantiated?

We try to deal with all these at once, by developing the thesis: ***the function of a memory is to re-create a state of mind.*** Then each memory must embody information that can later serve to re-assemble the mechanisms that were active when it was formed -- to recreate a "memorable" brain event. (See *Note 1.*) Our scheme is basically simple:

When you "get an idea", or "solve a problem", or have a "memorable experience", you create something we shall call a *K-line* for it.

This K-line gets connected to those "mental agencies" that were recently active -- i.e., which were involved in the memorable mental event.

When the K-line is later "activated", it reactivates those mental agencies, creating a "partial mental state" resembling the original.

To make this concrete, we must explain:

What are "mental agencies"?
How do K-lines interact with them?
What is a "partial mental state"?
How does this relate to "meaning"?

DISPOSITIONS vs. PROPOSITIONS

In this modern era of "information processing psychology" it may seem quaint to talk of mental states. The concept of "representation of knowledge" seems lucid enough when talking about memories of sentences, numbers, or even faces, for one can imagine how to formulate these in terms of propositions, frames, or semantic networks. But it is much harder to do this for feelings, insights and understandings, with all the attitudes, dispositions, and "ways of seeing things" that go with them. (See ***Note 2.***) Traditionally, such issues are put aside, with the excuse that we should understand simpler things first. But what if feelings and viewpoints *are* the simpler things -- the elements of which the others are composed? Then, I assert, we should deal with dispositions directly, using a "structural" approach that portrays memory as re-setting the states of parts of the nervous system.

We will view a memory as something that predisposes the mind to deal with a new situation in an old, remembered way. This is why I put "dispositions" ahead of "propositions". First we propose some "disposition representing" structures. Then we try to show that these can evolve into the more familiar kinds of cognitive constructs we know as adults. One should not assume that human memory has the same uniform, invariant character throughout development, nor attribute to infants abilities that develop only later. Our first model might serve for an infantile, dispositional memory. Later we try to see how it might evolve into a more adult system.

MENTAL STATES and the SOCIETY of MIND

One could say but little about "mental states" if one imagined the Mind to be a single, unitary thing. But if we envision a mind (or brain) as composed of many partially autonomous "agents" -- a "Society" of smaller minds -- then we can interpret "mental state" and "partial mental state" in terms of ***subsets of the states of the parts of the mind.*** To develop this idea, we will imagine first that this Mental Society works much like any human administrative organization.

On the largest scale are gross "Divisions" that specialize in such areas as sensory processing, language, long-range planning, and so forth.

Within each Division are multitudes of subspecialists -- call them "agents" -- that embody smaller elements of an individual's knowledge, skills, and methods.

No single one of these little agents knows very much by itself, but each recognizes certain configurations of a few associates and responds by altering its state.

In the simplest version of this, each agent has just two states, active and quiet. A *total mental state* is just a selection of which agents are active. A *partial mental state* is a partial such specification: it ***fixes the states of just some of the agents.***

It is easiest to think of partial states that constrain only agents within a single Division. Thus one could think of a partial state that specifies some "visual imagery" without saying anything about agents outside the visual division. In this paper our main concern will be with even "smaller" partial states, that constrain only *some* agents in one Division.

Note that the concept of partial state allows us to speak of entertaining ***several partial states at once*** -- to the extent they are compatible -- that is, they do not assign different states to the same individual agents. Even if they conflict, the concept may still have some meaning, if the conflicts can be settled within the Society. This could be important, because local mechanisms for resolving differences could be the antecedents of what we know later as ***reasoning*** -- useful ways to combine different fragments of knowledge.

In the next few sections we postulate that certain units -- the K-nodes and K-lines -- are the elements of memory. When activated, each such unit imposes a specific partial state upon the Society. Such effects are not always easy to describe, for we are most fluent at talking of arrangements of sights and sounds -- or motor patterns; these are "concrete" matters. Much more elusive seem our recollections of attitudes, points of view, and feelings. This does not mean that concrete recollection is fundamentally simpler! It may only reflect the enormous competence of the logical and linguistic parts of the adult mental society to communicate about concrete matters. That illusion of simplicity can fool us, as theorists, into trying first to solve the hardest problems. (See *Note* 3.)

The novice remembers "being at" a concert, and something of how it affected him. The amateur remembers more of what it "sounded like". Only the professional remembers much of the music itself, timbres, tones and textures. So, the most concrete recollection may require the most refined expertise. Thus, while our theory might appear to put last things first, I maintain that attitudes do really precede propositions, feelings come before facts. This seems strange only because we cannot remember what we knew in infancy.

MEMORIES and PARTIAL BRAIN STATES

Old answers never perfectly suit new questions, except in the most formal, logical circumstances. To deal with this, theorists have tried various ideas:

Encode memories in "abstract" form.
Search all memory for the "nearest match".
Use prototypes with detachable defaults.
Remember "methods", not answers.

Our theory resembles the latter. We propose remembering not the stimulus itself but part of the state of mind it caused. So we shall translate

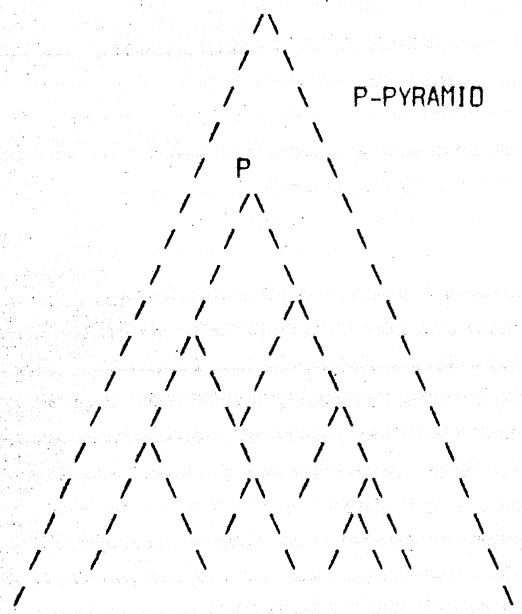
a "method" for solving a kind of problem

into

I once solved a similar problem. If I can get into that old state, I could probably handle this one the same way.

To carry out the translation we must sketch some of the architecture in which our Agents are embedded. (See *Note 4.*) We envision a brain containing a great lattice of "Agents", each connected to just a few others. In our model, we shall suppose that each agent's inputs come either from below or from the side, and its outputs go upwards or sideways. Thus information can move only upwards, on the whole. (See *Note 5.*) This is what one might imagine for the lower levels of a visual system, with simple feature- and texture-detectors at the bottom, then edge- and region-sensing agents, then identifiers of larger structures.

Given these constraints, if one "looks down" from the viewpoint of a given agent P, one will see other agents arranged roughly in a hierarchical Pyramid:



I emphasize that the network as a whole need not be pyramidal; the P-pyramid we speak of is an illusion of an agent's perspective.

CROSS-EXCLUSION and PERSISTENCE

In our concept of the Society of Mind, most agents are grouped in small "cross-exclusion" arrangements. Each sends inhibiting connections to the others in its group, so that it is hard for more than one to be "active" at a time. This kind of sub-structure, familiar in physiology, makes it particularly easy to re-set the state of a system; one need only force to "on" one agent in each cross-exclusion group. Then that agent will inhibit its associates -- reducing ***their*** inhibiting effect on itself.

The result: networks composed of cross-exclusion systems have a kind of built-in "short-term memory." Once such a system is forced into a partial state, even for a moment, then that state will tend to persist -- except for those agents under strong external pressure to change. Accordingly, such a system tends to have internal persistences. To an outside observer, these will appear as "dispositions" -- distinctive styles of behavior. To make a large change in such a disposition, one has to change many of the agents' states. Small changes will only slightly perturb the overall disposition.

Finally, we suppose that agents at the lowest levels tend to change states most frequently, in response to signals ascending from the outside or from other P-nets. In the scheme described below, the states of intermediate level agents will have the most effect on the longer-term dispositions, hence will be most deeply involved with memory; they will play relatively persistent roles in determining how agents below them influence agents above them. In my image of development, the region in which agents are in this sense "intermediate" will presumably move upwards during cognitive growth. (See ***Note 6.***)

For example, a "low-level" agent in the visual system would always compute the same function of retinal stimulation. But at higher levels, different dispositions would induce different "ways of seeing things". For example, the choice of perspective for the "Necker cube" is dictated, not by ascending sensory information but by preference signals coming from other agencies. Thus, one uses non-sensory information to dispose oneself to regard sound as noise or word -- or image as thing or picture. Each P-pyramid may have a repertory of such dispositions, defined by pre-activating different subsets of agents. And a single such system might maintain, at one time, fragments of ***several*** such dispositions -- but only if conflicts are not too serious.

K-LINES and LEVEL BANDS

Now imagine the whole brain to include many such P-structures, interconnected and overlapping according to intricate genetic constraints. Return to the psychological view for a moment, and suppose that one part P of your mind has just experienced a mental event EK which led to achieving some goal -- call it GK. Suppose another part G of your mind declares this to be "memorable". We postulate that two things happen:

> K-NODE ASSIGNMENT: A new agent -- call it the *K-node* AK -- is created and somehow linked with GK.
>
> K-LINE ATTACHMENT: Each K-node has a *K-line* -- a wire having potential connections to every Agent in the P-pyramid. The act of "memorizing" causes this K-line to make an "excitatory" attachment to every currently active P-agent.

The result: when AK is activated at a later time, its K-line will make P "re-enact" that partial state -- by arousing those P-agents that were active when EK was celebrated. So P will virtually "hallucinate" that event. (See ***Note 7.***)

This is the basic idea. But it might seem impractical because every K-line has to come near every P-agent. We now introduce a series of "improvements" that combine to form a powerful mechanism for abstraction and inference. First let us note that it is *not* the goal of Memory to produce a perfect hallucination. (See *Note 8.*) ***One wants to re-enact only enough to "get the idea".*** Indeed, the perfect hallucination would be harmful, for complete resetting of the P-net would erase all the work done in processing the recent data. It might even fool one into seeing the present problem as already solved. The new state must be sensitive to the new situation. ***A memory should induce a state through which we see current reality as an instance of the remembered event.*** The idea below of how to do this is probably the most important idea of this theory.

THE LEVEL-BAND PRINCIPLE

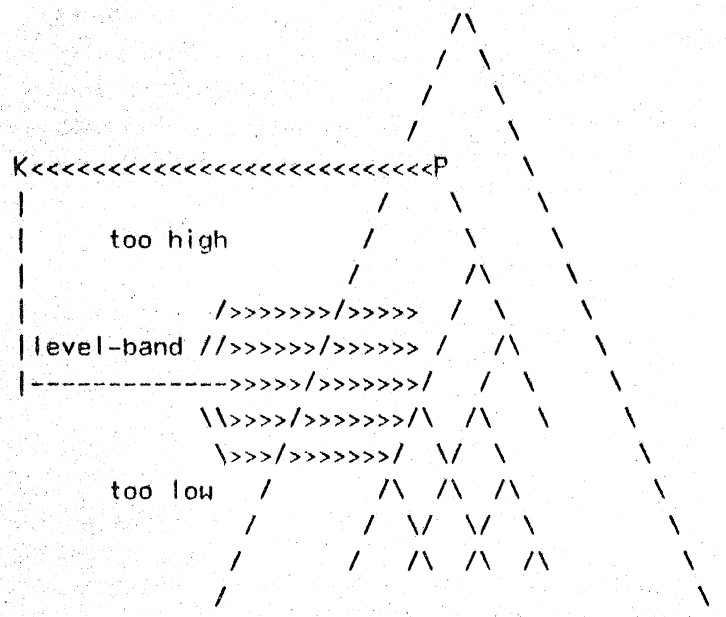
To obtain the desired metaphorical activity, we do not connect AK to all the P-agents that were active during EK, but only those within an intermediate band of levels. To explain this, we must suppose that AK is somehow associated with some agent PK at a certain level of the P-pyramid -- we discuss this "P-->K" association later. Then:

LOWER BAND-LIMIT: the K-line must not reach agents at levels far below PK, for this would impose false perceptions and conceal the real details of the present problem. (See *Note 9.*)

UPPER BAND-LIMIT: Nor should that K-line reach up close to the level of PK itself, for that would make us hallucinate the present problem as already solved, and impose too strongly the details of the old solution.

These two constraints combine to suggest:

LEVEL-BAND PRINCIPLE: A K-line should span only a band of levels somewhere below that of PK, leaving it free to (i) exploit higher level agents appropriate to current goals and (ii) be sensitive to current contigencies as perceived at lower levels.



To summarize: by activating agents only at intermediate levels, the system can ***perform a computation analogous to one from the memorable past, but sensitive to present goals and circumstances.***

CONNECTIONS AMONG K-NODES

A second important principle is this: if K-lines are to contribute to memory, they may as well benefit from memory! When forming a new K-node, we should not ignore the existence of other, previously defined K-nodes. Here is how we embody this idea:

> K-RECURSION PRINCIPLE: Whenever you solve a problem, you exploit memories from the past. So we can assume that when the memorable event EK occurred, this itself was in large part due to activation of some already-existing K-lines. Therefore ***it will suffice to attach the new K-line AK to just the currently-active K-nodes!***

In effect, this says that new memories are composed mainly of ingredients from earlier memories. By making connections to other K-nodes (rather than P-nodes) we need fewer connections and obtain (we shall argue) more meaningful cognitive structures. The level-band arguments apply just as before, hence:

We do not connect AK to all the K-nodes active during EK, but only to those in accord with the Level-Band Principle!

Taken literally, this has a fatal flaw: if K-lines contact only other K-nodes, they can have no ultimate contact with the P-pyramid. That process has to start somewhere! Our proposal: we envision that K-agents lie anatomically near the P-agents of corresponding levels. Then it is easy for K-lines to contact either P- or K- agents. Presumably genetics specifies the proportions and, during development, these preferences tend to shift over from P's to K's.

THE CROSSBAR PROBLEM

Even using the Recursion and Level-Band principles, still each K-node needs potential junctions with many agents. Every brain theory must deal in some way with this "crossbar" problem -- to make the mind capable of a great range of "associations". There may be no general solution. In the cerebral cortex, for example, the (potential) interconnections constitute almost the entire biomass and the actual computer is but a thin layer bordering a three dimensional mass of connecting fibres. But note that the Level-band principle does reduce by one the apparent dimensionality of the problem. (See *Note 10.*)

The crossbar issue is often ignored in traditional programming, because computer memory can be regarded as totally-connected in the sense that register "addresses" can connect any cell to any other in a single step. The problem returns in systems with multiple processors or more active kinds of memory.

In any case, I would not seek to solve the crossbar problem within the context of K-theory nor, for that matter, in any clever coding scheme, or holographic phase-detector -- although any such inventions might help make brains more efficient. Instead, I would seek the answer in the concept of the Society of Mind itself. If the mechanisms of thought can be divided into specialists that intercommunicate only sparsely, then the crossbar problem may need no general solution,: for most pairs of agents will have no real need to talk to one another. Indeed, because they speak (so to speak) different languages, they could not even understand each other. If most communication is local, the crossbar problem scales to more modest proportions.

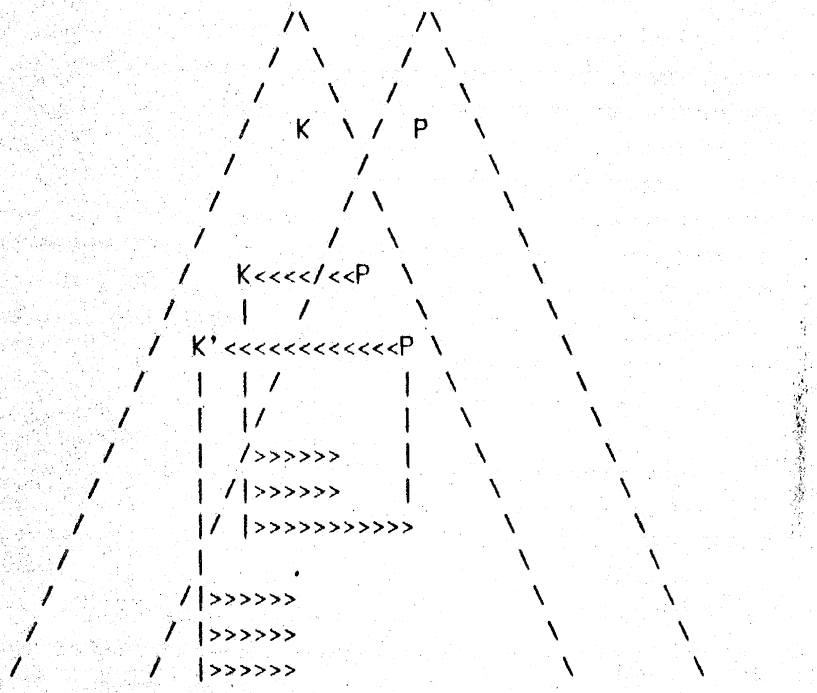
The reader might complain that communication limits within the Mind seem counter-intuitive: cannot one mentally associate *any* two ideas, however different?

Though the final answer is surely "yes", it would seem that unusual connections are usually "indirect" -- be it via words, images, or whatever. The bizarre structures used by mnemonists (and, presumably unknowingly, by each of us) suggest that arbitrary connections require devious pathways.

THE KNOWLEDGE-TREE

It will not have escaped the reader that we have arrived at an elegant geometry:

> ***The K-nodes grow into a structure whose connections mirror those of the P-pyramid, except that information flows goes the other way. P-nodes activate units above them, K-nodes activate units below them. Thus forms a K-pyramid, lying closely against the P-pyramid, each with convenient access to the level bands of the other.***



In terms of this diagram, the local pattern of computation forms a counterclockwise spiral. Globally, over several "cycles", the locus of activity can drift either upwards or down. This "computational architecture" seems very general and versatile.

But the apparent symmetry is deceptive, because I suppressed some hard questions. I gave adequate descriptions of the connections within K, and of those from K to P. I said little about the connections within P, but that is not part of this story, nor is it a problem here; this is discussed in [1]. But of the connections from P back to K, all I said was that "... AK is somehow associated with some agent PK at a certain level of the P-pyramid ...".

The idea was in some way to relate P-events with achievement of Goals represented elsewhere. The rest of the essay discusses various possible such relations but does not settle upon any particular one. In fact, this ends the constructive part of this essay and, from this point, the reader can assume that difficulties in understanding are my fault, not his. I hope only that the foregoing intuitions may stimulate others to construct a more complete theory.

It is tempting to try to find simple ways to restore the symmetry. For example, we talked only of making the K-tree learn to adapt to the P-tree, but the P-tree itself must once have been the learner. Could they take turns training each other? Was the P-tree once the K-tree for another P-system?

Alas, nothing so simple will do. We later argue that non-trivial learning requires at least *three* nets to be involved. For there must be some link from K and P to the rest of the Society, and the P-->K connection seems to want that role.

K-KNOWLEDGE

We started with a naive idea that "memories re-enact past states" -- without attempting to explain what they "mean". But now we come full circle: since the K-system forms a sort of hierarchical web, one can hardly escape asking what its nodes might mean. It seems natural to try to see it as some sort of abstraction lattice in which ***each K-node "represents" some relation among whatever its subordinates "represent".***

K-Knowledge seen as Logical. What kinds of relations? In the simplest case, when partial states do not interact much, a superior simply ***superposes*** the effects of its subordinates. Concurrent activation of two K-lines at ***comparable*** levels will dispose P to respond to *either* meaning. Thus, if P were a sensory system, and if detectors for "chair" *and* "table" are activated, then P will be disposed to react both to a chair *or* to a table. So K-terms at comparable levels tend to combine "disjunctively." If the P-net has multiple outputs at its top, it would tend to produce both outputs.

When the partial states of the subordinates *do* interact, the "logic" of combining K-lines depends upon the "logic" within P. In a version of cross-exclusion that Papert and I favor, the ***activation of two or more competitive P-units usually causes their entire cross-exclusion group simply to "drop out" completely***, defaulting to another group at the next higher level. We see this as a profound heuristic principle: if a single viewpoint produces two conflicting suggestions in a certain situation, it is often better not to seek a compromise, but to seek another, less ambiguous viewpoint! We introduced this idea as a general principle in [2] after Papert formulated it as a theory of how Piaget's Conservation develops in Children.

K-Knowledge seen as Abstract. Initially, we spoke only of creating an entirely new K-node for each memorable event. Now we begin to allow for more gradual and incremental ways to "accumulate" new subordinates to an existing node. A chimpanzee might reach the too-high banana by different means at different times -- first using a box, then a chair, later a table. If all these can be "accumulated" to the *same* node, it can become a powerful "how to reach higher" node. When re-activated, it will ***concurrently*** activate P-agents for boxes, chairs, or tables, so that perception of any of them will be considered relevant to the "reach higher" goal. In this crude way, such an "accumulating" K-node will acquire the effect of a class-abstraction -- an extensional definition of "something to stand on".

But it may do much better than that! If conflicts between details cancel one another out (because of conflict within cross-exclusion subgroups, as mentioned above) then decisions will default to the remaining non-conflicting details! ***This automatically produces a more abstract kind of abstraction -- the extraction of common, non-conflicting properties!*** Combining the concrete "accumulation" of particular instances with the rejection of strongly dissonant properties leads automatically to a rather abstract "unification". (See ***Note 11.***)

K-knowledge as Procedural. When K-lines interact at different vertical levels, the superposition of several partial states will produce various sorts of logical and "illogical consequences" of them. We already know they can produce simple disjuncts as well as "exclusive-ors" -- enough to make a universal propositional logic. For predicate logic, a lower K-line could affect the instantiation of a higher-level, "more abstract" K-line. For example, this could be a way to partly instantiate one frame [3] with other frames at its terminals. Thus, a group of K-lines could activate a frame displacing some of its "defaults assignments" by active sensory recognizers. (See ***Note 9.***)

What else can happen depends on the specifics of the P-logic. One might even be able to design a "detachment" operation to yield deduction chaining via the overall K-P-K-- operation cycle. But I have no detailed proposal about how to do that.

LEARNING and REINFORCEMENT

Generations of experiments have led to many theories about learning -- in animals -- via "reinforcement" of success. But I maintain that no such simplistic, centralized reward mechanism could suffice for human learning because:

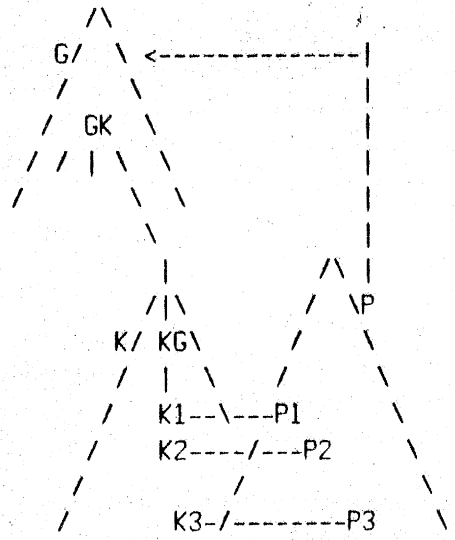
> ***The recognition of what events should be considered "memorable" -- in an intelligent system -- cannot be a single, uniform process. It requires too much intelligence. For the purposes of any particular division of the mind, such recognitions must usually be made by some other agency that has engaged the present one for a purpose.***

To solve hard problems, one needs strategies and tactics that span very different time scales. When a goal is finally achieved, one wants to "reinforce" not only the immediately preceding events, but also the longer range strategy that caused them. But, between selection and completion of a strategic plan, there usually intervene a variety of tactical failures. So at that final moment the traces that remain within the mind's state include all sorts of elements left over from bad decisions and futile experiments.

Traditional behavioristic learning theories rely on "recency" to sort these out. This could work for simple strategies in which the most recent mental events are indeed the best correlated with success. But for human problem solving, I am sure that the "credit assignment" problem is much too complicated for this to work. Instead, I conjecture, different scales of strategies and tactics are segregated in different agencies -- e.g., different P-nets. Then, the learning mechanisms can also be segregated to operate over different time scales. After all, many human cognitive strategies actually achieve their goals by assembling subsidiary learning systems that operate over hours and days. Strategies for dealing with loss and grief, acquisition and ambition, span the years -- yet, sometimes, in the end we learn something from them.

For what that is worth, we conclude that decisions about what and when to "reinforce" cannot be made on a global recency basis. Nor can it be done entirely locally within the K-P pair, for they lack enough information about the intentions of other centers. At the least, it would seem that control over formation of K-P links must be held by a third agency -- either one with innate, unlearned reinforcers, or one that has already learned something.

Consider a model based on these intuitions, in which a third network G, with an active goal-node GK, has the power to construct new K-nodes for P. Suppose that at some earlier time GK was achieved and was connected to a K-node KG that activates two subnodes K1 and K2. At some later time G achieves another instance of GK and celebrates this as memorable. If nothing new happened in P, there is no need to change KG. But suppose a new element K3-->P3 is involved this time: then we could add K3 to KG's K-line, so that P3 will be available for achieving GK in the future.



Of course, this raises all the issues about novelty, conflict, adaptation and saturation that any learning theory must face. (See *Note 12.*) What if P3 later became a direct competitor of P1 or P2? What if there were a mistake? How do we keep the web attached to KG within bounds? After all, there is always *something* new! One can try to invent local solutions to all these problems, but I doubt there is any single, adequate answer. Instead, it must be better always to leave link formation under the control of a distinct system that itself can learn, so that the mnemonic strategies in each locale can be made to suit their circumstances. Perhaps some people become smarter than others because they develop better mnemonic strategies. These might more affect the quality of intelligence than do the specific problem-solving strategies we can observe directly.

Returning to the three-part model, what activates KG? If the G-system could call on a variety of P-nets for its purposes, GK might be selected because of some "cue" involving P that suggests it as a plausible alternative -- e.g., KG is activated by an "*and*" of GK and that P-condition. Through such connections KG becomes part of the representation or meaning of GK -- a remembered solution to a problem. While this raises more questions than it answers, it seems clear that a minimal learning theory will involve at least ***three*** nets -- G, K, and P -- in which the first controls how the second learns to operate the third. This does not mean the system is made of distinct such triplets. Presumably, the same net could play a P-role in one domain and a G-role in another.

TACIT vs. ARTICULATE KNOWLEDGE.

It is commonplace to distinguish between "tacit" knowledge (like how to climb stairs) and "explicit" knowledge (like how to spell "spell"). In a "single-agent" theory, one might wonder how knowledge could possibly be tacit. In a "society of mind" theory, one might wonder how could knowledge ever become "explicit". One cannot expect positive answers in general; only where K-->P connections become somehow linked with such cognitive elements as particular senses of particular words.

It is better to regard the "tacit-explicit" distinction as merely a first approximation to some richer theory of the different kinds of remoteness between one mechanism and another. While surely some agencies in the mind have exceptional expressive roles, each sub-society of the mind must still have its own internal epistemology and phenomenology, with most details private, not only from those central processes, but from one another.

In my view, self-awareness is a complex, constructed illusion. As adults we rightly place high value on the work of those mental agencies that acquire powers to reflect on the behavior of other agencies -- especially our linguistic and ego-structure mechanisms. But probably no part of any mind can ever see very deeply into other parts; it can only use models it constructs of them. Any theory of intelligence must eventually explain the agencies that make models of others: such self-awareness is probably essential to highly intelligent thought, because thinkers must adapt their strategies to the available mental resources.

Each part of the mind sees only the surface products of some other parts. What little we can "directly" sense is swiftly refined, reformulated and "represented". We find it useful to believe that these fragments have

meanings in themselves apart from the great webs of structure from which they emerge. That illusion (valuable to people *qua* thinkers but not *qua* psychologists) leads us to think that expressible knowledge is the first thing to study. If the present theory is right, this is topsy-turvy; most knowledge stays more or less where it was formed, and does its work there. It is the exception, not the rule, that lets one speak of what one knows.

To say much more about this would engage a world of issues beyond the bounds of this little theory. I mean to indicate no pessimism in saying that explaining the meanings of memories will need many more little theories beyond this one. We can understand the "meanings" in the parts of our minds if -- and only if -- we can model enough of them inside others. But this is no different from understanding anything else -- except perhaps harder.

Cambridge, Massachusetts
January - June, 1979

NOTES

I gratefully acknowledge valuable discussions about K-lines with D. Hillis, G. J. Sussman, W. Richards, Jon Doyle, R. J. Solomonoff, R. Berwick, and especially S. Papert -- for the basic idea came in conversations with him.

Note 1: Background. The references to the "Society of Mind" relate to a theory I have been evolving jointly with S. Papert. That theory tries to explain thought in terms of many weakly interacting (and often conflicting) specialists, rather than in terms of a centralized and logically consistent system. It is described briefly in [1], which the present paper complements in several areas. The computational structures described therein were confusingly bidirectional, and the K-P duality clarifies that a little. The C-lines of that paper correspond roughly to the K-->P connections here. The discussion in [1] of *cognitive cases* and of *differences* supplement the discussion here of goals. But I do not mean to pretend that the reader should be able to figure out, even from both papers, exactly what happens in P-nets; we simply haven't fixed the details.

Note 2: Dispositions. The term "disposition" is used here in its ordinary language sense to mean "a momentary range of possible behaviors". I don't see a way to define it technically without making it synonymous with "state", which does not capture the same intuition. In a computer program, a disposition could be imposed by selecting which items are active in a data base, e.g., as in Doyle's [4] flagging of items that are "in" and "out" of current consideration.

The term "representation" also has problems. It always involves three agents -- *A* represents *B* as *C*. In a Mind theory, *A* might be either part of the mind or the theorist himself; one must be clear about that! In this paper, a "K-node" imposes a disposition on a P-net hence, ***for us***, that node can represent that disposition. But what it represents ***for the mind that contains it*** is another matter we touch on only at the very end of the paper.

Note 3: Modularity. This is not to say that understanding memories of feelings should be easier than understanding memories of facts. The latter appear simpler in the adult perspective of "modular" knowledge, because a lifetime of mental theory-construction builds for us our orderly, commonsense epistemological hierarchies. A fragment of incremental knowledge -- e.g., that ducks have webbed feet -- is easy to "represent", once we have only to link together a few already established structures. But should not mistake that surface smoothness for simplicity of underlying process. It captures little of the real quality of "meaning" -- of how such linkages participate in the total "web" of our dispositions.

Note 4: Brains. Some might object that we just don't know enough about brains to make such theories! But we are not proposing specific neurological details -- only that things are organized along the general lines of the Society theory. This architectural theory is just another form of information processing theory, emphasizing control structure and data flow rather than data structure.

Note 5: Unidirectionality. It is technically very difficult to make theories about systems that allow large degrees of circular behavior. On the other hand, one cannot base a theory of mind on unidirectional networks, because loops and feedback are essential for non-trivial behavior. This is why it has been so difficult to pursue the field of "neural net" models, and why so little has happened therein since the works of Hebb [5] and Marr [6].

What I find satisfying is the way the present theory introduces the required circularity in a controlled way. It begins with a nearly unidirectional network, avoiding mathematical universality and its usual theoretical intractability. (The lateral cross-exclusion of the P-nets still leaves basically unidirectional behavior.) Then, feedback loops are built up as steps in training the K-net. Surely this strategy lends itself to circuits that are manageable and debuggable. With the loops introduced a little at a time, one can watch for instability and oscillation, distraction and obsession.

We note that K-logic must be more complex than as described above. If activating a K-node recursively activates subordinates all the way down, this would vitiate the level-band idea. I do not see any easy local way to deal with this; it suggests that the activity band of a K-P pair should be controlled, not locally, but by some other agency -- using a facilitation signal (with low spatial resolution) that enhances the activity in a selected level band. Such an agency could bias the ascent or descent of the K-P computation, without needing to understand much of the details of the events within K-P. In effect it could instruct K-P to "try a more general method" or to "pay more attention to the input" or, perhaps, to "try another like *that*", and so forth.

Such an agency would provide a locus for high-level heuristic knowledge about how to use the knowledge within K-P, and would be useful for implementing plans, looking ahead, and backing up. It might be the natural place for our all-important knowledge about knowledge".

More speculatively, perhaps the difficulty of dealing with too-circular networks is no mere human limitation. Evolution itself probably cannot cope with the uncontrolled range of recursive network behaviors. So, we speculate, the individual nervous system *has* to evolve its circularities by separating the flows into distinct directional classes. If the present theory were correct, this would suggest an evolutionary pressure that might have led to it.

Note 6: Global Architecture. An entire brain would contain many such P-structures associated with different, genetically specified functions: sensory, motor, affective, motivational, and whatever. The present theory would apply only to the common properties of neocortex; the brain contains many other kinds of structures. Finally, I repeat that the "pyramid" image is only relative to any particular "agent". There is no reason to suppose that P-structures need either narrow or widen as they ascend.

Note 7: Excitation. G.A. Miller pointed out to me that this resembles the idea of "redintegration" popular in an earlier era of psychology. Note that we do not need to add "negative" K-line connections to agents that were inactive when EK occurred; many of them will be automatically suppressed by cross-exclusion via AK. Others may persist, so that the partial hallucination may include additional elements. It is perhaps of interest that (according to Mountcastle [7]) all lines entering the cortex from other centers are excitatory.

Note 8: Accuracy. Only a naive theory of memory would depend critically on first-time perfect recollection. Many agents active during EK will be "inessential" to most new situations, so we need not demand perfect and complete attachments; indeed we will need ways to correct serious errors later. In the early days of simple neural models one might have welcomed "sampling noise" as a desirable source of "variety." In systems as complicated as the present one, that view is obsolete; the problem is, rather, of finding heuristics to restrict excessive variation.

Note 9: Fringes and Frames. In this sense, a K-node acts like a "frame", as described in [3]. When a K-node activates agents in the level-band below it, these correspond to the essential, obligatory terminals of the frame. ***By making K-lines have "weaker" connections at its lower fringes, we obtain much of the effect of the loosely bound "default assignments" of the frame theory.*** For, weakly activated agents will be less persistent in cross-exclusion competition. What about the upper fringe? This might be related to the complementary concept of a "frame-system", emphasized in [3]. A failure of the P-net to do anything useful could cause it to default control to a slightly higher-level goal type; that is, to move up in abstraction level. All this comes simply from making weak connections at the fringes of the level-band.

I recognize that this argument about the upper limit is much less clear then for the lower limit. I have no really strong reason even to insist that the upper fringe end below PK, except for an overall feeling of consistency. In fact, PK wasn't defined clearly in the first place, except for locating the level bands. But the asymmetry really comes from the murkiness of my explanation of how the "P-->K" connections relate P-structures to goals and actions. At the end of the paper are a few incomplete suggestions about these matters.

Note 10: Crossbar Problem. I am not very much concerned about this problem's size, because I envision the mind as employing a few thousand P-nets, each with a few thousand Agents. So the local crossbar problem, which is the one that concerns me most, involves only thousands of lines, not millions; that is, K-lines must have access to that order of connections. As for interconnecting all the P-nets, this must be the function of the brain's white matter; we argue in [1] that one need not suppose all P-nets can or need to communicate with each other.

There exist communication-hardware schemes more physically efficient than point to point wiring. Since the density of actual K-line connections is surely sparse in the space of all possible such connections, they could use such schemes as those of Mooers [8] or Willshaw et al. [9]. To implement these, one would make available a large bundle of descending conductors -- call them M-lines. To simulate a K-line, attach the K-node to excite a small, fixed, but randomly assigned, subset of M-lines. To connect it to another K-node, the latter must first construct the corresponding "logical and", then "logical or" that into its excitation condition. Using 10-line subsets of a 100-line bundle would suffice for very large K-pyramids. The final chapter of Fahlman's thesis [10] speculates on other radical crossbar schemes.

Note 11: Winston Learning. Because I consider Winston's [11] the most interesting constructive theory of abstraction, I will try to relate it to the present theory. "Emphasis links" can be identified with K-lines to members of cross-exclusion groups. But "prevention pointers" must enable specific P-agents to disable higher level class-accepting agent; I do not see any easy way to do that. Crucial to Winston's scheme is the detection and analysis of Differences. To make our system able to do this, one might want K-line attachment to prefer P-agents whose activation status has recently changed; then, perhaps by some "blinking" of input contexts, the networks could be made to detect and learn differences.

Generally, in this essay, I have suppressed any discussion of sequential activity. Of course, a K-node could be made to activate a sequence of other K-nodes. But I considered such speculations to be obvious, and that they might obscure the simplicity of the principal ideas.

Winston's scheme emphasizes differences in "near miss" situations; in a real situation there must be a way to protect the agents from dissolution by responding too actively to "far misses". Perhaps a broader form of cross-exclusion could separate the different senses of a concept into families. When serious conflicts result from a "far miss", this should disable the confused P-net so that a different version of the concept can be formed in another P-net.

Note 12: Saturation. In the present theory, one only adds connections and never removes them. This might lead to trouble. Does a person have a way to "edit" or prune his cognitive networks? Well, the present theory, like any other simple psychological theory, must proceed through stages -- just as does its subject matter. Perhaps the Winston theory could be amended so that only imperative pointers long survive. Perhaps the cross exclusion mechanism is adequate to refer low-level confusions to higher level agents. Perhaps, when an area becomes muddled and unreliable, we replace it by another -- perhaps using a special revision mechanism. Perhaps in this sense we are all like the immortal people in Arthur Clarke's novel [12], who from time to time erase their least welcome recollections.

REFERENCES

[1] Minsky, M., "Plain Talk about Neurodevelopmental Epistemology". IJCAI, 1977. Condensed version in Winston and Brown (eds.), *Artificial Intelligence*, vol. 1, MIT Press, 1979

[2] --- and S. Papert, *Artificial Intelligence*, Univ. of Oregon Press, 1974

[3] Minsky, M., "A Framework for Representing Knowledge", in Winston (ed.) *Psychology of Computer Vision*, McGraw-Hill, 1975

[4] Doyle, J., *A Truth Maintenance System*, MIT AI Memo 521, 1979

[5] Hebb, D.O., *Organization of Behavior*, Wiley, 1949

[6] Marr, D., "A Theory of Cerebellar Cortex", *J. Physiology*, vol. 202, pp437-470 , 1969. Also, "A Theory for Cerebral Neocortex", *Proc. Roy, Soc. Lond.*, vol. 176B, pp161-234, 1970

[7] Mountcastle, V., in *The Mindful Brain*, F. Schmitt (ed.), MIT Press, 1978

[8] Mooers, C.E., "Zatocoding and Developments in Information Retrieval", *ASLIB proceedings*, vol. 8, no. 1, pp3-22, Feb, 1956

[9] Willshaw, P.J., Buneman, O.P. and H.C. Longuet-Higgins, "Non-Holographic Associative Memory, *Nature*, vol. 222, pp960-962, 1969.

[10] Fahlman, S., *NETL: a System for Representing and Using Real-World Knowledge*, MIT Press, 1979

[11] Winston, P., "Learning Structural Descriptions from Examples", in Winston (ed.), *Psychology of Computer Vision*, McGraw Hill, 1975

[12] Clarke, A.C., *The City and the Stars*, Signet, 1957