A Framework for Representing Knowledge

5

Marvin Minsky
1974

1 Frames

It seems to me that the ingredients of most theories both in artificial intelligence and in psychology have been on the whole too minute, local, and unstructured to account—either practically or phenomenologically—for the effectiveness of common-sense thought. The "chunks" of reasoning, language, memory, and perception ought to be larger and more structured; their factual and procedural contents must be more intimately connected in order to explain the apparent power and speed of mental activities.

Similar feelings seem to be emerging in several centers working on theories of intelligence. They take one form in the proposal of Papert and myself (1972) to divide knowledge into substructures, "microworlds". Another form is in the "problem spaces" of Newell and Simon (1972), and yet another is in the new, large structures that theorists like Schank (1973), Abelson (1973), and Norman (1973) assign to linguistic objects. I see all these as moving away from the traditional attempts both by behavioristic psychologists and by logic-oriented students of Artificial Intelligence in trying to represent knowledge as collections of separate, simple fragments.

I try here to bring together several of these issues by pretending to have a unified, coherent theory. The paper raises more questions than it answers, and I have tried to note the theory's deficiencies.

Here is the essence of the theory: when one encounters a new situation (or makes a substantial change in one's view of the present problem), one selects from memory a structure called a *frame*. This is a remembered framework to be adapted to fit reality by changing details as necessary

A *frame* is a data structure for representing a stereotyped situation, like being in a certain kind of living room, or going to a child's birthday party. Attached to each frame are several kinds of information. Some of this information is about how to use the frame. Some is about what one can expect to happen next. Some is about what to do if these expectations are not confirmed.

We can think of a frame as a network of nodes and relations. The top levels of a frame are fixed, and represent things that are always true about the supposed situation. The lower levels have many *terminals*—slots that must be filled by specific instances or data. Each terminal can specify conditions its assignments must meet. (The assignments themselves are usually smaller subframes.) Simple conditions are specified by *markers* that might require a terminal assignment to be a person, an object of sufficient value, or a pointer to a subframe of a certain type. More complex conditions can specify relations among the things assigned to several terminals.

Collections of related frames are linked together into *frame systems.* The effects of the important actions are mirrored by *transformations* between the frames of a system. These are used to make certain kinds of calculations economical, to represent changes of emphasis and attention, and to account for the effectiveness of imagery.

For visual scene analysis, the different frames of a system describe the scene from different viewpoints, and the transformations between one frame and another represent the effects of moving from place to place. For nonvisual kinds of frames, the differences between the frames of a system can represent actions, cause-effect relations, or changes in conceptual viewpoint. *Different frames of a system share the same terminals*; this is the critical point that makes it possible to coordinate information gathered from different viewpoints.

Much of the phenomenological power of the theory hinges on the inclusion of expectations and other kinds of presumptions. A *frame's terminals are normally already filled with "default" assignments.* Thus a frame may contain a great many details whose supposition is not specifically warranted by the situation. These have many uses in representing general information, most likely cases, techniques for bypassing "logic", and ways to make useful generalizations.

The default assignments are attached loosely to their terminals, so that they can be easily displaced by new items that better fit the current situation. They thus can serve also as variables or as special cases for reasoning by example, or as textbook cases, and often make the use of logical quantifiers unnecessary.

The frame systems are linked, in turn, by an *information retrieval network*. When a proposed frame cannot be made to fit reality—when we cannot find terminal assignments that suitably match its terminal marker conditions—this network provides a replacement frame. These interframe structures make possible other ways to represent knowledge about facts, analogies, and other information useful in understanding.

Once a frame is proposed to represent a situation, a *matching* process tries to assign values to each frame's terminals, consistent with the markers at each place. The matching process is partly controlled by information associated with the frame (which includes information about how to deal with surprises) and partly by knowledge about the system's current goals. There are important uses for the information, obtained when a matching process fails. I will discuss how it can be used to select an alternative frame that better suits the situation.

An apology: the schemes proposed herein are incomplete in many respects. First, I often propose representations without specifying the processes that will use them. Sometimes I only describe properties the structures should exhibit. I talk about markers and assignments as though it were obvious how they are attached and linked; it is not.

Besides the technical gaps, I will talk as though unaware of many problems related to "understanding" that really need much deeper analysis. I do not claim that the ideas proposed here are enough for a complete theory, only that the frame-system scheme may help explain a number of phenomena of human intelligence. The basic frame idea itself is not particularly original—it is in the tradition of the "schemata" of Bartlett and the "paradigms" of Kuhn; the idea of a frame-system is probably more novel. Winograd (1974) discusses the recent trend, in theories of AI, toward frame-like ideas.

In the body of the paper I discuss different kinds of reasoning by analogy, and ways to impose stereotypes on reality and jump to conclusions based on partial-similarity matching. These are basically uncertain methods. Why not use methods that are more logical and certain? Section 6 is a sort of appendix which argues that traditional logic cannot deal very well with realistic, complicated problems because it is poorly suited to represent *approximations* to solutions—and these are absolutely vital.

> Thinking always begins with suggestive but imperfect plans and images; these are progressively replaced by better—but usually still imperfect—ideas.

1.3* Artificial intelligence and human problem solving

In this essay I draw no boundary between a theory of human thinking and a scheme for making an intelligent machine; no purpose would be served by separating them today, since neither domain has theories good enough to explain, or produce, enough mental capacity. There is, however, a difference in professional attitudes. Workers from psychology inherit stronger desires to minimize the variety of assumed mechanisms. I believe this leads to attempts to extract more performance from fewer "basic mechanisms" than is reasonable. Such theories especially neglect mechanisms of procedure control and explicit representations of processes. On the other side, workers in AI have perhaps focused too sharply on just such questions. Neither has given enough attention to the structure of knowledge, especially procedural knowl-

is understandable that psychologists are uncomfortable with complex proposals not based on well established mechanisms, but I believe that parsimony is still inappropriate at this stage, valuable as it may be in later phases of every science. There is room in the anatomy and genetics of the brain for much more mechanism than anyone today is prepared to propose, and we should concentrate for a while longer on *sufficiency* and *efficiency* rather than on *necessity*.

1.11 Default assignment

Although both seeing and imagining result in assignments to frame terminals, imagination leaves us wider choices of detail and variety of such assignments. I conjecture that frames are never stored in long-term memory with unassigned terminal values. Instead, what really happens is that frames are stored with weakly bound default assignments at every terminal! These manifest themselves as often-useful but sometimes counterproductive stereotypes.

Thus if I say, "John kicked the ball", you probably cannot think of a purely abstract ball, but must imagine characteristics of a vaguely particular ball; it probably has a certain default size, default color, default weight. Perhaps it is a descendant of one you first owned or were injured by. Perhaps it resembles your latest one. In any case your image lacks the sharpness of presence because the processes that

* *Editor's note:* Section numbers have been retained from the original tech report, and hence are not always sequential in this abridged edition.

inspect and operate upon the weakly bound default features are very likely to change, adapt, or detach them.

Such default assignments would have subtle, idiosyncratic influences on the paths an individual would tend to follow in making analogies, generalizations, and judgements, especially when the exterior influences on such choices are weak. Properly chosen, such stereotypes could serve as a storehouse of valuable heuristic plan skeletons; badly selected, they could form paralyzing collections of irrational biases. Because of them, one might expect, as reported by Freud, to detect evidences of early cognitive structures in free-association thinking.

2 Language, understanding, and scenarios

2.1 Words, sentences, and meanings

> The device of images has several defects that are the price of its peculiar excellences. Two of these are perhaps the most important: the image, and particularly the visual image, is apt to go farther in the direction of the individualisation of situations than is biologically useful; and the principles of the combination of images have their own peculiarities and result in constructions which are relatively wild, jerky, and irregular, compared with the straightforward unwinding of a habit, or with the somewhat orderly march of thought. (Bartlett 1932/61)

The concepts of frame and default assignment seem helpful in discussing the phenomenology of "meaning". Chomsky (1957) points out that such a sentence as

(A) colorless green ideas sleep furiously

is treated very differently from the nonsentence

(B) furiously sleep ideas green colorless

and suggests that because both are "equally nonsensical", what is involved in the recognition of sentences must be quite different from what is involved in the appreciation of meanings.

There is no doubt that there are processes especially concerned with grammar. Since the meaning of an utterance is encoded as much in the positional and structural relations between the words as in the word choices themselves, there must be processes concerned with analyzing those relations in the course of building the structures that will more directly represent the meaning. What makes the words of (A) more effective and predictable than (B) in producing such a structure—putting aside the question of whether that structure should be called semantic or syntactic—is that the word-order relations in (A) exploit the (grammatical) conventions and rules people usually use to induce others to make assignments to terminals of structures. This is entirely consistent with theories of grammar. A generative grammar would be a summary description of the *exterior* appearance of those frame rules—or their associated processes—while the operators of transformational grammars seem similar enough to some of our frame transformations.

But one must also ask: to what degree does grammar have a separate identity in the actual working of a human mind? Perhaps the rejection of an utterance (either as ungrammatical, as nonsensical, or, most important, as *not understood*) indicates a more complex failure of the semantic process to arrive at any usable representation; I will argue now that the grammar/meaning distinction may illuminate two extremes of a continuum but obscures its all-important interior.

We certainly cannot assume that logical meaninglessness has a precise psychological counterpart. Sentence (A) can certainly generate an image! The dominate frame (in my case) is that of someone sleeping; the default system assigns a particular bed, and in it lies a mummy-like shape-frame with a translucent green color property. In this frame there is a terminal for the character of the sleep—restless, perhaps—and 'furiously' seems somewhat inappropriate at that terminal, perhaps because the terminal does not like to accept anything so "intentional" for a sleeper. 'Idea' is even more disturbing, because a person is expected, or a least something animate. I sense frustrated procedures trying to resolve these tensions and conflicts more properly, here or there, into the sleeping framework that has been evoked.

Utterance (B) does not get nearly so far because no subframe accepts any substantial fragment. As a result, no larger frame finds anything to match its terminals, hence, finally, no top level "meaning" or "sentence" frame can organize the utterance as either meaningful or grammatical. By combining this "soft" theory with gradations of assignment tolerances, I imagine one could develop systems that degrade properly for sentences with poor grammar rather than none; if the smaller fragments—phrases and subclauses—satisfy subframes well enough, an image adequate for certain kinds of comprehension could be constructed anyway, even though some parts of the top level structure are not entirely satisfied. Thus we arrive at a qualitative theory of "grammatical": *if the top levels are satisfied but some lower terminals are*

not, we have a meaningless sentence; if the top is weak but the bottom solid, we can have an ungrammatical but meaningful utterance.

I do not mean to suggest that sentences must evoke visual images. Some people do not admit to assigning a color to the ball in "He kicked the ball." But everyone admits (eventually) to having assumed, if not a size or color, at least some purpose, attitude, or other elements of an assumed scenario. When we go beyond vision, terminals and their default assignments can represent purposes and functions, not just colors, sizes, and shapes.

2.6 Scenarios

> Thinking ... is biologically subsequent to the image-forming process. It is possible only when a way has been found of breaking up the "massed" influence of past stimuli and situations, only when a device has already been discovered for conquering the sequential tyranny of past reactions. But though it is a later and a higher development, it does not supercede the method of images. It has its own drawbacks. Contrasted with imagining it loses something of vivacity, of vividness, of variety. Its prevailing instruments are words, and, not only because these are social, but also because in use they are necessarily strung out in sequence, they drop into habit reactions even more readily than images do. [With thinking] we run greater and greater risk of being caught up in generalities that may have little to do with actual concrete experience. If we fail to maintain the methods of thinking, we run the risks of becoming tied to individual instances and of being made sport of by the accidental circumstances belonging to these. (Bartlett 1932/61)

We condense and conventionalize, in language and thought, complex situations and sequences into compact words and symbols. Some words can perhaps be "defined" in elegant, simple structures, but only a small part of the meaning of "trade" is captured by:

$$\overset{\text{(first frame)}}{\boxed{\text{A has X}\quad\text{B has Y}}} \longrightarrow \overset{\text{(second frame)}}{\boxed{\text{B has X}\quad\text{A has Y}}}$$

Trading normally occurs in a social context of law, trust, and convention. Unless we also represent these other facts, most trade transactions will be almost meaningless. It is usually essential to know that each party usually wants both things but has to compromise. It is a happy but unusual circumstance in which each trader is glad to get rid of what he has. To represent trading strategies, one could insert the basic maneuvers right into the above frame-pair scenario: in order for A to make B want X more (or want Y less) we expect him to select one of the familiar tactics:

- Offer more for Y.
- Explain why X is so good.
- Create favorable side-effect of B having X.
- Disparage the competition.
- Make B think C wants X.

These only scratch the surface. Trades usually occur within a scenario tied together by more than a simple chain of events each linked to the next. No single such scenario will do; when a clue about trading appears, it is essential to guess which of the different available scenarios is most likely to be useful.

Charniak's thesis (1972) studies questions about transactions that seem easy for people to comprehend yet obviously need rich default structures. We find in elementary school reading books such stories as:

> Jane was invited to Jack's birthday party.
> She wondered if he would like a kite.
> She went to her room and shook her piggy bank.
> It made no sound.

Most young readers understand that Jane wants money to buy Jack a kite for a present but that there is no money to pay for it in her piggy bank. Charniak proposes a variety of ways to facilitate such inferences—a "demon" for 'present' that looks for things concerned with money, a demon for 'piggy bank' which knows that shaking without sound means the bank is empty, and so on. But although 'present' now activates 'money', the reader may be surprised to find that neither of those words (nor any of their synonyms) occurs in the story. 'Present' is certainly associated with 'party' and 'money' with 'bank', but how are the longer chains built up? Here is another problem raised by Charniak. A friend tells Jane:

> He already has a Kite.
> He will make you take it back.

Take *which* kite back? We do not want Jane to return Jack's old kite. To determine the referent of the pronoun 'it' requires understanding a lot about an assumed scenario. Clearly, 'it' refers to the proposed *new* kite. How does one know this? (Note that we need not agree on any single explanation.) Generally, pronouns refer to recently mentioned things, but as this example shows, the referent depends on more than the local syntax.

Suppose for the moment we are already trying to instantiate a "buying-a-present" default subframe. Now, the word 'it' alone is too small a fragment to deal with, but 'take it back' could be a plausible unit to match a terminal of an appropriately elaborate 'buying' scenario. Since that terminal would be constrained to agree with the assignment of 'present' itself, we are assured of the correct meaning of 'it' in 'take X back'. Automatically, the correct kite is selected. Of course, that terminal will have its own constraints as well; a subframe for the 'take-it-back' idiom should know that 'take X back' requires that:

- X was recently purchased.
- The return is to the place of purchase.
- You must have your sales slip.

And so on.

If the current scenario does not contain a 'take-it-back' terminal, then we have to find one that does and substitute it, maintaining as many prior assignments as possible. Notice that if things go well, the question of it being the old kite never even arises. *The sense of ambiguity arises only when a "near miss" mismatch is tried and rejected.*

Charniak's proposed solution to this problem is in the same spirit but emphasizes understanding that, because Jack already has a kite, he may not want another one. He proposes a mechanism associated with 'present':

(A) If we see that person P might not like a present X, then look for X being returned to the store where it was bought.

(B) If we see this happening, or even being suggested, assert that the reason why is that P does not like X.

This statement of "advice" is intended by Charniak to be realized as a production-like entity, to be added to the currently active data base whenever a certain kind of context is encountered. Later, if its antecedent condition is satisfied, its action adds enough information about Jack and about the new kite to lead to a correct decision about the pronoun.

Charniak in effect proposes that the system should watch for certain kinds of events or situations and inject proposed reasons, motives, and explanations for them. The additional interconnections between the story elements are expected to help bridge the gaps that logic might find it hard to cross, because the additions are only "plausible" default explanations, assumed without corroborative assertions. By assuming (tentatively) "does not like X" when X is taken back, Charniak hopes to simulate much of ordinary "comprehension" of what is happening. We do not yet know how complex and various such plausible inferences must be to get a given level of performance, and the thesis does not answer this because it did not include a large simulation. Usually he proposes terminating the process by asserting the allegedly plausible motive without further analysis unless necessary. To understand why Jack might return the additional kite, it should usually be enough to assert that he does not like it. A deeper analysis might reveal that Jack would not really mind having two kites but he probably realizes that he will get only one present; his utility for two different presents is probably higher.

2.7 Scenarios and "questions"

The meaning of a child's birthday party is very poorly approximated by any dictionary definition like "a party assembled to celebrate a birthday", where a party would be defined, in turn, as "people assembled for a celebration". This lacks all the flavor of the culturally required activities. Children know that the "definition" should include more specifications, the particulars of which can normally be assumed by way of default assignments:

Dress	Sunday best.
Present	Must please host. Must be bought and gift wrapped.
Games	Hide and seek; pin tail on donkey.
Decor	Balloons, favors, crepe-paper.
Party-meal	Cake, ice cream, soda, hot dogs.
Cake	Candles; blow out; wish; sing birthday song.
Ice-cream	Standard three-flavor.

These ingredients for a typical American birthday party must be set into a larger structure. Extended events take place in one or more days.

A Party takes place in a day, of course, and occupies a substantial part of it, so we locate it in an appropriate Day frame. A typical day has main events, such as:

Get-up Dress Eat-1 Go-to-Work Eat-2

but a School-Day has more fixed detail:

Get-up Dress
Eat-1 Go-to-School Be-in-School
Home-Room Assembly English Math (arrgh)
Eat-2 Science Recess Sport
Go-Home Play
Eat-3 Homework
Go-To-Bed

Birthday parties obviously do not fit well into school-day frames. Any parent knows that the Party-Meal is bound to Eat-2 of its Day. I remember a child who did not seem to realize this. Absolutely stuffed after the Party-Meal, he asked when he would get Lunch.

Returning to Jane's problem with the kite, we first hear that she is invited to Jack's birthday party. Without this party scenario, or at least an invitation scenario, the second line seems rather mysterious:

She wondered if he would like a kite.

To explain one's rapid comprehension of this, I will make a somewhat radical proposal: *to represent explicitly, in the frame for a scenario structure, pointers to a collection of the most serious problems and questions commonly associated with it.* In fact, we shall consider the idea that the frame terminals are exactly those questions. Thus, for the birthday party:

Y must get P for X	Choose P!
X must like P	Will X like P?
Buy P	Where to buy P?
Get money to buy P	Where to get money? (Sub-question of the Buy frame?)
Y must dress up	What should Y wear?

Certainly these are one's first concerns, when one is invited to a party.

The reader is free to wonder, with the author, whether this solution is acceptable. The question, "Will X like P?" certainly matches "She wondered if he would like a kite?" and correctly assigns the kite to P.

But is our world regular enough that such question sets could be precompiled to make this mechanism often work smoothly? I think the answer is mixed. We do indeed expect many such questions; we surely do not expect all of them. But surely "expertise" consists partly in not having to realize *ab initio* what are the outstanding problems and interactions in situations. Notice, for example, that there is *no* default assignment for the Present in our party-scenario frame. This mandates attention to that assignment problem and prepares us for a possible thematic concern. In any case, we probably need a more active mechanism for understanding *"wondered"* which can apply the information currently in the frame to produce an expectation of what Jane will think about.

The third line of our story, about shaking the bank, should also eventually match one of the present-frame questions, but the unstated connection between Money and Piggy-Bank is presumably represented in the piggy-bank frame, *not* the party frame, although once it is found, it will match our Get-Money question terminal. The primary functions and actions associated with piggy banks are Saving and Getting-Money-Out, and the latter has three principal methods:

(1) Using a key. (Most piggy banks don't offer this option.)

(2) Breaking it. (Children hate this.)

(3) Shaking the money out, or using a thin slider.

In the fourth line, does one know specifically that a *silent* Piggy Bank is empty, and hence out of money (I think, yes), or does one use general knowledge that a hard container which makes no noise when shaken is empty? I have found quite a number of people who prefer the latter. Logically, the "general principle" would indeed suffice, but I feel that this misses the important point that a specific scenario of this character is engraved in every child's memory. The story is instantly intelligible to most readers. If more complex reasoning from general principles were required, this would not be so, and more readers would surely go astray. It is easy to find more complex problems.

> A goat wandered into the yard where Jack was painting. The goat got the paint all over himself. When Jack's mother saw the goat, she asked, "Jack, did you do *that*?"

There is no one word or line, which is the referent of "that". It seems to refer, as Charniak notes, to "cause the goat to be covered with paint". Charniak does not permit himself to make a specific proposal to handle this kind of problem, remarking only that his "demon" model would need a substantial extension to deal with such a poorly localized "thematic subject". Consider how much one has to know about our culture, to realize that *that* is not the *goat-in-the-yard* but the *goat-covered-with-paint.* Charniak's thesis—basically a study rather than a debugged system—discusses issues about the activation, operation, and dismissal of expectation and default-knowledge demons. Many of his ideas have been absorbed into this essay.

In spite of its tentative character, I will try to summarize this image of language understanding as somewhat parallel to seeing. The key words and ideas of a discourse evoke substantial thematic or scenario structures, drawn from memory with rich default assumptions. The individual statements of a discourse lead to temporary representations—which seem to correspond to what contemporary linguists call "deep structures"—which are then quickly rearranged or consumed in elaborating the growing scenario representation. In order of "scale", among the ingredients of such a structure there might be these kinds of levels:

SURFACE SYNTACTIC FRAMES. Mainly verb and noun structures. Prepositional and word-order indicator conventions.

SURFACE SEMANTIC FRAMES. Action-centered meanings of words. Qualifiers and relations concerning participants, instruments, trajectories and strategies, goals, consequences, and side-effects.

THEMATIC FRAMES. Scenarios concerned with topics, activities, portraits, setting. Outstanding problems and strategies commonly connected with topic.

NARRATIVE FRAMES. Skeleton forms for typical stories, explanations, and arguments. Conventions about foci, protagonists, plot forms, development, and so on, designed to help a listener construct a new, instantiated thematic frame in his own mind.

A single sentence can assign terminals, attach subframes, apply a transformation, or cause a gross replacement of a high-level frame when a proposed assignment no longer fits well enough. A pronoun is comprehensible only when general linguistic conventions, interacting with defaults and specific indicators, determine a terminal or subframe of the current scenario.

In *vision* the transformations usually have a simple grouplike structure; in *language* we expect more complex, less regular systems of frames. Nevertheless, because *time, cause,* and *action* are so important to us, we often use sequential transformation pairs that replace situations by their temporal or causal successors.

Because syntactic structural rules direct the selection and assembly of the transient sentence frames, research on linguistic structures should help us understand how our frame systems are constructed. One might look for such structures specifically associated with assigning terminals, selecting emphasis or attention viewpoints (transformation), inserting sentential structures into thematic structures, and changing gross thematic representations.

Finally, just as there are familiar "basic plots" for stories, there must be basic superframes for discourses, arguments, narratives, and so forth. As with sentences, we should expect to find special linguistic indicators for operations concerning these larger structures; we should move beyond the grammar of sentences to try to find and systematize the linguistic conventions that, operating across wider spans, must be involved with assembling and transforming scenarios and plans.

2.8 Questions, systems, and cases

> Questions arise from a point of view—from something that helps to structure what is problematical, what is worth asking, and what constitutes an answer (or progress). It is not that the view determines reality, only what we accept from reality and how we structure it. I am realist enough to believe that in the long run reality gets its own chance to accept or reject our various views.
>
> (Newell 1973a)

Examination of linguistic discourse leads thus to a view of the frame concept in which the "terminals" serve to represent the questions most likely to arise in a situation. To make this important viewpoint more explicit, we will spell out this reinterpretation.

> A *frame* is a collection of questions to be asked about a hypothetical situation: it specifies issues to be raised and methods to be used in dealing with them.

The terminals of a frame correspond perhaps to what Schank (1973) calls "conceptual cases", although I do not think we should restrict them to as few types as Schank suggests. To understand a narrated or perceived action, one often feels compelled to ask such questions as

- What caused it (agent)?
- What was the purpose (intention)?

- What are the consequences (side-effects)?
- Whom does it affect (recipient)?
- How is it done (instrument)?

The number of such "cases" or questions is problematical. While we would like to reduce meaning to a very few "primitive" concepts, perhaps in analogy to the situation in traditional linguistic analysis, I know of no reason to suppose that that goal can be achieved. My own inclination is to side with such workers as Martin (1974) who look toward very large collections of "primitives", annotated with comments about how they are related. Only time will tell which is better.

For entities other than actions, one asks different questions; for thematic topics the questions may be much less localized, for instance:

- Why are they telling this to me?
- How can I find out more about it?
- How will it help with the "real problem"?

In a "story" one asks what is the topic, what is the author's attitude, what is the main event, who are the protagonists, and so on. As each question is given a tentative answer, the corresponding subframes are attached and the questions they ask become active in turn.

The "markers" we proposed for vision-frames become more complex in this view. If we adopt for the moment Newell's larger sense of "view", it is not enough simply to ask a question; one must indicate how it is to be answered. Thus a terminal should also contain (or point to) suggestions and recommendations about how to find an assignment. Our "default" assignments then become the simplest special cases of such recommendations, and one certainly could have a hierarchy in which such proposals depend on features of the situation, perhaps along the lines of Wilks's (1973) "preference" structures.

For syntactic frames, the drive toward ritualistic completion of assignments is strong, but we are more flexible at the conceptual level. As Schank (1973) says:

> People do not usually state all the parts of a given thought that they are trying to communicate because the speaker tries to be brief and leaves out assumed or inessential information ... The conceptual processor makes use of the unfilled slots to search for a given type of information in a sentence or a larger unit of discourse that will fill the needed slot.

Even in physical perception we have the same situation. A box will not present all of its sides at once to an observer, and, although this is certainly not because it wants to be brief, the effect is the same; the processor is prepared to find out what the missing sides look like and (if the matter is urgent enough) to move around to find answers to such questions.

Frame *systems*, in this view, become choice points corresponding (on the conceptual level) to the mutually exclusive choice "systems" exploited by Winograd (1970). The different frames of a system represent different ways of using the same information, located at the common terminals. As in the grammatical situation, one has to choose one of them at a time. On the conceptual level this choice becomes: *what questions shall I ask about this situation?*

View changing, as we shall argue, is a problem-solving technique important in representing, explaining, and predicting. In the rearrangements inherent in the frame-system representation (for example, of an action), we have a first approximation to Simmons's (1973) idea of "procedures which in some cases will change the contextual definitional structure to reflect the action of a verb".

Where do the "questions" come from? This is not in the scope of this paper, really, but we can be sure that the frame makers (however they operate) must use some principles. The methods used to generate the questions ultimately shape each person's general intellectual style. People surely differ in details of preferences for asking "Why?", "How can I find out more?", "What's in it for me?", "How will this help with the current higher goals?", and so forth.

Similar issues about the style of *answering* must arise. In its simplest form, the drive toward instantiating empty terminals would appear as a variety of hunger or discomfort, satisfied by any default or other assignment that does not conflict with a prohibition. In more complex cases we should perceive less animalistic strategies for acquiring deeper understandings.

It is tempting, then, to imagine varieties of frame systems that span from simple template-filling structures to implementations of the "views" of Newell—with all their implications about coherent generators of issues with which to be concerned, ways to investigate them, and procedures for evaluating proposed solutions. But I feel uncomfortable about any superficially coherent synthesis in which one expects the same kind of theoretical framework to function well on many different levels of scale or concept. We should expect very different question-processing mechanisms to operate our low-level stereotypes and our most comprehensive strategic overviews.

3 Learning, memory, and paradigms

> To the child, nature gives various means of rectifying any mistakes he may commit respecting the salutary or hurtful qualities of the objects which surround him. On every occasion his judgements are corrected by experience; want and pain are the necessary consequences arising from false judgement; gratification and pleasure are produced by judging aright. Under such masters, we cannot fail but to become well informed; and we soon learn to reason justly, when want and pain are the necessary consequences of a contrary conduct.
>
> In the study and practice of the sciences it is quite different: the false judgments we form neither affect our existence nor our welfare; and we are not forced by any physical necessity to correct them. Imagination, on the contrary, which is ever wandering beyond the bounds of truth, joined to self-love and that self-confidence we are so apt to indulge, prompt us to draw conclusions that are not immediately derived from facts. (Lavoisier 1789/1949)

How does one locate a frame to represent a new situation? Obviously, we cannot begin any complete theory outside the context of some proposed global scheme for the organization of knowledge in general. But if we imagine working within some bounded domain, we can discuss some important issues:

Expectation: How to select an initial frame to meet some given conditions.

Elaboration: How to select and assign subframes to represent additional details.

Alteration: How to find a frame to replace one that does not fit well enough.

Novelty: What to do if no acceptable frame can be found. Can we modify an old frame or must we build a new one?

Learning: What frames should be stored, or modified, as result of the experience?

In popular culture, memory is seen as separate from the rest of thinking; but finding the right memory—it would be better to say: finding a *useful* memory—needs the same sorts of strategies used in other kinds of thinking!

We say someone is "clever" who is unusually good at quickly locating highly appropriate frames. His information-retrieval systems are better at making good hypotheses, formulating the conditions the new frame should meet, and exploiting knowledge gained in the "unsuccessful" part of the search. Finding the right memory is no less a problem than solving any other kind of puzzle! Because of this, a good retrieval mechanism can be based only in part upon basic "innate" mechanisms. It must also depend largely on (learned) knowledge about the structure of one's own knowledge! Our proposal will combine several elements—a pattern-matching process, a clustering theory, and a similarity network.

In seeing a room or understanding a story, one assembles a network of frames and subframes. Everything noticed or guessed, rightly or wrongly, is represented in this network. We have already suggested that an active frame cannot be maintained unless its terminal conditions are satisfied.

We now add the postulate that *all satisfied frames must be assigned to terminals of superior frames.* This applies, as a special case, to any substantial fragments of "data" that have been observed and represented.

Of course, there must be an exception! We must allow a certain number of items to be attached to something like a set of "short term memory" registers. But the intention is that very little can be remembered unless embedded in a suitable frame. This, at any rate, is the conceptual scheme; in certain domains we would, of course, admit other kinds of memory "hooks" and special sensory buffers.

3.1 Requests to memory

We can now imagine the memory system as driven by two complementary needs. *On one side are items demanding to be properly represented by being embedded into larger frames; on the other side are incompletely filled frames demanding terminal assignments.* The rest of the system will try to placate these lobbyists, but not so much in accord with general principles as in accord with special knowledge and conditions imposed by the currently active goals.

When a frame encounters trouble—when an important condition cannot be satisfied—something must be done. We envision the following major kinds of accommodation to trouble:

MATCHING: When nothing more specific is found, we can attempt to use some "basic" associative memory mechanism. This will succeed by itself only in relatively simple situations, but should play a supporting role in the other tactics.

EXCUSE: An apparent misfit can often be excused or explained. A "chair" that meets all other conditions but is much too small could be a "toy".

ADVICE: The frame contains explicit knowledge about what to do about the trouble. Below, we describe an extensive, learned, "similarity network" in which to embed such knowledge.

SUMMARY: If a frame cannot be completed or replaced, one must give it up. But first one must construct a well-formulated complaint or summary to help whatever process next becomes responsible for reassigning the subframes left in limbo.

In my view, all four of these are vitally important. I discuss them in the following sections.

3.3 Excuses

We can think of a frame as describing an "ideal". If an ideal does not match reality because it is "basically" wrong, it must be replaced. *But it is in the nature of ideals that they are really elegant simplifications; their attractiveness derives from their simplicity, but their real power depends upon additional knowledge about interactions between them!* Accordingly we need not abandon an ideal because of a failure to instantiate it, provided one can explain the discrepancy in terms of such an interaction. Here are some examples in which such an "excuse" can save a failing match:

OCCLUSION: A table, in a certain view, should have four legs, but a chair might occlude one of them. One can look for things like T-joints and shadows to support such an excuse.

FUNCTIONAL VARIANT: A chair-leg is usually a stick, geometrically; but more important, it is functionally a support. Therefore, a strong center post, with an adequate base plate, should be an acceptable replacement for all the legs. Many objects are multiple purpose and need functional rather than physical descriptions.

BROKEN: A visually missing component could be explained as in fact physically missing, or it could be broken. Reality has a variety of way to frustrate ideals.

> PARASITIC CONTEXTS: An object that is just like a chair, except in size, could be (and probably is) a toy chair. The complaint "too small" could often be so interpreted in contexts with other things too small, children playing, peculiarly large "grain", and so forth.

In most of these examples, the kinds of knowledge to make the repair—and thus salvage the current frame—are "general" enough usually to be attached to the thematic context of a superior frame. In the remainder of this essay, I will concentrate on types of more sharply localized knowledge that would naturally be attached to a frame itself, for recommending its own replacement.

3.5 Clusters, classes, and a geographic analogy

> Though a discussion of *some* of the attributes shared by a *number* of games or chairs or leaves often helps us to learn how to employ the corresponding term, there is no set of characteristics that is simultaneously applicable to all members of the class and to them alone. Instead, confronted with a previously unobserved activity, we apply the term 'game' because what we are seeing bears a close "family resemblance" to a number of the activities we have previously learned to call by that name. For Wittgenstein, in short, games, chairs, and leaves are natural families, each constituted by a network of overlapping and crisscross resemblances. The existence of such a network sufficiently accounts for our success in identifying the corresponding object or activity. (Kuhn 1962/70, p. 45)

To make the similarity network act more "complete", consider the following analogy. In a city, any person should be able to visit any other; but we do not build a special road between each pair of houses; we place a group of houses on a "block". We do not connect roads between each pair of blocks, but have them share streets. We do not connect each town to every other, but construct main routes, connecting the centers of larger groups. Within such an organization, each member has direct links to some other individuals at its own "level", mainly to nearby, highly similar ones; but each individual has also at least a few links to "distinguished" members of higher-level groups. The result is that there is usually a rather short sequence between any two individuals, if one can but find it.

To locate something in such a structure, one uses a hierarchy like the one implicit in a mail address. Everyone knows something about the largest categories, in that he knows where the major cities are. An inhabitant of a city knows the nearby towns, and people in the towns know the nearby villages. No person knows all the individual routes between pairs of houses; but, for a particular friend, one may know a special route to his home in a nearby town that is better than going to the city and back. *Directories* factor the problem, basing paths on standard routes between major nodes in the network. Personal shortcuts can bypass major nodes and go straight between familiar locations. Although the standard routes are usually not quite the very best possible, our stratified transport and communication services connect everything together reasonably well, with comparatively few connections.

At each level, the aggregates usually have distinguished foci or *capitals.* These serve as elements for clustering at the next level of aggregation. There is no nonstop airplane service between New Haven and San Jose because it is more efficient overall to share the trunk route between New York and San Francisco, which are the capitals at that level of aggregation.

As our memory networks grow, we can expect similar aggregations of the destinations of our similarity pointers. Our decisions about what we consider to be primary or trunk difference features and which are considered subsidiary will have large effects on our abilities. Such decisions eventually accumulate to become epistemological commitments about the conceptual cities of our mental universe.

The nonrandom convergences and divergences of the similarity pointers, for each difference *d,* thus tend to structure our conceptual world around

- the aggregation into *d*-clusters, and
- the selection of *d*-capitals.

Note that it is perfectly all right to have *several capitals in a cluster,* so that there need be no one attribute common to them all. The "criss-cross resemblances" of Wittgenstein are then consequences of the local connections in our similarity network, which are surely adequate to explain how we can feel as though we know what a chair or a game is—yet cannot always define it in a logical way as an element in some class-hierarchy or by any other kind of compact, formal, declarative rule. The apparent coherence of the conceptual aggregates need not reflect explicit definitions, but can emerge from the success-directed sharpening of the difference-describing processes.

The selection of capitals corresponds to selecting the stereotypes or typical elements whose default assignments are unusually useful. There are many forms of chairs, for example, and one should choose carefully the chair-description frames that are to be the major capitals of chair-land. These are used for rapid matching and assigning priorities to the various differences. The lower-priority features of the cluster center then serve either as default properties of the chair types or, if more realism is required, as dispatch pointers to the local chair villages and towns. Difference points could be "functional" as well as geometric. Thus, after rejecting a first try at "chair", one might try the functional idea of "something one can sit on" to explain an unconventional form. This requires a deeper analysis in terms of forces and strengths. Of course, that analysis would fail to capture toy chairs, or chairs of such ornamental delicacy that their actual use would be unthinkable. These would be better handled by the method of excuses, in which one would bypass the usual geometrical or functional explanations in favor of responding to contexts involving art or play.

It is important to reemphasize that there is no reason to restrict the memory structure to a single hierarchy; the notions of "level" of aggregation need not coincide for different kinds of differences. The *d*-capitals can exist, not only by explicit declarations, but also implicitly by their focal locations in the structure defined by convergent *d*-pointers. (In Newell and Simon's GPS framework, the "differences" are ordered into a fixed hierarchy. By making the priorities depend on the goal, the same memories could be made to serve more purposes; the resulting problem solver would lose the elegance of a single, simple-ordered measure of "progress", but that is the price of moving from a first-order theory.)

Finally, we should point out that we do not need to invoke any mysterious additional mechanism for *creating* the clustering structure. Developmentally, one would assume, the earliest frames would tend to become the capitals of their later relatives, unless this is firmly prevented by experience, because each time the use of one stereotype is reasonably successful, its centrality is reinforced by another pointer from somewhere else. Otherwise, *the acquisition of new centers is in large measure forced upon us from the outside: by the words available in our language; by the behavior of objects in our environment; by what we are told by our teachers, family, and general culture.* Of course, at each step the structure of the previous structure dominates the acquisition of the later. But in any case such forms and clusters should emerge from the interactions between the world and almost any memory-using mechanism; it would require more explanation were they *not* found!

3.6 Analogies and alternative descriptions

We have discussed the use of different frames of the same system to describe the same situation in different ways: for change of position in vision and for change of emphasis in language. Sometimes, in "problem solving", we use two or more descriptions in a more complex way to construct an analogy or to apply two radically *different* kinds of analysis to the same situation. *For hard problems, one "problem space" is usually not enough!*

Suppose your car battery runs down. You believe that there is an electricity shortage and blame the generator.

The generator can be represented as a mechanical system: the rotor has a pulley wheel driven by a belt from the engine. Is the belt tight enough? Is it even there? The output, seen mechanically, is a cable to the battery or whatever. Is it intact? Are the bolts tight? Are the brushes pressing on the commutator?

Seen electrically, the generator is described differently. The rotor is seen as a flux-linking coil, rather than as a rotating device. The brushes and commutator are seen as electrical switches. The output is current along a pair of conductors leading from the brushes through control circuits to the battery.

We thus represent the situation in two quite different frame systems. In one, the armature is a mechanical rotor with pulley; in the other, it is a conductor in a changing magnetic field. The same—or analogous—elements share terminals of different frames, and the frame transformations apply only to some of them.

The differences between the two frames are substantial. The entire mechanical chassis of the car plays the simple role, in the electrical frame, of one of the battery connections. The diagnostician has to use both representations. A failure of current to flow often means that an intended conductor is not acting like one. For this case, the basic transformation between the frames depends on the fact that electrical continuity is in general equivalent to firm mechanical attachment. Therefore, any conduction disparity revealed by electrical measurements should make us look for a corresponding disparity in the mechanical frame. In fact, since "repair" in this universe is synonymous with "mechanical repair", the diagnosis *must* end in the mechanical frame. Eventually, we might locate a defective mechanical junction and discover a loose connection, corrosion, wear, or whatever.

Why have two separate frames, rather than one integrated structure to represent the generator? I believe that in such a complex problem, one can never cope with many details at once. At each moment one must work within a reasonably simple framework. I contend that any problem that a person can solve at all is worked out at each moment in a small context and that the key operations in problem solving are concerned with finding or constructing these working environments.

Indeed, finding an electrical fault requires moving between at least three frames: a visual one along with the electrical and mechanical frames. If electrical evidence suggests a loose mechanical connection, one needs a visual frame to guide oneself to the mechanical fault.

Are there general methods for constructing adequate frames? The answer is both yes and no! There are some often-useful strategies for adapting old frames to new purposes; but I should emphasize that humans certainly have no magical way to solve *all* hard problems! One must not fall into what Papert calls the superhuman-human fallacy and require a theory of human behavior to explain even things that people cannot really do!

One cannot expect to have a frame exactly right for
expect always to be able to invent one. But we do have a good deal to work with, and it is important to remember the contribution of one's culture in assessing the complexity of problems people seem to solve. *The experienced mechanic need not routinely invent*; he already has engine representations in terms of ignition, lubrication, cooling, timing, fuel mixing, transmission, compression, and so forth. Cooling, for example, is already subdivided into fluid circulation, air flow, thermostasis, and the like. Most "ordinary" problems are presumably solved by systematic use of the analogies provided by the transformations between pairs of these structures. The huge network of knowledge, acquired from school, books, apprenticeship, or whatever, is interlinked by difference and relevancy pointers. No doubt, the culture imparts a good deal of this structure by its conventional *use of the same words* in explanations of different views of a subject.

3.8 Frames and paradigms

> Until that scholastic paradigm [the medieval 'impetus' theory] was invented, there were no pendulums, but only swinging stones, for

> scientists to see. Pendulums were brought into the world by something very like a paradigm-induced gestalt switch.
>
> Do we, however, really need to describe what separates Galileo from Aristotle, or Lavoisier from Priestly, as a transformation of vision? Did these men really *see* different things when *looking at* the same sorts of objects? Is there any legitimate sense in which we can say they pursued their research in different worlds? ...
>
> I am ... acutely aware of the difficulties created by saying that when Aristotle and Galileo looked at swinging stones, the first saw constrained fall, the second a pendulum. ... Nevertheless, I am convinced that we must learn to make sense of sentences that at least resemble these. (Kuhn 1962/70, pp. 120-121)

According to Kuhn's model of scientific evolution, normal science proceeds by using established descriptive schemes. Major changes result from new paradigms, new ways of describing things that lead to new methods and techniques. Eventually there is a redefining of "normal".

Now while Kuhn prefers to apply his own very effective redescription paradigm at the level of major scientific revolutions, it seems to me that the same idea applies as well to the microcosm of everyday thinking. Indeed, in the above quotation, we see that Kuhn is seriously considering that the paradigms play a substantive rather than metaphorical role in visual perception, just as we have proposed for frames.

Whenever our customary viewpoints do not work well, whenever we fail to find effective frame systems in memory, we must construct new ones that bring out the right features. Presumably, the most usual way to do this is to build some sort of pair-system from two or more old ones and then edit or debug it to suit the circumstances. How might this be done? It is tempting to formulate the requirements, and then solve the construction problem.

But that is certainly not the usual course of ordinary thinking! Neither are requirements formulated all at once, nor is the new system constructed entirely by deliberate preplanning. Instead we recognize unsatisfied requirements, one by one, as deficiencies or "bugs", in the course of a sequence of modifications made to an unsatisfactory representation.

I think Papert (1972; see also Minsky 1970) is correct in believing that the ability to diagnose and modify one's own procedures is a collection of specific and important "skills". *Debugging,* a fundamentally important component of intelligence, has its own special techniques and procedures. Every normal person is pretty good at them; or otherwise he would not have learned to see and talk! Although this essay is already speculative, I would like to point here to the theses of Goldstein (1974) and Sussman (1973/75) about the explicit use of ***knowledge about debugging*** in learning symbolic representations. They build new procedures to satisfy multiple requirements by such elementary but powerful techniques as:

(1) Make a crude first attempt by the first order method of simply putting together procedures that *separately* achieve the individual goals.

(2) If something goes wrong, try to characterize one of the defects as a *specific* (and undesirable) kind of interaction between two procedures.

(3) Apply a debugging technique that, according to a record in memory, is good at repairing that *specific kind* of interaction.

(4) Summarize the experience, to add to the "debugging techniques library" in memory.

These might seem simple minded, but if the new problem is not too radically different from the old ones, they have a good chance to work, especially if one picks out the right first-order approximations. If the new problem *is* radically different, one should not expect *any* learning theory to work well. Without a structured cognitive map—without the "near misses" of Winston or a cultural supply of good training sequences of problems, we should not expect radically new paradigms to appear magically whenever we need them.

What are "kinds of interactions", and what are "debugging techniques"? The simplest, perhaps, are those in which the result of achieving a first goal interferes with some condition prerequisite for achieving a second goal. The simplest repair is to reinsert the prerequisite as a new condition. There are examples in which this technique alone cannot succeed because a prerequisite for the second goal is incompatible with the first. Sussman presents a more sophisticated diagnosis and repair method that recognizes this and exchanges the order of the goals. Goldstein considers related problems in a multiple description context.

If asked about important future lines of research on artificial or natural intelligence, I would point to the interactions between these ideas and the problems of using multiple representations to deal with the same situation from several viewpoints. To carry out such a study, we need better ideas about interactions among the transformed relationships. Here the frame-system idea by itself begins to show limitations. Fitting together new representations from parts of old ones is clearly a complex process itself, and one that could be solved within the framework of our theory (if at all) only by an intricate bootstrapping. This, too, is surely a special skill with its own techniques. I consider it a crucial component of a theory of intelligence.

We must not expect complete success in the above enterprise; there is a difficulty, as Newell (1973) notes in a larger context:

> Elsewhere is another view—possibly from philosophy—or other "elsewheres" as well, since the views of man are multiple. Each view has its own questions. Separate views speak mostly past each other. Occasionally, of course, they speak to the same issue and then comparison is possible, but not often and not on demand.

Appendix: criticism of the logistic approach

> If one tries to describe processes of genuine thinking in terms of formal traditional logic, the result is often unsatisfactory; one has, then, a series of correct operations, but the sense of the process and what was vital, forceful, creative in it seems somehow to have evaporated in the formulations. (Wertheimer 1959)

I here explain why I think more "logical" approaches will not work. There have been serious attempts, from as far back as Aristotle, to represent common-sense reasoning by a "logistic" system—that is, one that makes a complete separation between

(1) "propositions" that embody specific information, and

(2) "syllogisms" or general laws of proper inference.

No one has been able successfully to confront such a system with a realistically large set of propositions. I think such attempts will continue to fail, because of the character of logistic in general rather than from defects of particular formalisms. (Most recent attempts have used variants of "first order predicate logic", but I do not think *that* is the problem.)

A typical attempt to simulate common-sense thinking by logistic systems begins in a micro-world of limited complication. At one end are high-level goals such as "I want to get from my house to the airport". At the other end we start with many small items—the *axioms*—

like “The car is in the garage”, “One does not go outside undressed”, “To get to a place one should (on the whole) move in its direction”, and so on. To make the system work, one designs heuristic search procedures to “prove” the desired goal, or to produce a list of actions that will achieve it.

I will not recount the history of attempts to make both ends meet—but merely summarize my impression: in simple cases, one can get such systems to “perform”, but as we approach reality, the obstacles become overwhelming. The problem of finding suitable axioms—the problem of “stating the facts” in terms of always-correct, logical assumptions—is very much harder than is generally believed.

Formalizing the required knowledge: Just constructing a knowledge base is a major intellectual research problem. Whether one's goal is logistic or not, we still know far too little about the contents and structure of common-sense knowledge. A “minimal” common-sense system must “know” something about cause and effect, time, purpose, locality, process, and types of knowledge. It also needs ways to acquire, represent, and use such knowledge. We need a serious epistemological research effort in this area. The essays of McCarthy (1969) and Sandewall (1970) are steps in that direction. I have no easy plan for this large enterprise; but the magnitude of the task will certainly depend strongly on the representations chosen, and I think that “logistic” is already making trouble.

Relevance: The problem of selecting relevance from excessive variety is a key issue! A modern epistemology will not resemble the old ones! Computational concepts are necessary and novel. Perhaps the better part of knowledge is not propositional in character, but interpropositional. For each “fact” one needs meta-facts about how it is to be used and when it should not be used. In McCarthy's “Airport” paradigm we see ways to deal with some interactions between “situations, actions, and causal laws” within a restricted micro-world of things and actions. But though the system can make deductions implied by its axioms, it cannot be told when it should or should not make such deductions.

For example, one might want to tell the system to “not cross the road if a car is coming”. But one cannot demand that the system “prove” no car is coming, for there will not usually be any such proof. In PLANNER, one can direct an *attempt* to prove that a car *is* coming, and if the (limited) deduction attempt ends with “failure”, one can act. This cannot be done in a pure logistic system. “Look right, look left” is a first approximation. But if one tells the system the real truth about speeds, blind driveways, probabilities of racing cars whipping around the corner, proof becomes impractical. If it reads in a physics book that intense fields perturb light rays, should it fear that a mad scientist has built an invisible car? We need to represent "usually"! Eventually it must understand the trade-off between mortality and accomplishment, for one can do nothing if paralyzed by fear.

Monotonicity: Even if we formulate relevance restrictions, logistic systems have a problem in using them. In any logistic system, all the axioms are necessarily "permissive"—they all help to permit new inferences to be drawn. Each added axiom means more theorems; none can disappear. There simply is no direct way to add information to tell such a system about kinds of conclusions that should *not* be drawn! To put it simply: if we adopt enough axioms to deduce what we need, we deduce far too many other things. But if we try to change this by adding axioms about relevance, we still produce all the unwanted theorems, plus annoying statements about their irrelevance.

Because logicians are not concerned with systems that will later be enlarged, they can design axioms that permit only the conclusions they want. In the development of intelligence the situation is different. One has to learn which features of situations are important and which kinds of deductions are not to be regarded seriously. The usual reaction to the "liar's paradox" is, after a while, to laugh. The conclusion is not to reject an axiom, but to reject the deduction itself! This raises another issue.

Procedure-controlling knowledge: The separation between axioms and deduction makes it impractical to include classificational knowledge about propositions. Nor can we include knowledge about management of deduction. A paradigm problem is that of axiomatizing everyday concepts of approximation or nearness. One would like nearness to be transitive:

$$(\text{A near B}) \text{ and } (\text{B near C}) \rightarrow (\text{A near C})$$

but unrestricted application of this rule would make everything near everything else. One can try technical tricks like

$$(\text{A near}_1\text{ B}) \text{ and } (\text{B near}_1\text{ C}) \rightarrow (\text{A near}_2\text{ C})$$

and admit only (say) five grades: near_1, near_2, ... near_5. One might invent analog quantities or parameters. But one cannot (in a logistic system) decide to make a new kind of "axiom" to prevent applying transitivity after (say) three chained uses, conditionally, unless there is a "good excuse". I do not mean to propose a particular solution to the transitivity of nearness. (To my knowledge, no one has made a creditable proposal about it.) My complaint is that, because of acceptance of logistic, no one has freely explored this kind of procedural restriction.

Combinatorial problems: A human thinker reviews plans and goal lists as he works, revising his knowledge and policies about using them. One can program some of this into the theorem-proving program itself; but one really wants also to represent it directly, in a natural way, in the declarative corpus—for use in further introspection. Why then do workers try to make logistic systems do the job? A valid reason is that the systems have an attractive simple elegance; if they worked, this would be fine. An invalid reason is more often offered: that such systems have a mathematical virtue because they are:

(1) Complete: All true statements can be proven; and

(2) Consistent: No false statements can be proven.

It seems not often realized that completeness is no rare prize. It is a trivial consequence of any exhaustive search procedure, and any system can be "completed" by adjoining to it any other complete system and interlacing the computational steps. Consistency is more refined; it requires one's axioms to imply no contradictions. But I do not believe that consistency is necessary or even desirable in a developing intelligent system. No one is ever completely consistent. What is important is how one handles paradox or conflict, how one learns from mistakes, how one turns aside from suspected inconsistencies.

Because of this kind of misconception, Gödel's incompleteness theorem has stimulated much foolishness about alleged differences between machines and men. No one seems to have noted its more "logical" interpretation: that enforcing consistency produces limitations. Of course there will be differences between humans (who are demonstrably inconsistent) and machines whose designers have imposed consistency. But it is not inherent in machines that they be programmed only with consistent logical systems. Those "philosophical" discussions all make these quite unnecessary assumptions! (I regard the recent demonstration of the consistency of modern set theory, thus, as indicating that set theory is probably inadequate for our purposes—not as reassuring evidence that set theory is safe to use!)

A famous mathematician, warned that his proof would lead to a paradox if he took one more logical step, replied "Ah, but I shall not take that step." He was completely serious. A large part of ordinary (or even mathematical) knowledge resembles the cautions in dangerous professions: When are certain actions unwise? When are certain approximations safe to use? When do various measures yield sensible estimates? Which self-referential statements are permissible if not carried too far? Concepts like "nearness" are too valuable to give up just because no one can exhibit satisfactory axioms for them.

In summary:

(1) "Logical" reasoning is not flexible enough to serve as a basis for thinking: I prefer to think of it as a collection of heuristic methods, effective only when applied to starkly simplified schematic plans. The consistency that logic absolutely demands is not otherwise usually available—*and probably not even desirable!*—because consistent systems are likely to be too weak.

(2) I doubt the feasibility of representing ordinary knowledge effectively in the form of many small, independently true propositions.

(3) The strategy of complete separation of specific knowledge from general rules of inference is much too radical. We need more direct ways for linking fragments of knowledge to advice about *how* they are to be used.

(4) It was long believed that it was crucial to make all knowledge accessible to deduction in the form of declarative statements; but this seems less urgent as we learn ways to manipulate structural and procedural descriptions.

I do not mean to suggest that "thinking" can proceed very far without something like "reasoning". We certainly need (and use) something like syllogistic deduction; but I expect mechanisms for doing such things to emerge in any case from processes for "matching" and "instantiation" required for other functions. Traditional formal logic is a technical tool for discussing either *everything that can be deduced from some data* or *whether a certain consequence can be so deduced*; it cannot discuss at all what *ought* to be deduced under ordinary circumstances. Like the abstract theory of syntax, formal logic without a powerful procedural semantics cannot deal with meaningful situations.

I cannot state strongly enough my conviction that the preoccupation with consistency, so valuable for mathematical logic, has been incredibly destructive to those working on models of the mind. At the popular level it has produced a weird conception of the potential capabilities of machines in general. At the "logical" level it has blocked efforts to represent ordinary knowledge, by presenting an unreachable image of a corpus of context-free "truths" that can stand almost by themselves. And at the intellect-modeling level it has blocked the fundamental realization that *thinking begins first with suggestive but defective plans and images that are slowly (if ever) refined and replaced by better ones.*