

K-Lines: A Theory of Memory *

Marvin Minsky

Massachusetts Institute of Technology

June 1, 1979 (First Issued Date)

(Reformatted by Yida Xin on September 28, 2019)

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 retrieved?
- How is it stored?
- 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 partial mental state that created that *K-line*. 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.

*Minsky, Marvin. “K-Lines: A theory of Memory.” *Cognitive Science* 4 (1980): 117-133.

1 Introduction

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?

New situations are never exactly the same as old, so if 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, via the thesis that *the function of a memory is to recreate a state of mind*. Hence, each memory must embody information that can later serve to reassemble the mechanisms that were active when the memory was formed, thus recreating a “memorable” brain event. (See Note 1.) More specifically:

When you “get an idea” or “solve a problem” or have a “memorable experience,” you create what we shall call a “**K-line**.” This **K-line** gets connected to those “mental agencies” that were actively involved in the memorable mental events. When that K-line is later “activated”, it re-activates some of those mental agencies, creating a “partial mental state” resembling the original.

To turn this intuitive idea into a substantive theory, we have to explain (1) “mental agencies”, (2) how “**K-lines**” interact with them, (3) “partial mental states”, and (4) how all this relates to conventional ideas about meaning and memory.

1.1 Dispositions vs. Propositions

In this modern era of “Information-processing Psychology,” it may seem quaint to talk about *mental states*; it’s more fashionable to speak of representations, frames, scripts, or semantic networks. But while I find it lucid enough to speak in such terms about memories of things, sentences, or even faces, it is much harder so to deal with feelings, insights, and understandings—and all the attitudes, dispositions, and ways of seeing things that go with them. (See note 2.) We usually put such issues aside, saying that one must first understand simpler things. But what if feelings and viewpoints **are** the simpler things? If such dispositions are the *elements* of which the others are composed, then we must

deal with them directly. So we shall view memories as entities that predispose the mind to deal with *new* situations in *old, remembered* ways—specially, as entities that reset the states of parts of the nervous system. Then, they can cause that nervous system to be “disposed” to behave as though it remembers. This is why I put “dispositions” ahead of “propositions.”

The idea proposed here—of a primitive disposition-representing structure—would probably serve only for a rather infantile dispositional memory; the present theory does not go very far toward supporting the more familiar kinds of cognitive constructs we know as adults. But I would not expect to capture all that at once in one simple theory; I doubt that human memory has the same uniform and invariant character throughout development, and I don’t want to attribute to infants capacities that develop only later.

1.2 Mental States and The Society of Mind

One could say little about “mental states”, if one imagined the Mind to be a single unitary thing. Instead, we shall envision the mind (or brain) as composed of many partially autonomous “agents,” as a “Society” of small minds. This allows us to interpret “mental states” and “partial mental states” in terms of *subsets of the states of the parts of the mind*. To give this idea substance, we must propose some structure for that “Mental Society.” In fact, we’ll suppose that it works much like any human administrative organization.

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

Each Division is itself a multitude of sub-specialists—call them “agents”—that embody smaller elements of an individual’s knowledge, skills, and methods. No single one of these little agents need know 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 specification of all the agents that are active. A **partial mental state** is a partial such specification: It specifies the activity-state of just some of the agents.

It is easiest to think about *partial states* that constrain only agents within a single Division. Thus, a visual partial state could describe some aspects of an imagery process, without saying anything about agents that are outside the Visual Division. In this paper, our main concern will be with yet “smaller” partial states that constrain only **some** agents within one Division.

This concept of partial state allows us to speak of entertaining *several partial states* at once—to the extent that they do not assign different states to the same individual agents. And even if there is such a conflict, the concept may still be meaningful, if that conflict can be settled within the Society. This is important, because (we suggest) the local mechanisms for resolving such conflicts could be the precursors of what we know later as *reasoning*—useful ways to combine different fragments of knowledge.

In the next few sections, we describe, in more detail, the **K-nodes** and **K-lines** that are proposed as the elements of memory. Activating a **K-node** will impose a specific partial state upon the Society by activating the agents connected to its **K-line**, and this will induce a certain *computational disposition*. Now, while it’s fairly easy to see how

such elements could be used in systems that learn to recognize arrangements of sights and sounds, the reader might suppose that it must be much harder so to capture recollections of attitudes, points of view, or feelings. But one must not assume that “concrete” recollections are basically the simplest—that’s an illusion reflecting the enormous competence of the adult mental systems that we have evolved for communicating about concrete matters. I mention this, lest that illusion fools us, as theorists, into trying to solve the hardest problem first.

Concrete concepts are *not* necessarily the simplest ones. (See Note 3.) A novice best remembers “being at” a concert, and something about how it affected him. The amateur remembers more about what it “sounded like.” Only the professional remembers much about the music itself, such as timbres, tones, and textures. So, the most concrete recollection may require the most refined expertise. Thus, while our theory might appear to be putting last things first, I maintain that attitudes do really precede propositions, and feelings do come before facts. This seems strange only because we can’t remember what we knew in infancy.

1.3 Memories and Partial Brain States

Old answers never perfectly suit new questions, except in the most formal, logical circumstances. To explain how memories could then be useful, one might consider various theories:

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

Our theory most resembles the latter, in remembering not the stimulus itself but the part of the state of mind that the stimulus caused. When one is faced with a new problem, one may be able to solve it, if one is “reminded of” some similar problem that was solved in the past. How does this help? It’s not enough to just “remember the solution” for that similar problem in the past — unless the current situation is exactly the same as the old situation, some work will have to be done to adapt the current situation to the old situation. What’s better, we suggest, is to get the mind into the (partial) state that solve the old problem, and then the mind might be able to handle the new problem in “the same way”. To be more specific, we must sketch more of the architecture in which our Agents are embedded. (See Note 4.)

We envision the brain to contain a great lattice of “agents,” each one connected to only a few others. We further suppose that an agent’s inputs come either from below or from the side, while 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: Simple feature- or texture-detectors at the bottom, edge- and region-sensing agents above them, and identifiers of more specific objects or images at higher levels.

Given these connection constraints, if one “looks down” from the viewpoint of a given agent **P**, one will see other agents arranged roughly in a hierarchical **Pyramid** (See Figure 1). Note that, although we shall thus talk about “pyramids,” that shape is merely an illusion from the agent’s perspective. The network, as a whole, need not have any particular shape.

1.4 Cross-Exclusion and Persistence

We mentioned that information can flow laterally, as well as upwards, in a pyramid. Unrestricted lateral connections would permit feedback and reverberatory activity. However, we shall assume that all the cross-connections are essentially *inhibitory*, which rules out such feedback or reverberatory activity. We assume this, because in our concept of *The Society of Mind*, agents tend to be grouped in small “cross-exclusion” arrangements, where each member within such a cross-exclusion group sends inhibiting signals to other members within the same cross-exclusion group. This makes it hard for more than one agent, in each group, to be **active** at any given time. Any active agent, then, tends to suppress its associates, which in turn weakens those associates’s inhibiting effect on that active agent itself. This kind of sub-structure is familiar in physiology.

A network composed of cross-exclusion groups has a kind of built-in “short-term memory.” Once such a network is forced into a partial state, even for a moment, that partial state will tend to persist. To reset the whole network’s state, one needs only activate, transiently, one agent in each cross-exclusion group. Afterwards, the new sub-states will tend to persist—except for those agents under strong external pressure to change. To an outside observer, these internal persistences will appear as “dispositions” —distinctive styles of behavior. Changing the states of many agents grossly alters the behavior, but changing only a few agents’s states just perturbs the overall disposition a little.

The temporal span of an agent’s disposition will depend on its place in the hierarchy: The states of low-level agents change frequently in response to signals that ascend from outside or from other **P-nets**; but the states of high-level agents are presumably bound to plans and goals of longer duration. In the following theory, it will be the *intermediate-level* agents that are most involved with the memories associated with each particular **P-net**, because they must help to determine how the agents below them influence the agents above them. Of course, this notion of “intermediate” probably must be defined in terms of development; its locus will presumably move upward during cognitive growth. (See Note 6.) For example, a lowest-level agent in the visual system would always compute the same function of retinal stimulation;. But at higher levels, different dispositions induce different “ways of seeing things.” Thus, the choice between the three natural perspectives for seeing the Necker Cube is dictated not by ascending sensory information, but by decisions in other agencies. Similarly, one needs non-sensory information to dispose oneself to regard a certain sounds as noise or word—an image as thing or picture.

2 K-Lines and Level Bands

Our theory will propose that around each **P-pyramid** grows another structure—the “**K-pyramid**”—that embodies a repertory of dispositions, each of which is defined by preactivating a different subset of **P-agents**. The **K-pyramid** is made of “**K-nodes**,” each of which can excite a collection of **P-agents**, via its “**K-lines**.” To explain the idea, suppose that one part, **P**, of your mind has just experienced a mental event **E**, which led the mind to achieve some goal—call it **G**. Suppose another part of your mind declares this to be “memorable.” Then, we postulate that these two following things must take place:

- **K-node Assignment:**
A new agent, called a **K-node KE**, is created and somehow linked with **G**.
- **K-line Attachment:**
Every **K-node** comes with a wire, called its **K-line**, that has potential connections to every agent in the **P-pyramid**. The act of “memorizing” causes this **K-line** to create an *excitatory* attachment to every currently active **P-agent**.

Consequently, when **KE** is activated at some later time, this will make **P** “reenact” that partial state—by arousing those **P-agents** that were active when **E** was “memorized.” Thus, the activation of **KE** causes the **P-net** to become “disposed” to behave the way it was working when the original goal **G** was achieved. What happens if *two K-nodes* are activated? Since we are talking of partial (not *total*) states, it is possible for a single **P-pyramid** to maintain fragments of *several* dispositions at one time. Of course, if the two dispositions send conflicting signals to an agent, there is a problem, which we discuss later.

It might also seem impractical to require that every **K-line** come near to every **P-agent**. Now we introduce a series of “improvements” that not only alleviate this requirement, but also combine to form a powerful mechanisms for abstraction and inference.

To begin with, the schema that we just described would tend to reset the entire **P-net**. This would amount to making **P** to virtually “hallucinate” the event **EK**. (See Note 7.) But, on reflection, one sees that it is *not* the purpose of memory to produce hallucinations. (See Note 8.) Rather, *one wants to reenact only enough to “recapture the idea.”* Complete hallucinations would be harmful; resetting the whole **P-net** would erase all the work recently done—and might even fool one into seeing the present problem as already solved. Instead, memory must induce a state that remains sensitive to the new situation. We conclude that *a memory should induce a state through which we see current reality as an instance of the remembered event*—or, equivalently, see the past as an instance of the present.

2.1 The Level-Band Principle

We propose to accomplish establishing such a memory by connecting **KE** not to all the **P-agents** that were active during **E**, but only to those within an intermediate band of levels. To explain this, I must assume here that **KE** is somehow associated with some agent **PE** at a certain level of the **P-pyramid**. I will return later to discuss this somewhat

obscure **P**→**K** association. But, assuming that for the moment, we can formulate two important restrictions:

- **Lower Band-Limit**

The **K-line** should not affect agents at levels too far below **PE**, for this would impose false perceptions and conceal the real details of the present problem. (See Note 9.)

- **Upper Band-Limit**

The **K-line** should not reach up to or above the level of **PE** itself, for that would make us hallucinate the present problem as already solved. The **K-line** also should neither change **PE** nor impose onto **PE** too strongly the details of the old solution.

These two constraints combine to suggest the first of the paper's two principal ideas:

PRINCIPAL IDEA #1 – The Level-Band Principle

A **K-line** should span only a certain band of levels somewhere below **PK**.

This induces a disposition that can

- (1) exploit high-level agents that are appropriate to current goals, and
- (2) be sensitive to the current situation, as perceived at lower levels.

So, by activating agents only at intermediate levels, the system can *perform a computation that's analogous to a computation from the memorable past, and yet still remain sensitive to present goals and circumstances*. (See Figure 2).

2.2 Connections among **K-Nodes**

The second idea: If memories partially recreate previous states, and if those states are in turn based on other memories, this suggests that **K-lines** should exploit other memories, i.e., other **K-lines**. We can do this via attachments to previously constructed **K-nodes**. In fact, this idea provides a second way to make the scheme more physically plausible:

PRINCIPAL IDEA #2 – The K-Recursion Principle

When you solve a problem, it is usually by exploiting memories from the past. The occurrence of a memorable event **E** is itself usually due, in large part, to activation of already-existing **K-lines**. So, to “memorize” that state, it will usually suffice to attach the new **K-line KE** just to activate **K-nodes**—rather than to all active **P-nodes**!

Connecting **K-lines** to **K-nodes** (rather than **P-nodes**) allows us to compose new memories mainly from ingredients or earlier memories. This should lead to meaningful cognitive structures, especially when combined with the level-band constraint. So, finally:

We connect **KE**, not to *all* recently active **K-nodes**, but only to those in a Level Band (of **K-nodes**) below **KE**! Note that this creates a “**K-pyramid**,” as discussed below.

Of course, if **K-lines** were connected *only* to other **K-nodes**, they would ultimately have no contact with the **P-pyramid**: The process has to start somewhere! I envision the **K-agents** to lie anatomically near the **P-agents** of corresponding levels, making it easy for **K-lines** to contact *either* **P-agents** or **K-agents**. Perhaps in early stages of growth, the connections are primarily to **P-agents**. Then later, under genetic control, the preferences tend to shift over from **P's** to **K's**.

2.3 The Crossbar Problem

We digress for a moment into an issue that concerns the physical “hardware.” Even using both the Recursion and the Level-Band principles, each **K-node** still must have potential junctions with many agents. This problem confronts every brain-theory that tries to explain how the mind is at all capable of any great range of “associations.” We shall call it the “crossbar” problem. The problem is often ignored in traditional programming, because computer memory can be regarded as totally connected, in the sense that any register “address” can connect to any other in a single step. But the problem returns in systems with multiple processors or more active kinds of memory.

One need not expect to find any general solution to this problem. In the cerebrum, the (potential) interconnections constitute almost the entire biomass; the computational hardware itself—the cortical neurons and synapses—is but a thin layer that borders the biomass. *The Level-band principle would have a large effect by lowering the dimensionality of the problem by one.* The advantage of using the recursion principle is not so obvious, but it suggests that local, short connections should suffice for most purposes (See Note 10.)

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 chemical diffusion, or holographic phase-detector—although any such invention might make a brain more efficient). Instead, I would seek the answer within 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 then, most pairs of agents will have no real need to talk to one another; indeed, since they speak (so to speak) different languages, they could not even understand each other. If most communication is local, then the crossbar problem scales to more modest proportions.

Still, the reader might complain that any communication limits within the Mind would seem counter-intuitive: Can't one mentally associate *any* two ideas, however different? Perhaps, but it would seem that making unusual connections is unusually difficult and, often, rather “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.

2.4 The Knowledge-Tree

It will not have escaped the reader that we have arrived at an elegant and suggestive geometry: The **K-nodes** grow into a structure whose connections mirror those of the **P-pyramid**, except that information flows in the other direction. The **K-nodes** form a **K-pyramid**, lying closely against the **P-pyramid**, each with convenient access to the

level-bands of the other. **P-nodes** activate units **above** them, while **K-nodes** activate units **below** them. A typical path of computation within the diagram of Figure 3 tends to traverse a counterclockwise spiral. Over time, the locus of this activity could drift either up or down—presumably controlled by other agencies who demand more generality or specificity.

While this “computation architecture” seems very general and versatile, its apparent symmetry is deceptive, because I suppressed some hard questions. I described the connections **K**, and of those from **K** to **P**. And while I have said little here about the connections within **P**, that is not a major problem—it is discussed in more detail in Minsky (1977). The real problem concerns the link from **P** back to **K**; I said only that “...**KE** is somehow associated with some agent **PE** at a certain level of the **P-pyramid**. ...” We need to provide some relation between **P-events** and the achievement of **Goals** represented elsewhere, and the rest of the essay discusses various possible such relations but does not settle upon any particular one; from this point on, the reader can assume that difficulties in understanding are my fault, not his. But I hope I have supplied an adequate enough framework so as to make plausible these further speculations.

It is tempting to try to find simple ways to restore the symmetry. For example, our **K-trees** learn to adapt to the **P-tree**. But the **P-tree** itself must once have been the learner. Was the **P-tree** once the **K-tree** for another **P-system**? Could they take turns training each other? Alas, nothing so simple will do. We shall argue that nontrivial learning requires at least *three* nets to be involved, because 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.

3 K-Knowledge

We started with a naïve idea that “memories reenact past state”—without attempting to explain what they “mean.” Now, we come full circle: Because 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.*

3.1 K-Knowledge Seen as Logical

What kinds of relations? In the simplest case, when the partial states do not interact much, a superior simply *superposes* the effect of its subordinates. Concurrent activations 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 either to a chair or to a table. So, **K-terms** at comparable levels tend to combine “disjunctively.”

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 the entire cross-exclusion group simply to “drop out” completely*, defaulting to another element at the next-higher level. Returning to the previous example, if the dispositions for “chair”

and for “table” were in some local conflict (e.g., by requiring both “back” and “no back”), then the conflicting agencies should disarm each other—leaving the remaining harmonious elements to accept anything in the next higher “furniture” class!

Papert and I 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 between them, **but** to find another viewpoint! We introduced this idea as a general principle in *Minsky and Papert* (1974), after Papert had observed how it might explain how Piaget’s Conservation develops in children.

3.2 **K-Knowledge Seen as Abstract**

Earlier, we spoke only of creating an entirely new **K-node** for each memorable event. But surely there are more gentle ways to “accumulate” new subordinates to already-existing nodes. Suppose that a chimpanzee achieves the too-high banana by using different means at different times—first using a box, then a chair, and later a table. One could remember these separately. But, if they were all “accumulated” to one single **K-node**, this would lead to creation of a more powerful “**how-to-reach-higher**” **node**: when activated, **this node** would concurrently activate **P-agents** for boxes, chairs, or tables, so that the 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.”

Indeed, it may do much better than that, in view of our proposed cross-cancellation principle. Suppose, as mentioned above, those conflicts among details cause decision to be made—by default—by those remaining, non-conflicting agents. *The effect is that of a more abstract kind of abstraction—the extraction of common, non-conflicting properties!* Thus, combining the concrete “accumulation” of a particular instances with the rejection of strongly dissonant properties leads automatically to a rather abstract “unification.” (See Note 11.)

3.3 **K-Knowledge Seen as Procedural**

This is more speculative: When **K-lines** interact at different vertical levels, the superposition of several partial states will produce various sort of logical and “illogical” consequences of them. We already know that they can produce simple disjuncts and mutual exclusions. This is probably enough for simple forms of propositional logics. I think that it is possible for such structures also to simulate some kinds of predicate logic. A lower **K-line** could affect the *instantiation* of a higher-level, “more abstract” **K-line**, just as one can partly instantiate one frame (Minsky, 1975) by attaching other frames to some of its terminals. Thus, a **K-line** could “displace” one of a **P-agent**’s “default assignments” by activating a specific sensory recognizer. (See Note 9.) Other specific kinds of logic could be architecturally embedded in the **P-logic**. One might even be able to design a “detachment” operation to perform deduction chaining during the overall **K-P-K** operation-cycle. But I have no detailed proposal about how to do that.

4 Learning and Reinforcement

Most theories of learning have been based on ideas about “reinforcement” of success. But all these theories postulate a single, centralized reward mechanism. I doubt this could suffice for human learning, because the recognition of which events should be considered memorable cannot be a single, uniform process. It requires too much “intelligence.” Instead, I think that such recognitions must be made, for each division of the mind, by some other agency that has engaged the present one for a purpose.

Hard problems require strategies and tactics that span different time scales. When a goal is achieved, one must “reinforce” not only the most recent events, but also the strategies that caused them. At such a moment, the traces that remain within the mind’s state include all sorts of elements left over from both good and bad decisions. Traditional behavioristic learning theories rely on “recency” to sort these out, but strategy-based activities create “credit assignment” problems too complex for this to work. However, if we segregate different strategic time-scales in different **G-P-K** systems, then they can operate over appropriate time-scales. Our everyday activities seem to involve agencies that operate and learn over seconds, minutes, hours, and days. Strategies for dealing with ambition and acquisition, loss and grief, may span years. Furthermore, decisions about what and when to “reinforce” cannot be made within the **K-P** pairs, for those decisions must depend, to some extent, on the goals of other centers.

We conclude that control over formation of links between **K** and **P** must be held by yet a third agency. Based on these intuitions, suppose that a third network, **N**, has the power to construct new **K-nodes** for **P**. Suppose that, at some earlier time, some goal **G** (represented in **N**) is achieved and was connected to a **K-node**, **KE**, that (for example) activates two sub-nodes, **K1** and **K2**. Suppose that, at a later time, **N** achieves another instance of **G** and celebrates this as memorable. If nothing new has happened in **P**, there is no need to change **KE**. But if a new element **K3**→**P3** is involved, we could add **K3** to **KE**’s **K-line**, making **P3** available for achieving **G** in the future. (See Figure 4.)

This raises all the issues about novelty, conflict, adaptation, and saturation that any learning theory must face. (See Note 12.) What if **P3** were a direct competitor of **P1** and **P2**? What if there were a mistake? How do we keep the attachments to **KE** 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 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. What activates **KE**? It should be possible for the goal-type, **G**, to call upon some variety of **P-nets**. Selecting **P** (that is, **KE**) in particular would presumably depend on use of some “cue” involving **P**—e.g., by making **KE**’s activation depend on an “and” condition involving **G** and that **P-condition**. Through such connections, **KE** becomes part of the meaning of **G**—a remembered solution to a problem. This sketchy argument suggests that a minimal learning theory needs at least *three* nets, in which the first controls how the second learns to operate the third. The triplets may not be distinct, because the same net might play a **P-role** in one domain and a **G-role** in another.

It is commonplace to distinguish between “tacit” knowledge (e.g., how to walk) and “explicit” knowledge (e.g., to solve a quadratic). In a “single-agent” theory, one might wonder how knowledge could possibly be tacit. In a “society of mind” theory, one might

wonder how any knowledge could ever become “explicit”—this might require **K-nodes** to become linked with such cognitive elements as particular senses of particular words. I discuss some such issues in Minsky (1977).

In any case, the “tacit-explicit” distinction is only a simplistic approximation to some richer theory of internal connections. 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, but carefully constructed illusion: We rightly place high value on the work of those mental agencies that appear able to reflect on the behavior of other agencies—especially our linguistic and ego-structure mechanisms. Some form of self-awareness is surely essential to highly intelligent thought, because thinkers must adapt their strategies to the available mental resources. On the other hand, I doubt that any part of a 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: Each part of the mind sees only a little of what happens in some others, and that little is swiftly refined, reformulated, and “represented.” We like to believe that these fragments have meanings in themselves, apart from the great webs of structure from which they emerge; and indeed, this illusion is valuable to us *qua* thinkers, but not to us as psychologists, because it leads us to think that expressible knowledge is the first thing to study. According to the present theory, this is topsy-turvy; most knowledge stays more-or-less where it was formed, and does its work there. Only in the exception, not the rule, can one really speak of what one knows. To explain the meanings of memories will need many more little theories beyond this one; we can understand the relations among our mental agencies if—and only if—we can model enough of some of them inside the others. But this is no different from understanding anything else—except perhaps harder.

Acknowledgements

I gratefully acknowledge valuable discussions about K-lines with D. Hillis, G. J. Sussman, W. Richards, Jon Royle, R. J. Solomonoff, R. Berwick, and especially S. Papert.

References

1. Clarke, A.C., *The city and the stars*. New York: Signet, 1957.
2. Doyle, J., *A truth maintenance system*. MIT Artificial Intelligence Memo. No. 521, 1979.
3. Hebb, D. O., *Organization of behavior*. New York: Wiley, 1949.
4. Marr, D., A theory of cerebellar cortex. *Journal of Physiology*, 1969, *202*, 437-470.
5. Marr, D., A theory of cerebellar neocortex. *Proceedings of the Royal Society of London*, 1970, *176B*, 161-234.
6. Minsky, M., A framework for representing knowledge. In P. Winston (Ed.), *The psychology of computer vision*. New York: McGraw-Hill, 1975.
7. Minsky, M., Plain talk about neurodevelopmental epistemology. Proceedings of the Fifth International Joint Conference on Artificial Intelligence, Boston, Mass., 1977. [Condensed version in P. Winston and R. Brown (Eds.), *Artificial Intelligence*, Vol. 1, Boston: MIT Press, 1979.]
8. Minsky, M. & Papert, S., *Artificial Intelligence*, Condon Lectures, University of Oregon, Eugene, Oregon, 1974.
9. Mooers, C. E., Datocoding and developments in information retrieval. *ASLIB Proceedings*, 1956, *8*, 3-22.
10. Mountcastle, V. In F. Schmitt (Ed.), *The mindful brain*, Boston: MIT Press, 1978.
11. Willshaw, P. J., Buneman, O. P. & Longuet-Higgins, H. C., Nonholographic associative memory. *Nature*, 1969, *222*, 960-962.
12. Winston, P., Learning structural descriptions from examples. In P. Winston (Ed.), *The psychology of computer vision*. New York: McGraw-Hill, 1975.

Notes

Note 1: Background

The references to the “Society of Mind” relate to a theory I have been evolving, jointly with S. Papert, in which we try to explain thought in terms of many weakly interacting (and often conflicting) specialists. It is described briefly in Minsky (1977), which the present paper complements in several areas. The **C-lines** of that paper correspond roughly to the **K→P** connections here. The discussion in Minsky (1977) of *cognitive cases* and of *differences* supplement the discussion here of goals, but there is not enough detail, even in both papers, to specify exactly what happens in **P-nets**. We hope to clarify this in a forthcoming book.

Note 2: Dispositions

I use “disposition” to mean “a momentary range of possible behaviors;” technically, it is the shorter-term component of the state. In a computer program, a disposition might depend upon which items are currently active in a database, e.g., as in Doyle’s (1979) flagging of items that are “in” and “out” in regard to making decisions.

The term “representation” always should involve three agents—**A** represents **B** as **C**. In a mind theory, **A** might be part of the mind or part of the theorist himself; most discussions are muddled about this. In the present paper, **K-nodes** impose dispositions on **P-nets** hence, *for us as theorists*, **K-nodes** can represent dispositions. But what they represent *for the mind that contains them* is another matter we address only in passing at the end of the paper.

Note 3: Modularity

Most people would assume that understanding memories of feelings should be harder than understanding memories of facts. But I think the latter appears simpler only in the adult perspective of “modular” knowledge, based on a lifetime of constructing our elderly, commonsense epistemological hierarchies. A fragment of incremental knowledge—e.g., that ducks have webbed feet—seems easy to “represent,” once we only have to link together a few already-established structures. But this apparent, surface smoothness should not be mistaken for underlying simplicity, for it conceals the deeper ways in which each event’s meanings become involved in the total “web” of our dispositions. I think it no accident that in popular culture, feelings are considered inexplicably complex, while thoughts are simple. But, in the culture of psychiatry, of professional concern with real mental activity, it is feelings that are analyzed (more or less successfully) while thoughts are found to be too intricate to understand in any useful detail.

Note 4: Brains

Not enough is known about the nervous system to justify proposing specific details. In our references to brains, our intention is to suggest that it might be useful to consider architectural hypotheses compatible with the general ideas of the society-of-mind approach.

Note 5: Unidirectionality

It is technically very difficult to theorize about systems that allow large degrees of circular behavior. On the other hand, no mind can be based on unidirectional networks, because loops and feedback are essential for nontrivial behavior. This, I think, is why so little has happened in the field of “neural net” models, since the works of Hebb (1949) and Marr (1969; 1970). A feature of the present theory is how it introduces the required circularity in a controlled way: It begins with a nearly unidirectional network, avoiding some of those problem. (The later cross-exclusion still leaves basically unidirectional behavior.) Then, feedback loops are built up as steps in training the **K-net**, yielding a strategy that 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.

Now consider a speculation: Perhaps the difficulty of dealing with too-circular networks is no mere human limitation. Evolution itself probably cannot cope with uncontrolled recursive behaviors. If the present theory were correct, this suggests an evolutionary pressure behind its development: Even the individual nervous systems must evolve its circularities by controlled interconnection of unidirectional flows.

Finally, we note that **K-logic** must be more complex than as described here, because **K-node** activation should not propagate to subordinates all the way down. That would vitiate the level-band idea. This suggests that perhaps the activity band of a **K-P** pair should be controlled, not locally, but by some other agency that uses a low spatial resolution signal to enhance the activity in a selected level-band. Such an agency could control the ascent or descent of the **K-P** computation—e.g., to instruct **K-P** to “try a more general method” or to “pay more attention to the input” or, perhaps, to “try another like that.” 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.

Note 6: Global Architecture

An entire brain would contain many different **P-structures** associated with different functions: sensory, motor, affective, motivational, and whatever, interconnected according to genetic constraints. The present theory might apply only to the common properties of neocortex; the brain contains many other kinds of structure. Incidentally, the idea of “middle level” agent is not precisely defined. In my image of mental development, the definition of this intermediate region of agents will tend to move upwards during cognitive growth.

Note 7: Excitation

We do not need to add “negative” **K-line** connections to agents that were inactive when **E** 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. According to Mountcastle (1978), all lines entering the cortex from other centers are excitatory.

Note 8: Accuracy

Only a naïve theory of memory would require first-time perfect recollection. Many agents active during **E** will be “inessential” to new situations, so we need not demand exact replication. (Indeed, the theory will need a way to undo serious errors.) In the early days of neural modeling, one found some workers who welcomed “sampling noise” as a desirable source of “variety.” I consider that view obsolete now, when the problem is, instead, to find control structures to *restrict* excessive variation.

Note 9: Fringes and Frames

In this sense, a **K-node** acts like a “frame,” as described in Minsky (1975). When a **K-node** activates agents in the level-band below it, these correspond to the essential, obligatory terminals of the frame. If **K-lines** have “weaker” connections at their lower fringes, we should obtain much of the effect of the loosely bound “default assignments” of the frame theory, for then the weakly activated agents will be less persistent in cross-exclusion competition. What about the upper fringe? This might relate to the complementary concept of a “frame-system,” emphasized in Minsky (1975): A failure of a **P-net** to do anything useful could cause control to pass, by default, to a competitive goal or plan (via the weak cross-exclusion), or to move upwards to a slightly higher-level goal type. It would be interesting if all this could emerge simply from making weaker connections at the fringes of the level-band.

I recognize that my arguments concerning upper fringes are weaker than those for lower fringes. My intuition that a level-band not include those agents that activate it embodies the idea of controlled circularity, but it is responsible also for the murkiness of my explanation of how the “**P-K**” connections relate **P-structures** to goals and actions. In fact, **P-K** was defined, in the first place, to give the reader a mental reference point for describing the level structures, but the **P-K** connection itself, involving at least another network, is only a functional concept. Incidentally, in regard to *motor* behavior, some of the image probably must be inverted, because action is somewhat dual to perception, with flow from intent to detail rather than from detail to recognition.

Note 10: Crossbar Problem

I conjecture that the popularly conceived need for holistic mechanisms may be ameliorated if we envision the mind as employing a few thousand **P-nets**, each with a few thousand **Agents**. This would factor the problem into two smaller crossbar problems, each involving only thousands of lines, not millions. In fact, we argue in Minsky (1977) that one need not suppose all **P-nets** can or need to communicate with each other.

While, in this view, there might well be enough white matter for the connections among **P-nets**, there do exist communication-hardware schemes more physically efficient than point-to-point wiring—e.g., the schemes of Mooers (1956) or Willshaw et al. (1969). To implement one of these within a **K-net**, one might use a 100-line bundle of descending conductors. To simulate a **K-line**, attach the **K-node** to excite a small, fixed, but randomly assigned, subset of these. Then, connection to another **K-node** needs a conjunctive recognizer for that subset. Ten-line subsets of a 100-line bundle would suffice for very large **K-pyramids**, and the recognizer might be a rather elementary perceptron.

Note 11: Winston Learning

Because Winston (1975) describes the most interesting constructive theory of abstraction, I will try to relate it to the present theory. “Emphasis links” are easily identified with **K-lines** to members of cross-exclusion groups, but “prevention pointers,” which must enable specific **P-agents** to disable higher-level class-accepting agents, are a problem that perhaps must be handled within **P**—rather than within the **K-line system**. Perhaps more basic to Winston’s scheme is the detection and analysis of Differences; this suggests that **K-line** attachment should be sensitive to **P-agents** whose activation status has recently changed.

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, however, 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. Then, when a serious conflict results from a “far miss,” this would 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? I presume that the present theory will have to be modified to allow for this. 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 confusion 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 (1957), who, from time to time, erase their least welcome recollections.

(Figures on the remaining pages.)

