Progress Report on Artificial Intelligence
Marvin Minsky and Seymour Papert
Dec 11, 1971

This report is the same as Artificial Intelligence Memo AIM-252, and as pages 129-224 of the 1971 project Mac progress report VIII. It was later published as a book, *Artificial Intelligence*, Condon lectures, Univ. of Oregon Press, 1974, now out of print. I have corrected a few misprints and awkward expressions. A scanned image of the original is at ftp://publications.ai.mit.edu/ai-publications/pdf/AIM-252.pdf.

This work was conducted at the Artificial Intelligence Laboratory, a Massachusetts Institute of Technology research program supported in part by the Advanced Research Projects Agency of the Department of Defense and monitored by the Office of Naval Research under Contract number N00014-70-A-0362-0002 and in part by the National Science Foundation under Grant GJ-1049.

At the time of this report, the main foci of attention of the MIT AI Laboratory included

Robotics: Vision, mechanical manipulation. Advanced automation.
Models for learning, Induction, and analogy.
Schemata for organizing bodies of knowledge.
Development of heterarchical programming control structures.
Models of structures involved in commonsense thinking.
Understanding meanings, especially natural language narrative.
Study of computational geometry.
Computational trade-offs between memory size, and parallelism.
Theories of complexities of various algorithms and languages.
New approaches to education.

These subjects were all closely related. The natural language project was intertwined with the commonsense meaning and reasoning study, in turn essential to the other areas, including machine vision. Our main experimental subject worlds, namely the "blocks world" robotics environment and the children's story environment, are better suited to these studies than are the puzzle, game, and theorem-proving environments that became traditional in the early years of AI research. Our evolution of theories of Intelligence has become closely bound to the study of development of intelligence in children, so the educational methodology project is symbiotic with the other studies, both in refining older theories and in stimulating new ones.

The main elements of our viewpoint were to study the use of symbolic descriptions and description-manipulating processes to represent a variety of kinds of knowledge-about facts, about processes about problem solving, and about computation itself. Our goal was to develop heterarchical control structures in which control of problem-solving programs is affected by heuristics that depend on the meanings of events.

The ability to solve new problems ultimately requires the intelligent agent to conceive of, debug, and execute new procedures. Such an agent must know to a greater or lesser extent how to plan, produce, test, edit, and adapt procedures. In short, it must know a lot about computational processes. We are not saying that an intelligent machine, or person must have such knowledge available at the level of overt statements or consciousness, but we maintain that the equivalent of such knowledge must be represented in an effective way somewhere in the system.

This report illustrates how these ideas can be embodied into effective approaches to many problems, into shaping new tools for research, and into new theories we believe important for Computer Science in general, as well as for Robotics, Semantics, and Education.

1.0 Vision and Description

When we enter a room, we feel we see the entire scene. Actually, at each moment most of it is out of focus, and doubly imaged; our peripheral vision is weak in detail and color; one sees nothing in his blind spot; and there are many things in the scene we have not understood. It takes a long time to find all the hidden animals in a child's puzzle picture, yet one

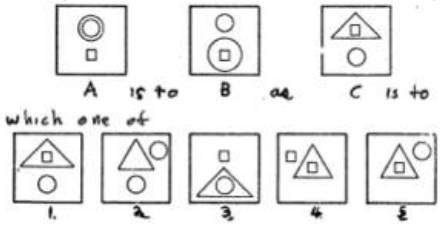
feels from the first moment that he sees everything. People can tell us very little about how the visual system works, or what is really "seen". One explanation might be that visual processes are so fast, automatic, and efficient that there is no place for introspective methods to operate effectively. We think the problem is deeper. In general, and not just in regard to vision, people are not good at describing mental processes; even when their descriptions seem eloquent, they rarely agree either with one another or with objective performances. The ability to analyze one's own mental processes, evidently, does not arise spontaneously or reliably; instead, suitable concepts for this must be developed or learned, through processes similar to development of scientific theories.

Most of this report presents ideas about the use of descriptions in mental processes. These ideas suggest new ways to think about thinking in general, and about imagery and vision in particular. Furthermore, these ideas pass a fundamental test that rejects many traditional notions in psychology and philosophy; if a theory of Vision is to be taken seriously, one should be able to use it to make a Seeing Machine!

1.1 Reasoning by Analogy

To emphasize that we really mean "seeing" in the normal human sense, we shall begin by showing how a computer program -- or a person -- might go about solving a problem of "reasoning by analogy". This might seem far removed from questions about ordinary "sensory perception". But as our thesis develops, it will become clear that there is little merit in trying to distinguish "sensation" or "perception" as separate and different from other aspects of thought and knowledge.

When we give an "educated person this kind of problem from an IQ test, he usually chooses the answer "figure 3":



A is to B as C is to which one of these?

People do not usually consider such puzzles to be problems about "vision." But neither do they regard them as simply matters of "logic". They feel that other, very different mental activities must be involved. Many people find it hard to imagine how a computer program could solve this sort of problem. Such reservations stem from feelings we all share; that choosing an answer to such a question must come from an intuitive comprehension of shapes and geometric relations, rather than from the mechanical use of some rigid, formal rules.

However, there is a way to convert the analogy problem to a much less mysterious kind of problem. To find the secret, one has merely to ask any child to justify his choice of Figure 3. The answer will usually be something like this!

"You go from A to B by moving the big circle down.
"You go from C to 3 in the same way by moving the big triangle."

On the surface this says little more than that something common was found in some transformations relating A with B AND C with 3. As a basis for a theory of the child's behavior it has at least three deficiencies:

It does not say how the common structure was discovered.

It appears to beg the question by relying on the listener to understand that the two sentences describe rules that are identical in essence although they differ in details.

It passes in silence over the possibility of many other such statements (some choosing different proposed answers). For example, the child might just as well have said:

"You go from A TO B by putting the circle around the square..."
"You go from A to B by moving the big figure down," etc.

Aha! If that last statement were applied also to C and 3 the rules would in fact be identical! This leads us to suggest a procedure for a computer and also a "mini-theory" for the child:

Step 1. Make up a description DA for Figure A and a description DC for C.

Step 2. Change DA so that it now describes FIGURE B.

Step 3. Make up a description D for the way that DA was changed in step 2.

Step 4. Use D TO CHANGE DC. If the resulting description describes one of the answer choices much better than any of the others, we have our answer. Otherwise start over, but next time use different descriptions for DA, DC and (perhaps) for D.

Notice that Step 3 asks for a description at a higher level! The descriptions in Steps 1 and 2 describe pictures, e.g. "There is a square below a circle". The description in Step 3 describes changes in descriptions, e.g., "the things around the upper figure in DA is around the lower figure in DB." Our thesis is that one needs both of these kinds of description-handling

mechanisms to solve even simple problems of vision. And once we have such mechanisms, we can easily solve not only harder visual problems but we can adapt them to use in other kinds of intellectual problems as well -- for learning, for language, and even for kinesthetic coordination.

This schematic plan was the main idea behind a computer program written in 1964 by T. G. Evans. [See *Semantic Information Processing*.]

Its performance on "standard" geometric analogy tests was comparable to that of fifteen-year old children! This came as a great surprise to many people, who had assumed that any such "mini-theory" would be so extreme an oversimplification that no such scheme could approach the complexity of human performance. But experiment does not bear out this impression. To be sure, Evans' program could handle only a certain kind of problem, and it does not become better at it with experience. Certainly, we cannot propose it as a complete model of "general intelligence." Nonetheless, analogical thinking is a vital component of thinking, hence having this theory [Evans, 1964], or some equivalent, is a necessary and important step.

In developing our simple schematic outline into a concrete and complete computer program, one has to fill in a great deal of detail: one must decide on ways to describe the pictures, ways to change descriptions, and ways to describe those changes. One also has to define a policy for deciding when one description "fits much better" than another. One might fear that the possible variety of plausible descriptions is simply too huge to deal with; how can we decide which primitive terms and relations should be used? This is not really a serious problem. Try, yourself, to make a great many descriptions of the relation between A and B that might be plausible (given the limited resources of a child) and you will see that it is hard to get beyond simple combinations of a few phrases like "inside of", "left of," "bigger than," "mirror-image of," and so on.

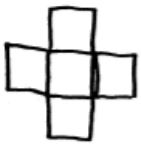
But let us postpone details of how this might be done [see Evans, 1964] and continue to develop our central thesis: by operating on descriptions (instead of on the things themselves), we can bring many problems that seem at first impossibly non-mechanical into the domain of ordinary computational processes.

What do we mean by "description"? We do not mean to suggest that our descriptions must be made of strings of ordinary-language words (although they might be). The simplest kind of description is a structure in which some features of a situation are represented by single ("primitive") symbols, and relations between those features are represented by other symbols -- or by other features of the way the description is put together. Thus the description is itself a MODEL -- not merely a name -- in which some features and relations of an object or situation are represented explicitly, some implicitly, and some not at all. Detailed examples are presented in 4.3 for pictures, and in 5.5 for verbal descriptions of physical situations. In 5.6 there are some descriptions which resemble computer programs. If we were to elaborate our thesis in full detail we would put much more emphasis on procedural (program-like) descriptions because we believe that these are the most useful and versatile in mental processes.

1.2 Children's Use of Descriptions

The theory of analogy we have just proposed might seem both too simpleminded and too abstract to be plausible as a theory of how humans make analogies. But there is other evidence for the idea that mental visual images are descriptive rather than iconic. Paradoxically, it seems that even young children (who might be expected to be less abstract or formal than adults) use highly schematic descriptions to represent geometric information.

We asked a little boy of 5 years to draw a cube. This is what he drew.



"Very good," we said, and asked: *"How many sides has a cube?" "Four, of course,"* he said.

"Of course," we agreed, recognizing that he had understood the ordinary meaning of "side," as of a box, rather than the mathematical sense in which top and bottom have no special status. "How many boards to make a whole cube, then?" "Six," he said, after some thought. We asked how many he had drawn. "Five." "Why?" "Oh, you can't see the other one!"

Then we drew our own conventional "isometric" representation of a cube.

We asked his opinion of it.

"It's no good."
"Why not?"
"Cubes aren't slanted!"

Let us try to appreciate his side of the argument by considering the relative merits of his "construction-paper" cube against the perspective drawing that adults usually prefer. We conjecture that, in his mind, the central square face of the child's drawing, and the four vertexes around it, are supposed in some sense to be "typical" of all the faces of the cube. Let us list some of the properties of a real three-dimensional cube:

Each face is a square.

Each face meets four others.
All plane angles are right angles.
Each vertex meets 3 faces.
Opposite edges on faces are parallel.
All trihedral angles are right angles, etc.

Now, how well are these properties realized in the child's picture?

Each face is a square.
The "typical" face meets four others!
All angles are right!
Each typical vertex meets 3 faces.
Opposite face edges are parallel!
There are 3 right angles at each vertex!

But in the grown-up's pseudo-perspective picture we find that:

Only the "typical' face is square.
Each face meets only two others.
Most angles are not right.
One trihedral angle is represented correctly in its topology, but only one of its angles is right.
Opposite edges are parallel but only in "isometric," not in true perspective.

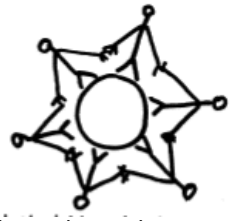
And so on. In the balance, one has to agree that the geometric properties of the cube are better depicted in the child's drawing than in the adult's! Or, perhaps, one should say that the properties depicted symbolically in the child's drawing are more directly useful, without the intervention of a great deal more knowledge.

One could argue that in the adult's drawing, the square face and the central vertex are understood to be "typical." We gave him the benefit of the doubt. Also, one never sees more than 3 sides of a cube, but children don't seem to know this, or feel that it is important. The parallelisms and the general "four-ness" surely dominate.

Incidentally, we do not mean to suggest that our child had in his mind anything like the graphical image of his drawing, but rather that he has a structural network of properties, features, and relations of aspects of the cube, and that what he drew matches this structure better than does the adult's more iconic picture. In 4.4 we will show how such structural networks can be used a program that learns new concepts as a result of experience.

Not all children will draw a cube just this way. They usually draw some arrangement of squares, however, and this sort of representation is typical of children's drawings, which really are not "pictures" at all, but attempts to set down graphically what they feel are the important relations between things and their parts.

Thus *"a ring of children holding hands around the pond"* is drawn like this, perhaps because the correct perspective view would put some of the children in the water.



Also, in the child's drawing the people are all at right angles to the ground, as they should be! For the same reason, perhaps, *"Trees on a Mountain"* is drawn this way

This is presumably because trees usually grow straight out of the ground. It doesn't matter if an actual scene is right in front of the child; he will still draw the trees sideways!

A person is often drawn this way, perhaps partly because the body that is so important to the adult doesn't really do much for the child except get in his way, partly because it does not have an easily-described shape.

From all this we are led to a new view of what children's drawings mean. The child is not trying to draw "the thing itself" -- he

is trying to make a drawing whose description is close to his description of that thing -- or, perhaps, is constructed in accord with that description. Thus the drawing problem and the analogy problem are related.

We hope no reader will be offended by the schematic simplicity of our discussion of "typical children's drawings". Certainly we are focusing on some common phenomena, and neglecting the fantastic variety and plasticity of what children do and learn. Yet even in that plasticity we see the dominance of symbolic description over iconic imitation.

Most children before 5 or 6 years old draw faces like this.

Find such a child and ask, *"Where is his hair?"* and draw some hair,

or say, *"Why doesn't his nose stick out?"* and draw an angular line in the middle of the face.

Chances are that if the child pays any attention at all and likes your idea, these features will appear in every face he draws for the next few months.

The hair is obviously symbolic. The new nose is no better, optically, than the old, but the child is delighted to learn a symbolism to depict protrusion.

There is a vast literature describing phenomena and theories of "learning" in terms of the gradual modification of behavior (or behavioral "dispositions") over long sequences of repetition and tedious "schedules" of reward, deprivation and punishment. There is only a minute amount of attention to the kind of "one-trial" experience in which you tell a child something, or in which he asks you what some word means. If you tell a child, just once, that the elephants in Brazil have two trunks, and meet him again a year later, he may tell you indignantly that they do not.

The success of Evans' program for solving analogy problems does not prove anything, in a strict sense, about the mechanisms of human intelligence. But such programs certainly do provide the simplest (indeed, today the only) models of this kind of thinking that work well enough to justify serious study.

It is natural to ask whether human brains "really" use symbolic descriptions or, instead, manage somehow to work more "directly" with something closer to the original optical image. It would be hard to design any direct experiment to decide such a question in view of today's limited understanding of how brains work. Nevertheless, the formalistic tendencies shown in the children's drawings point clearly toward the symbolic side. The phenomena in the drawings suggest that they are based on a rather small variety of elementary object-symbols, positioned in accord with a few kinds of relations involving those symbols, perhaps taken only one or two at a time. These phenomena are not seen so clearly in the pictures of sophisticated artists, but even so we think the difference is only a matter of degree. While it is possible to train oneself to draw with quantitative accuracy some aspects of the "true" visual image, the very difficulty of learning this is itself an indicator that the symbolic mode is the more normal manner of performance. Even sophisticated adults often show a preference for unreal but tidy "isometric" drawings over more "realistic" perspective drawings,

even though a cube is never seen exactly as in (1). In any case, all this suggests that "graphic" visual mechanisms become operative later (if at all) in human intellectual development than do methods based on structural descriptions. This conclusion seems surprising because in our culture we are prone to think of symbolic description as advanced, abstract, and intellectual, hence characteristic of more advanced stages of maturation.

2.1 Appearance and Illusion

Now consider some phenomena that might seem to be more visual, less intellectual. These two figures show the same rectangle.

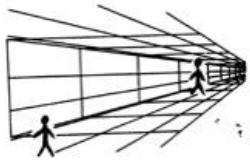
But on the right, the diagonal stripes affect its appearance so that (to most people) the sides appear to lean out and no longer seem perfectly parallel. Such phenomena have been studied with great intensity by psychologists. In the next two figures, the central squares actually have the same grey color, but everyone sees the one at the left as darker.

A good deal is known about the effects of nearby figures or backgrounds on another figure. Perhaps most familiar is the

phenomenon in which the directions of the oblique segments make the horizontal line in the left figure to be shorter than that in the right figure.

But the strangest illusion of all is this: to many psychologists these phenomena of small perceptual distortions have come to seem more important than the question of why we see the figures at all, as "rectangle," or "square," or as "double-headed arrow!" Surely this problem of how we analyze scenes familiar objects is a more central issue.

Thus one finds much more discussion why the smaller figure looks larger in pictures like this than about why one sees the figures as people at all.

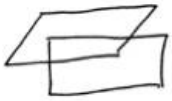


We agree that the study of distortions, ambiguities, and other "illusions" can give valuable clues about visual and other mechanisms. To resolve two or more competing theories of vision, such evidence might become particularly useful. First, however, we need to develop at least one satisfactory theory of how "normal" visual problems might be handled, particularly scenes that are complicated but not especially pathological.

Let us look at a few more visual phenomena. Both of the two figures below appear at first sight to be reasonable pictures of the bases of (triangular) pyramids -- that is, of simple flat-surfaced five-faced bodies that could be pyramids with their tops cut off. But, in fact, figure B cannot be a picture of such a body. For its three ascending edges (if extended) would not meet at a single point, whereas those of figure A do form a vertex for a pyramid.

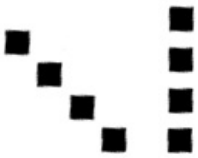
So here we have a sort of negative illusion—because figure B would not "match" a real photograph of any pyramid-base. However, it could match quite well an abstract description of a pyramid base—say, one that describes how its faces and edges fit together (qualitatively, but not quantitatively).

Another topic concerns "camouflaged" figures. The figure "4" embedded in this drawing is not normally seen as such—because, we presume, one describes the scene as a square and parallelogram.



Study of this kind of concealment can tell us something about the "principles" according to which our visual system "usually" describes scenes as made up of objects. But once the "4" has been pointed out or discovered, it is then "seen" quite clearly! A good theory must also account for phenomena in which it is possible to change and elaborate one's "image" of the same scene in ways that depend on changes in his interpretation and understanding of the structure "shown" in the picture.

A simpler—and more interesting—example of a figure with two competitive descriptions is the ordinary square! Young children know the square and the diamond as two quite distinct shapes, and the ambiguity persists in adults, as seen in these two figures by Fred Attneave.

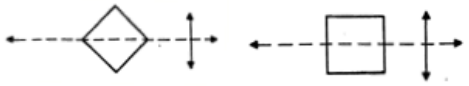


The four objects at the left are usually seen as diamonds, while those on the right are seen as squares. How can we explain this? Since the individual objects are in fact identical, the effect must have something to do with their arrangement. It is tempting to incant the phrase—"the whole is more than the sum of the parts."

Now consider a descriptive theory. If one is asked to describe this scene, he will say something like:*"There are two rows, each with four objects. One is a horizontal row of (etc.)".* We ignore details here, but suggest that the description is dominated by the grouping into rows, as indicated by their priority in the verbal presentation of the description. In section 4.6 we discuss a program that does something of this sort.

By "description" we do not usually mean "verbal description"; we mean an abstract data structure in which are represented features, relations, functions, references to processes, and other information. Besides representing things and relations between things, descriptions often contain information about the relative importance of features to one another, e.g., commitments about which features are to be regarded as essential and which are merely ornamental. For example, much of linguistic structure is concerned with the ability to embed hierarchies of detail into descriptions: subordinate clause formation and other word-order choices often reflect priorities and progressions of structural detail in the descriptions that are "meant." We will return to this in section 5.

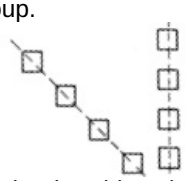
Once committed to describing a row of things, the choice between seeing squares and diamonds begins to make more sense. Which description does one choose? Apparently, the way one describes a square figure depends very much on how one chooses (in one's mind) the axis of symmetry. Consider the differences in how one might describe the same figure in these two different orientations:



points on axis sides parallel to axis
one point on each side two points on each side
made of two triangles made of two rectangles
unstable on ground stable, flat bottom
hurts when squeezed safe to pick up

These two descriptions could hardly be more different! No wonder that most 3 year olds do not believe that they are the same. In fact, children's drawings of diamonds often come out as

indicating that their descriptive image is a composition of two triangles, or at least that the most important features are the points on the symmetry axes. Our mystery is then almost solved: whatever process set up the description in terms of rows set up also a spatial frame of reference for each group.



Since one has to choose an axis for each square and "other things being equal" there is no strong reason locally for either choice, one tends to use the axis inherited from the direction of its "row." The fact that you can, if you want, choose to see any of the objects as either diamond or square only confirms this theoretical suggestion -- the choice is by default only, and hence would be expected to carry little force.

Once this door is opened, it suggests that other choices one has to make in visual description also can depend on other alien elements in one's thoughts—as well as on other things in the picture! Every simple figure is highly ambiguous. In a face, a circle can be an eye, a mouth, an ear, or the whole head. There should be no difficulty in admitting this to our theory —or to the computer programs that demonstrate its consistency and performance. Traditional theories directed toward physical (rather than on computational, or symbolic) mechanisms were inherently unable to account for the influence of other knowledge and ideas upon "perception".

2.2 Sensation, Perception and Cognition

Our discussion of how images depend on states of mind is part of a broader attack on the conventional view of the structure of mind. In today's culture we grow up to believe that mental activity operates according to some scheme in which information is transformed through a sequence of stages like:

World ==> Sensation ==> Perception ==> Recognition ==> Cognition

Although it is hard to explain exactly what these stages or levels are, everyone comes to believe that they exist. The "new look" in ideas about thinking rejects the idea that there are separate activities like "perception" that precede and are independent of "higher" intellectual activities. What one "sees" depends very much on one's current motives, intentions, memories, and acquired processes. We do not meant to say either that the old layer-cake scheme is entirely wrong or that it is useless. Rather, it represents an early concept that was once a clarification but is now a source of obscurity, for it is technically inadequate against the background of today's more intricate and ambitious ideas about mechanisms.

> The higher nervous system is embryologically, and anatomically divided into stages of some sort and this might suggest a basis for the popular-science hierarchy. This makes sense for the most peripheral sensory and motor systems, in which transmission between anatomical stages is chiefly unidirectional. But (presumably) when we go further in the central direction this is no longer true, and one should not expect the geometrical parts of a cybernetic machine to correspond very well to its "computational parts.

Indeed, the very concept of "part", as in a machine, must be rebuilt when discussing programs and processes. For example, it is quite common in computer programs—and, we presume, in thought processes—to find that two different procedures use each other as subprocedures! We shall see this happening throughout section 5. In such a case one can hardly think of either process as a proper part of the other. So the traditional view of a mechanism as a HIERARCHY of parts, subassemblies and sub-sub-assemblies (e.g., the main bearing of the fuel pump of the pitch vernier rocket of the second ascent stage) must give way to a HETERARCHY of computational ingredients.

It is unfortunate that technical theories, and even practical guidelines, for such heterarchies are still in their infancies. The rest of this chapter discusses some aspects of this problem.

2.3 Parts and Wholes

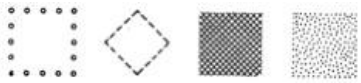
A recurrent theme in the history of psychological thinking involves recognizing an important distinction without having the

technical means to give it the appropriate degree of precision. Consequently, the dividing line becomes prematurely entrenched in the wrong place. An influential example was the concept of "Gestalt". This word is used in attempts to differentiate between the simplest immediate and local effects of stimuli, and those effects that depend on a much more "global" influence of the whole stimulus "field".

Here is a visual example in which this kind of distinction might be considered to operate: In one sense, this arch is "nothing but" three blocks.

But the arch has properties -- as a single whole -- that are not inherited directly from properties of its parts in any simple way. Some of those arch properties are shared also by these structures:

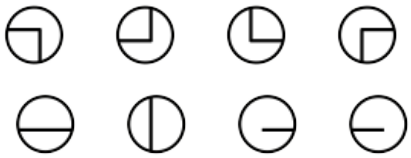
Obviously the properties one has in mind do not reside in the individual building blocks, they "emerge" from the arrangements of those parts. And one finds this in even simpler situations. Obviously we react to a simple outline square in a way that is very different from our reactions to four separate lines, and rather similar to how we react to such graphically different figures as these:



The question "whence comes the square if not from its parts" is not really very serious here, for it is easy to make theories about how one might "perceive" a shape if there are enough easily-detected features to approximately delineate its geometric form. But there is no similarly easy solution to the kinds of problems that arise when one looks at three-dimensional scenes.

The next two figures are "locally identical" in the following precise sense: Imagine innumerable experiments, in each of which we choose a different point of the picture to look at, and record what we see only within a very small circle around that point.

Both pictures would produce identical collections of data -- provided that we keep no records of the locations of the viewpoints. So in this sense both pictures have the same "parts". They are obviously very different, however. One particularly outstanding difference is that one picture is all in one piece—it is CONNECTED—while the other is not. In fact, both pictures are composed of just these kinds of "micro-scenes" and both figures have exactly the same numbers of each.



In our book *Perceptrons* we prove that in general one cannot use statistics about such local evidence to distinguish between figures that are "connected" and those that are not.

From this one might conclude that one can tell very little about a picture from such "spatially local" evidence. But this is not true. For example, we can completely define the property of being "made-entirely-of- separate, *solid*, rectangles" by requiring that all very small parts of the scene look like one or another of these micro-scenes: that is, *every micro-scene must be either homogeneous, a simple edge, or a right-angle corner*.

It is not hard to see that this definition will accept any picture that contains only solid rectangles, but no other kind of picture. So in this sense "rectangle-ness" can be defined in terms of local properties, while connectedness cannot. Try to define *"composed-of-a-single-solid-rectangle"* in this way. It cannot be done! So we see a difference between two kinds of categories of pictures, in regard to the relations between their parts and their wholes!

The question "Is the whole more than the sum of its parts" is certainly provocative and insightful. But it must be recognized also as vague, relative, and metaphorical. What is meant by "parts" and, more important, what is meant by "sum"?

In the case of the rectangles a trivial sense of "sum" will suffice: not even adding up evidence is necessary, for we can make the decision in favor of rectangle, and let any single exception to our condition on the local "micro-scenes" have absolute veto power. So the "sum of the parts" is simply the agreement of all local evidence. For connectedness we seem to need something more complicated, computationally. We have studied this situation rather deeply in *Perceptrons*: connectedness is a property that is quite important and very thoroughly understood in classical mathematics; it is in fact the central concern of the entire subject of Topology.

For example, here are several quite different-looking conditions each of which can be used to define the same concept of connectedness:

PATH-CONNECTION. For any two black points of the picture, there is a path connecting them that lies entirely in black points.

PATH-SEPARATION. There is no closed path, entirely in white points, such that there are some black points inside the path and some black points outside the path.

SET-SEPARATION. The black points cannot be divided into two non-empty sets which are separated by a non-zero distance -- that is, no pair of points, one from each set, are closer than a certain distance.

TOTAL-CURVATURE. Assume that there are no "holes" in the black set -- that is, white points that are cut off from the outside by a barrier of black points. Then compute the sum of all the boundary curvatures (direction-changes at all edges of the figure), taking convex curves as positive and concave curves as negative. The picture is connected if this sum is exactly 360 degrees.

Each of these suggests different computational approaches. Depending upon what resources are available, one or another will be more efficient, use more or less memory, time, hardware, etc. Each definition involves very large calculations in any case, except the fourth, in which one computes simply a sum of what one observes in each small neighborhood. However, the fourth definition does not work in general, but only for figures without holes. And, to be sure that condition is satisfied one must have another source of information (e.g., if one knows he is counting pennies) or else the definition is somewhat circular, because to be able to see that there are no holes is really equivalent to being able to see that the background is connected!

We know exactly what it means for the number seven to be the sum of the numbers three and four. But when we ask whether a house is just the sum of its bricks, we are in a more complicated situation. One might answer:

> "Yes, there is nothing but bricks there".

But another kind of answer could be

> "No, for the same bricks arranged differently would have made a very different house."

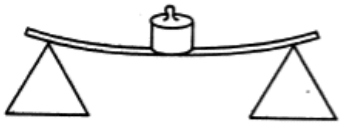
The answer must depend on the purpose of the question. If we admit only "yes" or "no", there is no room for refinement and subtlety of discussion. We do not really want either of the answers "Yes, it is nothing but the sum" or "No, it is a Gestalt, a totally different and new thing". We really want to know exactly how the response, image, or interpretation of the situation is produced: we want an explanation of the phenomenon. And the terms of the explanation must be appropriate to the kind of technical question we have in mind. Sometimes one wants the result in terms of a particular set of psychological concepts, sometimes in terms of the interconnections of some perhaps hypothetical neural pathways, and sometimes in terms of some purely computational schemata. Thus one might ask, about some aspect of a person's behavior:

> COMPONENTS: Can the phenomenon be produced in a certain kind of theoretical neural network?
>
> LEARNING: Can it be learned by a certain kind of reinforcement schedule according to certain proposed laws of conditioning?
>
> COMPUTATIONAL STRUCTURE: Can this result be computed by a computer-like system subject to certain restrictions, say, on the amount of memory, or on the exclusion of certain kinds of loops interconnecting its components?
>
> COMPUTATIONAL SCHEMATA: Can the outer behavior of this individual reasonably be imitated by a program containing such-and-such a data-structure and such-and-such a syntactic analyzer and synthesizer?

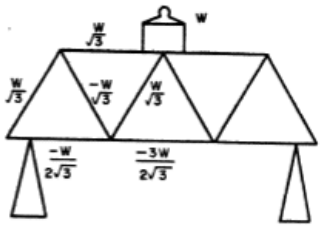
The way in which the whole depends upon its parts, for any phenomenon, has a direct bearing on how such questions can be answered. But to supply sensible answers, one needs a stock of crisp, precise, ideas about how parts and wholes may be related!

It is important to recognize that these kinds of problems are not special to Psychology. Water has properties that are not properties either of hydrogen or oxygen, yet chemistry is no longer plagued by fights between two camps—say, "Atomist" vs. "Gestalt". This is not at all because the problem is unimportant: exactly the opposite! The reason there are no longer two camps in Chemistry is because all workers recognize that the central problems of the field lie in developing good theories of the different kinds of interactions involved, and that the solution of such problems lie in constructing adequate scientific and mathematical models rather than in defending romantic but irrelevant philosophical overviews. But in Psychology and Biology, there remains a widespread belief that there are phenomena of mind or of cell that are not "reducible" to properties and interactions of the parts. They are saying, in essence, that there can be no adequate theory of the interactions.

Consider a concrete example. It is relatively easy to bend a thin rod, but much harder to bend this structure made of several such rods. Where does the extra stiffness come from?



The answer, in this case, is that the "new property" is indeed inherited from the parts, because of the arrangement, but in a peculiar way. In the truss, a force at the middle is resisted -- not by bending-forces across the rods -- but by compression and tension forces along the rods.



The resistance of a thin rod to forces along it is much greater than the resistance to forces across it. So the increased strength is indeed "reduced", in the Theory of Static Mechanics, to the interactions of stresses between members of the structure. Even the properties of a single rod itself can be explained in terms of more microscopic interactions of the tensile and compressive forces between its own "parts", when it is strained. By imagining the rod itself to be a truss (a heuristic planning step that helps one to write down the correct differential equation) we can analyze stress-strain relations inside the rod. Thus one obtains such a beautiful and accurate model that there remains no mysterious "Gestalt" problem at all. This is not to say that special arrangements have no special properties. In some of Buckminster Fuller's work, the dodecahedral sphere is yields a kind of structural stiffness rather different than that in the triangular truss. Here the rigidity does not come directly from that of small or "local" triangular substructures, and it takes a different kind of mathematical analysis to see why it is hard to distort it. Even so, there remains no mysterious "emergent" property here that cannot be deduced from the classical theory of statics.

Of course, our real concern is with problems of intelligence, rather than with engineering mechanics. But many problems that seem at first to be "purely psychological" often turn out to center around just such problems of wholes and parts. And with such an interpretation, we may replace an elusively ill-defined psychological puzzle by a much sharper problem within the theory of computation.

The computer is the example par excellence of mechanisms in which one gets complex results from simple interactions of simple components. In asking how thought-like activity could be embedded in computer programs, scientists for the first time really came to grips with understanding how intelligent behavior could be made to emerge from simple interactions.

The issue seems really to be fundamentally one of assessing the complexity of processes. The content of the gestalt discoveries is that certain psychological phenomena require forms of computation that lie outside the scopes of certain models of the brain -- and outside certain conjectures about the "elementary" units of which behavior is supposed to be composed. So, the whole discussion must be considered in relation to some overt or covert commitment about what units of behavior, or of brain-anatomy, or of computational capacity, are supposed to be "atomic".

To illustrate extreme versions of atomism vs. gestaltism one might consider these caricatures:

> Extreme ATOMISM: all behavior can be understood in terms of simple functions of neural paths that run from single receptors, through internuncials, to effectors.
>
> Extreme GESTALTISM: The essence in is the whole pattern. Many simple examples show that the response is made to the whole stimulus and cannot be represented as simple sums or products of simple local stimulations.

Clearly one does not want to set a threshold between these; one wants to classify intermediate varieties of interactions that might be involved, arranged if possible in some natural order of complexity. Thus in *Perceptrons* we studied a variety of simple schemas such as these:

> EXTREMELY ATOMIC ALGORITHM: One of the input wires is connected to the output, the others to nothing.
>
> VETO ALGORITHM: If every input says "yes", the output is "yes". If any input says "no", the output is "no".
>
> MAJORITY ALGORITHM: If M or more of N inputs say "yes", output is "yes".
>
> LINEAR SUM ALGORITHM: To each input is assigned a "weight". Add together the weights for just those inputs that say "yes". The output is just this sum.
>
> LINEAR THRESHOLD ALGORITHM: Use the LINEAR SUM algorithm, except, make the output "yes" if the sum is greater than a certain "threshold", otherwise the output is "no".

Exercise: the reader should convince himself that "extremely atomic", "veto", and "majority" are special cases of "linear threshold".

> EQUIVALENT-PAIR ALGORITHM: The input is considered to be grouped in pairs. The output is "yes" only when, for every pair, the two members have the same input value.

The reader should convince himself that this is not a special case of "linear threshold"!

> SYMMETRICAL ALGORITHM: The response is "yes" if the pattern of inputs is symmetrical about some particular center, or about some particular linear axis.

This is a special case of the equivalent-pair algorithm. They are both examples of perceptrons in which the global function can be expressed as a linear threshold function of intermediate functions of two variables. Here the whole is only trivially more than the sum of the parts.

PERCEPTRON ALGORITHM: First some computationally very simple functions of the inputs are computed, then one applies a linear threshold algorithm to the values of these functions.

Many different classes of perceptrons have been studied; such a class is defined by choosing a meaning for the phrase "very simple function." For example, one might specify that such a function can depend on no more than five of the stimulus points. This would result in what is called an order-five perceptron. All of the examples above had order one or two. The next example has no "order restriction", but the functions are very simple in another sense; they are themselves "order one" or linear-threshold functions.

GAMBA PERCEPTRON: A number of linear threshold systems have their outputs connected to the inputs of a linear threshold system. Thus we have a linear threshold function of many linear threshold functions.

Virtually nothing is known about the computational capabilities of this latter kind of machine. We believe that it can do little more than can a low order perceptron. (This, in turn, would mean, roughly, that although they could recognize some relations between the points of a picture, they could not handle relations between such relations to any significant extent.) That we cannot understand mathematically the Gamba perceptron very well is, we feel, symptomatic of the early state of development of elementary computational theories.

Which of these are atomic and which are gestaltist? Rather than muddle through a philosophical discussion of which cases "really" do more than add the parts, we should try to classify the kinds of mechanisms needed to realize each in certain "hardware" frameworks, chosen for good mathematical reasons. Then for each such framework, we might try to see which admit simple reinforcement mechanisms for learning, which admit efficient descriptive teaching (see section 4), and which admit the possibility of the cognitive machinery "figuring out for itself" what are the important aspects of a situation!

To supply such ideas, we have to make theoretical models and systems. One should not expect to handle complex systems until one thoroughly understands the phenomena that may emerge from their simpler subsystems. This is why we focused so much attention on the behavior of Perceptrons in problems of Computational Geometry. It is important to emphasize that we want to understand such systems for the reasons explained above, rather than as possible mechanisms for practical use. When a mathematical psychologist uses terms like "linear", "independent", or "Markoff Process", etc., he is not (we hope!) proposing that a human memory is one of those things; he is using it as part of a well-developed technical vocabulary for describing the structure of more complicated schemata. But until recently there was a serious shortage of ways to describe more procedural aspects of behavior.

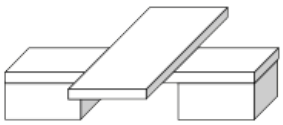
The community of ideas in the area of computer science makes a real change in the range of available concepts. Before this, we had too feeble a family of concepts to support effective theories of intelligence, learning, and development. Neither the finite-state and stimulus-response catalogs of the Behaviorists, the hydraulic and economic analogies of the Freudians, or the holistic insights of the Gestaltists supplied enough technical ingredients to develop such an intricate subject. It needs a substrate of debugged theories and solutions to related but simpler problems. Computer science has brought a flood of such ideas, well defined and experimentally implemented, for thinking about thinking; only a fraction of them have distinguishable representations in traditional psychology:

symbol table	closed subroutine
pure procedure	pushdown list
time-sharing	interrupt
calling sequence	communication cell
functional argument	common storage
memory protection	decision tree
dispatch table	hardware-software trade-off
error message	serial-parallel trade-off
function-call trace	time-memory trade-off
breakpoint	conditional breakpoint
formal language	asynchronous processing
compiler	interpreter
indirect address	garbage collection
macro language	list structure
property list	block structure
data type	look-ahead
hash coding	look-behind (cache)
micro-program	diagnostic program
format matching	executive program
syntax-direction	operating system

These are only a few ideas from the environment of general "systems programming" and debugging—and we have mentioned none of the much larger set of concepts specifically relevant to programming languages, artificial intelligence research, computer hardware and design, or other advanced and specialized areas. All these serve today as tools of a curious and intricate craft, programming. But just as astronomy succeeded astrology, following Kepler's discovery of planetary regularities, the discoveries of these many principles in empirical explorations of intellectual processes in machines should lead to a science, eventually.

3.0 Analysis of Visual Scenes

No one could have any doubt about what this picture is supposed to show: "Four blocks, three forming a bridge with the fourth lying across it."



We would like to program a machine to be able to understand scenes to at least this level of comprehension. Notice that our description involves recognizing the "bridge" as well as the blocks that comprise it, and that the phrase "lying across it" indicates knowing that the block is actually resting on the bridge. Indeed, in the pronoun reference to the bridge, rather than to the top block of the bridge, there is implied a further level of functional analysis.

In our earlier progress reports we described the SEE program [Guzman 1968], which was able to assemble the thirty vertices, forty segments and thirteen regions of this picture into four objects, using a variety of relatively local "linkage" cues. A new program, [Winston 1970] goes further in the analysis of three-dimensional support and can recognize groups of objects as special structures (such as "bridge") to yield just the kind of functional description we are discussing. Winston's program is even able to LEARN to recognize such configurations, using experience with examples and non-examples, as shown in chapter 4

Before discussing scene-analysis in detail, we have a few remarks about the nature of problems in this area. In the early days of cybernetics [McCulloch-Pitts 1943, Wiener 1949] it was felt that the hardest problems in apprehending a visual scene were concerned with questions like "why do things look the same when seen from different viewpoints", when their optical images have different sizes and positions.

How does one capture the "abstraction" or "concept" common to all the particular examples? For two-dimensional character-recognition, this kind of problem is usually handled by a two-step process in which the image is first "normalized" to standard position and then "matched" -- by a correlation or filtering process -- to one of a set of standard representatives. In practical engineering applications, the "normalizing" often failed because it could not disarticulate parts of images that touch together, and "matching" often failed because it is hard to make correlation-like processes attend to "important" parts of the figures instead of to ornaments. Even so, such methods work well enough for reasonably standardized symbols.

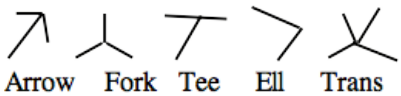
If, however, one wants the machine to read the full variety of typography that a literate person can, the problem is harder, and if one wants to deal with hand-printing, quite different methods are needed. One is absolutely forced to use exterior knowledge involving the pictures' contexts, in situations like this. [Selfridge, 1955]

Here the distinction between the "H" and the "A" is not geometric at all, but exists only in one's knowledge about the language. An early program that could do this was described in [Bledsoe and Browning 1959]. But we will not stop to review the field of character-recognition, for its technology is quite alien to the problems of three-dimensional scenes. This is because the problems that concern us most, like how to separate objects that overlap, or how to recognize objects that are partially hidden (either by other objects or by occluding parts of their own surfaces), simply do not occur at all in the two-dimensional case. Some more interesting two-dimensional problems require description when geometric matching fails; a conceptual "A" is not simply a particular geometric shape; it is

"Two lines of comparable length that meet at an acute angle, connected near their middles by a third line."

3.1 Finding Bodies in Scenes

Let us review quickly how Guzman's SEE program works. First a collection of “lower level” programs is made to operate directly on the optical data. Their job is to find geometric features of the picture -- regions, edges and vertices -- so that the scene can be described in a simple way in the program's data-structure. Next, the vertices are classified into "types". The most important kinds are these:

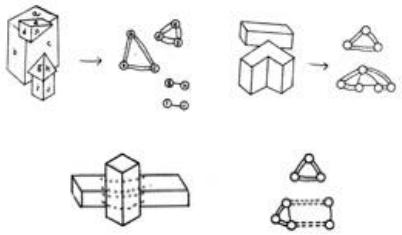


The main goal of the program is to divide the scene into "objects", and its basic method is to group together regions that probably belong to the same object. Each type of vertex is considered to provide some evidence about such groupings, and can be used to create "links" between regions.

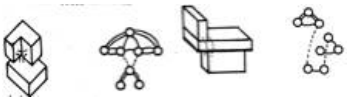
For example, the ARROW type of vertex usually is caused by an exterior corner of an object, where two of its plane surfaces form an edge. So we insert a "link" between the two regions that are bounded by the two smaller angles:

Similarly, the FORK type of vertex, which is usually due to three planes of one object, causes three links between those regions.

Using these clues, and representing the resulting relations by simple abstract networks, many scenes are "correctly" analyzed into objects.

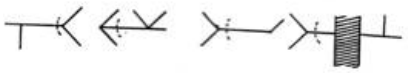


Many scenes are handled correctly by just these simple rules, but many are not. For example, the basic assumption about the FORK linking its three regions is not true of concave corners, and the "matching TEE" assumption may be false by coincidence, so that "false links" may be produced in such cases as these:



Guzman introduced several methods for correcting such errors. One method involves a conservative procedure in which groupings are considered to have different qualities of connectedness. Two high-quality groups that are connected together by only a single link are broken apart -- the link is deleted.

A second error-correction method is more interesting. Here we observe that the TEE vertex really has a special character, quite opposed to that of the FORK and the ARROW. The most usual physical cause of a TEE is that an edge of one object has disappeared under an edge of another object. Hence we should regard the TEE joint as evidence against linking the corresponding regions! Guzman's implementation of this was to recognize certain kinds of configurations as special situations in which the existence of one kind of vertex-type causes inhibition or cancellation of a link that would otherwise be produced by the other vertex-type. That would happen, for example, in these figures:



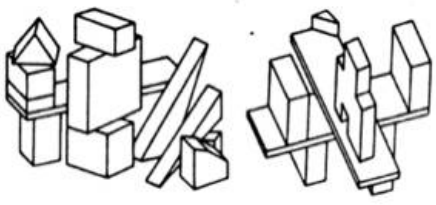
This technique corrects many errors that the more "naive" system makes, especially in objects with concavities. Note that this program attempts to compute Connectedness (for the notion of "object" used here is exactly that)—by extremely local methods, while the (better) system with cancellation is less local because of the effects of vertex-types of contiguous or closely related geometric features.

Guzman's method might seem devoid of the normalization and matching operations. Indeed, in a sense it has nothing to do with "recognizing" at all; it is concerned with the separation of bodies rather than with their shapes. But both normalization and matching are more or less inherent in the descriptive language itself, since the very idea of vertex-type is that of a micro-scene which is invariant of orientation, scale, and position.

This scheme of Guzman is very much in accord with the Gestaltists' conceptual scheme in which the separation of figure from background is considered prior to and more primitive than the perception of form.

The "cancellation" scheme has a more intelligible physical meaning. It has been pointed out by D. Huffman [1970] that each line in a line drawing may be interpreted as a physical edge formed (we assume) by the intersection of two planes, at least locally. In some cases one can see parts of both planes, but in other cases only one. A T-joint is good evidence that the edge involved is of the latter kind, and once one assigns such an interpretation to an edge, then it follows immediately that the adjacent Guzman links to the alien surface ought to be rejected. Accordingly, Huffman developed a number of procedures for making detailed global interpretations from local edge-region assignments.

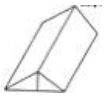
We will not give further details of the SEE program here. As an example of its performance, it correctly separates all the objects in this scene.



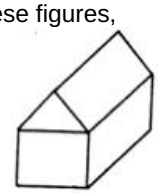
But SEE has faults, which include:

ORDINARY "MISTAKES": Certain simple figures are not handled "correctly." To be sure, all figures are inherently ambiguous (any scene with n regions could conceivably arise from a picture of n objects). Our real goal is to find an

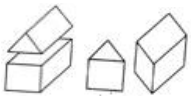
analysis that makes sense in everyday situations. Normally one would not suppose that this is a single body, but SEE says it is, because all regions get linked together.



INFLEXIBILITY: If its very first proposal is not acceptable, the body-aggregation program ought to be able to respond to complaints from other higher and lower level programs and thus generate some alternative "parsings" of the scene. For example, SEE finds a single body in the top one of these figures,



but it should be able to produce the two other alternatives shown below it. (It is interesting how difficult it is for some humans to see the third parsing.)

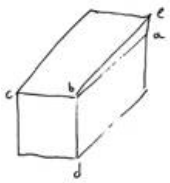


IGNORANCE: It has no way to use knowledge about common or plausible shapes. While it is a virtue to be able to go so far without using such exterior information, it is a fault to insist on this!

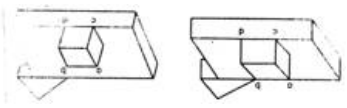
Following Guzman's work, Martin Rattner has described a procedure, called SEEMORE, that can handle some of these problems. [Rattner 1970] While it uses linking heuristics much as did Guzman, SEEMORE puts more emphasis on local evidence that an edge might separate two bodies. These "splitting heuristics" operate initially at certain kinds of vertices, notably TEE-vertices and vertices with more than three edges (which were not much used in earlier programs). When there is more than one plausible alternative, SEEMORE uses other evidence to make tentative choices of how to continue a splitting line, but stores these choices on back-up lists that can later be used to generate alternative parsings.

Here is a simple example. In this figure, one might imagine splitting either along the line a-b-c or along the line d-b-e.

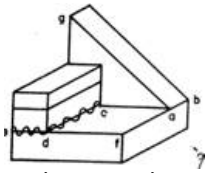
The central vertex 'b' suggests (locally) either of these; on the other hand, such splits as a-b-d or a-b-e are considered much less likely.



Degenerate situations like this, in which a small change in viewing angle produces a different topology, are likely to lead to "incorrect" analyses. Rattner uses a rather conservative linking phase, in which links are placed more cautiously than in SEE, but using similar "inhibiting" rules. Regions that are doubly linked to one another by these are considered "strongly" bound; then the heuristic rule is to attempt to split around these "nuclei," and to avoid splitting through them.



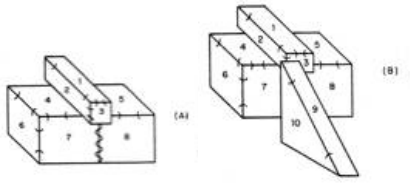
It would be tedious to give full details here, partly because the subject is so specialized, but primarily because the procedure has not been tested and debugged in a wide enough variety of situations. A few examples follow.



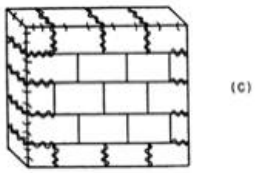
An initial split is made along e-d, extended to d-c. Then, between the possible splits g-a-f and c-a-b, the latter is preferred because it completes the unfinished split ending at 'c'.

In this situation, B is the procedure's first choice, C its second:

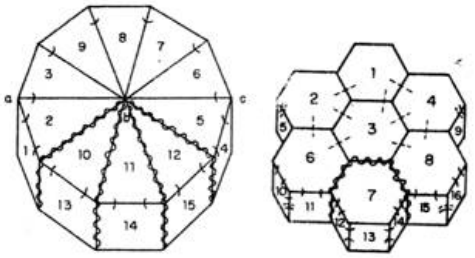
In A below, we get three bodies, (4-6-7), (5-8), and (1-2-3). SEE does not split between regions 7 and 8.



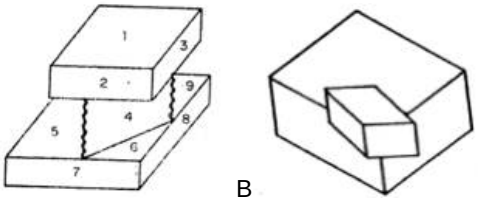
In B, one gets the plausible three-body analysis. If there is any complaint, SEE MORE will propose to separate (4-6-7) and (5-8).
In C, all the bricks are properly separated. While SEE would have to put in many spurious links because of the coincidentally matching TEE's, SEEMORE inhibits these on the basis of other splitting evidence.



The procedure divides these into the "natural" parts:



But in figure A below it finds three bodies 1-2-3, 5-7-8-9, and 4-6. The latter is perhaps not the first way a person would see it. And the procedure cannot aggregate the outer segments of the larger cube in figure B because its initial grouping process is so conservative.



A B

Clearly, such problems eventually must be gathered together in a "commonsense" reasoning system; the multiple T-joints all would meet, if "extended" in such a way as to suggest the proper split, and the program ought to realize this.

4.0 Description and Learning

The concepts we used to analyze ANALOGY and SEEING are just as vital in understanding LEARNING. It was traditional to try to account for learning in terms of such primitives as "conditioned reflex" or "stimulus-response bond". The phenomena of learning become much more intelligible when seen in terms of "description" and "procedure".

There might seem a world of difference between activities involving permanent changes in behavior—and the rest of thinking and problem solving. But even the temporary structures one obviously uses in imagining and understanding have to be set up and maintained for a time. We feel that the differences in degree of permanence are of small importance compared to the problems of deciding what to remember. It is not the details of how recording is done, but the details of how one solves the problem of what to record, that must be understood first.

As we develop this idea, we find ourselves forced to question the whole tradition in which one distinguishes a special sub-set of mental or behavioral processes called "learning". Nothing but disaster can come from looking for three separate theories to explain (for example)

How one learns mathematics,
how one thinks mathematically once one has learned to, and
what mathematics is, anyway.

We are not alone in trying to replace such subdivisions—but perhaps more radical and thoroughgoing. In this chapter we shall argue that many problems about "learning" really are concerned with the problem of finding a description that satisfies some goal. Gestalt psychologists also often emphasized the similarity between solving apparently abstract problems and situations that intuitively feel like simple perception; the same relation that is dimly reflected in ordinary language by expressions like

"I suddenly saw the solution!"

We thoroughly agree about bringing these phenomena together, but we have a very different way of dealing with the newly

united couple. We might caricature this difference by saying that the Gestaltists might look for simple and fundamental principles about how perception is organized, and then attempt to show how symbolic reasoning can be seen as following the same principles, while we might construct a complex theory of how knowledge is applied to solve intellectual problems and then attempt to show how the symbolic description that IS what one "sees" is constructed according to similar such processes. Indeed, we think that ideas that have come from the study of symbolic reasoning have done more to elucidate visual perception than ideas about perception have clarified our thoughts about abstract thinking—but the whole comparison is too dialectical to try to develop technically.

In any case we differ from the Gestaltists more deeply in problems of learning, which they neglected almost entirely -- perhaps because that was the favorite subject of the abominable behaviorists! Let us now explain why we feel that learning, technically, cannot usefully be separated from other aspects either of perception or of symbolic reasoning. As usual, we present first a caricature; then point to where the extreme positions might be softened.

Learning -- or "Keeping Track":

Everyone would agree that getting to know one's way around a city is "learning". Similarly, we see solving a problem often as getting to know one's way around a "micro-world" in which the problem exists. Think, for example, of what it is like to work on a chess problem (or on a geometry puzzle, or trying to fix something). Here the microworld consists of the network of situations on the chessboards that arise when one moves the pieces. Solving the chess problem consists largely of getting to know the relations between the pieces, and how the moves affect things. One naturally uses words like "explore" in this context. As the exploring goes on, one experiences events in which one suddenly "sees" certain relations. A grouping first seen as three pieces playing different roles is now described in terms of a single relation between the three, such as "pin", "fork", or "defense." The experience of re-description can be as "vivid" as if the pieces involved suddenly changed color or position.

One might object that the difference between getting to know the city and solving the chess problem is that one remembers the city and forgets the chess situation (assuming that one does). Isn't that what brings one into the domain of learning and excludes the other? Only to a degree! The chess analysis has to be remembered long enough, within the rest of the analysis. To take an extreme form of the argument, one would repeat one's first steps forever unless one remembered which positions had been analyzed, what relations were observed, and how their descriptions were summarized. What is stored during problem-solving is as vital to the immediate solution as what is retained afterwards is to the solution of the presumably larger-scale problems one is embedded in throughout life. Of course there is a problem about how long one retains what one learns -- but perhaps that belongs to the theory of forgetting rather than of learning!

In our laboratory the chess program written by R. Greenblatt plays fairly good chess, by amateur tournament standards. But visitors are always disappointed to find that this program does not "learn", in the sense that it carries no permanent change away from the games it plays. They are even more disappointed in our attempts to explain why this does not disturb us very much. We claim that there is indeed an important kind of learning within the program; this is in the position-description summaries that are constructed and used as it analyses the positions it is playing. But because board positions do not often repeat exactly in subsequent games (except for opening positions and end-games) and because the kinds of descriptions the program now uses do not have good qualities for dealing with broader classes of positions, there would be no point in keeping such records permanently.

We do not yet understand how to make the higher-level strategy-oriented descriptions that would make sense in the context of learning to improve. When we, ourselves, learn how to construct the right kind of descriptions, then we can make programs construct and remember them, too, and the problem of "learning" will vanish. In the past, our Laboratory avoided experiments with learning systems that seemed theoretically unsound, although we did NOT avoid studying them theoretically. This was because we believed that learning itself was not the real problem; what was needed was more knowledge about the intelligent shaping of description-handling processes. For the same reasons we avoided linguistic exercises such as Mechanical Translation, in favor of studying systems that could deal with limited fragments of meaning, and we avoided "creative" systems based on uninterpreted stochastic processes in favor of analyzing the interactions of design goals and constraints. Now we think we know enough to begin such experiments.

In the rest of this chapter we will discuss some systems that do exhibit some non-trivial learning functions. It should be understood from the start that these are not to be thought of as "self-organizing systems". They are equipped with very substantial initial structures; —they are provided with many built-in "innate ideas".

Because of this, some readers might object that although these programs learn, they do not significantly "learn to learn". Is this a serious objection? We do not think so, but the question is really one of degree and we are still much too uncertain about it to take a decisive position. In one view learning to learn would be an extremely advanced problem compared to what we now understand. In another view it is just one more problem about certain kinds of program-writing processes, not strikingly different from the static structural situations we already understand rather well. Our position is intermediate between these, at present.

We think that learning to learn is very much like debugging complex computer programs. To be good at it requires one to know a lot about describing processes and manipulating such descriptions. Unfortunately, work in Artificial Intelligence has not, up to now, been pointed very much in that direction, so today we have little real knowledge about such matters.

Consequently we are in a poor position to estimate how complex must be the initial endowment of intelligent learners—ones that could develop as rapidly as human minds rather than requiring evolutionary epochs. We certainly cannot assume from what we know that the "innate structure" required must be very, very complex as compared to present programs. It might be much simpler.

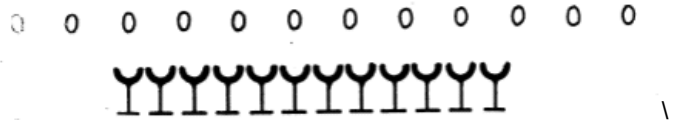
Even in the case of humans we have no useful guidelines. There is probably enough potential genetic structure to supply large innate behavioral programs but no one really knows much about this, either, at present. So let us proceed, instead, to

discuss our present understanding. We begin with some experiments on natural intelligence.

4.1 Learning and Piaget's Conservation Experiments

A classical experiment of Jean Piaget shows remarkably repeatable patterns of response of children (in the age range of 4-7 years) to questions about this sort of material:

Question: "Are there more eggs or more egg-cups?"
Typical Answer: "No, the same."



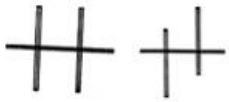
Question: "Are there more eggs or more egg-cups?"
Typical Five-Year-Old's Answer: "More eggs."
Typical Seven-Year-Old's Answer: "Of course not!"

Further questioning makes it perfectly clear that the younger child's comparison is based on the greater "spread" or space occupied by the eggs. The older child ignores or rejects this aspect of the situation and is carried along by the "conservationist" argument: before we spread them out there were the same number of eggs and egg-cups; we neither added or subtracted any, so the number must still be the same. Before constructing a theory of this we describe some other situations that are similar; nothing is more dangerous than to base a theory on just one example and we want the reader to have enough material to participate and, amongst other things, make rival theories.

Here is another relatively repeatable experiment. One shows the child three jars.

He agrees that the first two contain the same amount of liquid. Then, before his eyes, we pour the second jar into the third and ask again about the amounts. Usually, the younger child will say that the tall jar contains more; the older child says, *"Of course they have the same amount. It is the same water so it could not have changed."*

If we perform the pouring behind a screen, telling him what we are doing without his seeing it, the younger child also may say the amounts are the same, but may change his mind when he sees it.



In this experiment, younger children agree the rods are equal at first, but when displaced as shown at the right the "upper" one is usually said to be longer.

How can we explain the difference between the less and more mature children? We see two problems here from the point of view of learning. First, how is the pre-conservationist view acquired (and executed); then how is it replaced by a conservationist one? To many psychologists only the second seems interesting. This is because it is tempting to explain the earlier response in terms like "the child is carried away by appearances," or "the child is dominated by its perception," that is, instead of logic. The usual interpretation, then, is that the transition requires the development of some sort of reasoning capacity that allows it to "ignore the appearance" in favor of reasoning about "the thing itself".

There are serious problems with this view, we feel. First, the "appearance" theory is too incomplete; the notion of appearance is not structured enough. Second, we know that much younger children are quite secure (in other circumstances) about the properties of "permanent objects"; they are sufficiently surprised by magic that there is no reason to suppose they lack the required "logic". We do not think they lack any really basic or primitive intellectual ingredient; rather, they lack some particular kinds of knowledge and/or procedures that are appropriate here. Our view is most easily explained by proposing a more detailed mini-theory for the performance of the non-conservation child.

Behind the "appearance" theory lies some sort of assumption that the water in the tall jar, the upper one of the rods, and the spread-out eggs appear to be "more" than their counterparts, because of some basic law of perception. We think things are more complicated than that, and postulate that the younger child, when asked to make a quantitative comparison, CHOOSES to describe the things being compared in terms of "how far they reach, preferably upwards or in some other direction if necessary". That this description comes from a choice is clear from the fact that he can reliably tell which is "wider" or "taller", when it is not a question of which is "more". Indeed, if we asked the younger child to describe the situation in detail BEFORE asking which has more, he might say something like this:

(A) "There is a tall, thin column of water in the tall, thin jar and a short, wide column in the short, wide jar"

Actually, a four year old will not say anything of the sort. His syntactic structure will not be so elaborate, but more important, he is unlikely to produce that many descriptive elements in any one description. If we ask him "what is this", he might say any of "high glass", "almost full", high water", "round", etc., depending on what he imagines at the moment as a purpose for the question or the object. In any case, if we ask him for a description AFTER telling him we want to know which has more

he will probably say the equivalent of:

(B) "There is a high column of water in the tall jar and a low column of water in the short jar"

To answer the question "which has more" one has to apply some process to the description of the situation. Once we have the second description (B) almost any process would choose the "high column of water". We still need a theory of what symbolic rules delete preferentially the horizontal descriptive elements from the first description (A).

Another possibility is that perhaps the child is misinterpreting "more"; if he were strongly "motivated" by being thirsty or hungry he might give better answers. The experiments are, however, always careful about this, and one gets similar results if the eggs are replaced by candy actually to be eaten, or the water by a delicious beverage.

In suggesting that the child converts description "A" to description "B" we are proposing an analogy with ANALOGY! Is this too neat? Are we inventing this process for the child, who does not really do anything so simple? Certainly, we are making a mini-theory much simpler than what really happens. But what really happens is, we believe, correspondingly simpler than what most observers of children imagine is happening! The following kind of dialog is typical of what goes on in another situation that Piaget and his colleagues have studied, and illustrates explicitly the same striking kind of transformation of descriptions:

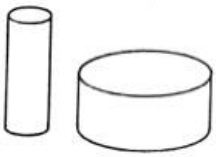
INTERVIEWER: How many animals are there?
CHILD: Five. Three horses and Two cows.
INTERVIEWER: Are there more horses or more animals?
CHILD: More horses. Three horses and two animals.
I: Now listen carefully: ARE THERE MORE HORSES OR MORE ANIMALS?
I: What did I ask you?
C: Are there more horses or more animals?
I: What is the answer?
C: More horses.
I: What was the question again?
C: Are there more horses or more cows?

We explain this phenomenon on a similar basis; again the child has to make a comparison of quantity. He has learned that it is generally correct to do this by counting mutually exclusive classes and the worst thing is to count anything more than once. So he proceeds to describe the situation "correctly" for such purposes, and (in this frame) gets the correct answer. It is often said that the pre-conservation child gets the answer wrong to "inclusion" questions. No. He gets the answer right. He gets the QUESTION wrong! Of course, inclusion comparisons are never natural, so we can agree with the child that these are silly "trick" questions, anyway.

Returning to judging "amount" by height alone, we must ask what "learning" process could cause a child to acquire this "false" idea?

Our mini-theory begins not by trying to explain the particular fact (why the child says this about water or that about eggs) but to look for a general rule for comparing quantities that combines simplicity with widespread utility. Who is bigger; the child or his cousin? Stand back to back! How do you divide a bottle of coke between two glasses? By the level—and generally this is fine because the glasses are identical. Finally, the child can afford to be wrong some of the time; this rule serves very well for many purposes and it would be hard to find a better one without taking a giant step.

A confirmation of this is given by the children who judge there is more water in the thinner container of this pair.



Although fewer children will say this, the fact that there are *any* of them who do disproves the "appearance" theory, for one can hardly maintain that an unalterable law of perception is operating here.

Clearly the (heuristic) symbolic rule of vertical extent here overrides "perception" of dimensions.

One could make a case for the "appearance" theory, in the water-jar experiment as follows: The water is MUCH higher where it is high, but only somewhat wider where it is wide. The most plausible kind of comparison algorithm would look first for a unique term or quality upon which to base its decision—as is easily found in (B).

If there is none—as in (A)—then a subprocess has to make a "quantitative" comparison. But even this seems less symbolic than quantitative, for if we compare "much higher" with "somewhat thinner", the former will surely win! In any case, even

adults can hardly believe that these two solids could have the same volume. So if the child were really faced with the problem of comparing quantitative dimensions, this would be almost impossible for him.

We next have to ask, how was this rule acquired, and how can we explain the transition to conservationist thinking? The simplest theory would assert that the child specifically learns each conservation (and, earlier, each comparison technique) as isolated pieces of knowledge. However, this theory is incomplete because it postulates some agent or specific circumstance responsible for the specific act of learning A more satisfactory kind of theory would let the child himself play the part of the "teaching agent" in the weak theory, and find his own strategies for making descriptions adequate for his problems.

Consider again the original conservation-of-number experiment. Suppose that we wanted to TELL the child how to behave. An authoritarian approach would shout at him: no, no, no, they are equal. But most teachers would prefer the gentler approach of explaining what he is doing wrong. One could say: "Yes, you are right, the eggs take up more space than the egg-cups so you could say that SPATIALLY there are more eggs; but NUMERICALLY there are still as many eggs as egg-cups."

We hope readers are objecting that no child of five will understand this little speech. Indeed, one can go a step further and say that the attempted lesson begs the entire question. The non-conservation child seems to lack a sharp distinction between "numerical" and "spatial". That's his problem! If he knew how to use the distinction well enough he would not need us to teach him about conservation. Our child has already a variety of concepts about quantities; we maintain that his problem is in knowing which to use when (instead of, or combined with others) in describing situations. His real problem is that he does not yet know good enough ways to describe his descriptors! If he learned how to describe his descriptors—for example, to label some as "spatial" and some as "numerical"—and if he could use these descriptions of descriptors to choose the appropriate ones, then the specific problem of learning conservations would dissolve away. As it should! For "conservation" is not a single thing, and "its development is typically spread out over several years as a child learns to deal with number, mass, volume, and other descriptive concepts.

Assuming a structure for classifying descriptions we can imagine an internal scenario, for the egg experiment, in which many descriptions are considered by a supervising process:

Choose a kind of rule. Choices are QUANTITATIVE vs. HISTORICAL RULES
QUANTITATIVE is chosen.
Select a kind. Choices are SPATIAL vs. NUMERICAL RULES
SPATIAL is chosen.
Select a kind. Choices are EXTENT (implies more) or SPARSENESS (implies less)
Try EXTENT. The spread out eggs have more extent. This means MORE.
Test for coherence with other SPATIAL rules.
Try SPARSENESS. The eggs are sparser. This means LESS!
An inconsistency. Reject method.
Instead, Try NUMERICAL. Try COUNTING. Too many to count.
Reject method.
Reject choice of quantitative rules!
Try the next choice, HISTORICAL.
When HISTORICAL is tried, one might first choose IDENTITY.
Some eggs were moved, but none added or taken away. This means SAME!
Test for coherence with other HISTORICAL rules. Try REVERSIBILITY.
The operation SPREADING-OUT is reversible. This means SAME!
We conclude that HISTORICAL seems consistent!

The same sort of scenario could be constructed for the water experiment; there the counting descriptions cannot be invoked, but instead other quantitative descriptions must be available. In each attempt, the description of the scene takes on a different form: the successful historical form will resemble

"The water that was in the second jar is now in the third jar"

and "of course" it has the same amount as the first jar! Well! This gives the right answer, because he has obtained an adequate description. What kinds of processes must the child have in order to do this? We have already proposed that he has a procedure for selecting descriptions; in what kind of environment could this operate? One kind of model would assume that the mature child's description is at first more elaborate, including both geometric and historical elements,

"The amounts of water in the first and second jars were equal. The water that was in the second jar is now in the third jar. The water in the third jar is higher and thinner than that in the first jar."

The mature child, we might theorize, will eliminate elements from his description until there are no serious conflicts. This will yield a tentative answer, which he can maintain if he can explain away any problems that arise from reconsidering other details. Alternatively, one might imagine a process that begins with a very primitive description and elaborates it. But in any case, the process must have facilities for such functions as:

Choosing among the most plausible methods for answering the question. To apply a method he must bring the description into a useable form. For example, when he chooses a "history" method he suppresses some features of the spatial appearance. This means he must have a good classification of the different kinds of description elements.

The selection of the description involves common-sense knowledge. This, in a word, means that his entire cognitive structure is potentially engaged—language, goals, logic, even interpersonal situational processes.

If the situation is at all novel, then any commitment to "ignore" a class of elements may require a reason or "excuse", for conflicts in the original description that remain unexplained. A standard strategy is "compensation"—knowing when it is reasonable to propose tradeoff between such pairs as height and width when manipulating fluids.

One cannot balance an arbitrary pair of dimensions, and particular pairs compensate only under suitable conditions. Ideas like "geometric property" are necessary, so that one isn't tempted to trade height with color, for example. What features of histories might correspond to such static properties as "spatial" and "numerical?"

Most important, the directing process in which the history of the situation wins out over the unusable geometric features, must exist and be debugged well enough that it can be relied upon! The child needs to have and trust the higher-order knowledge about which kinds of knowledge should have priority in each situation.

We have intentionally not specified the time scale of this scenario; some of it occurs over long periods, while some in the course of solving a particular problem. Furthermore, these conditions are still incomplete, yet our structure is already quite complicated. But so is the situation! Remember, our child can already carry on an intelligent conversation. This is not a good place to encourage the use of Occam's Razor. The time for that is when one has several good competing theories, not before one has any! It takes the child several years to work out all of this, and a theory that explained it away on too simple a basis might be therefore suspect! We do not, we repeat, want to explain the different conservations either on completely separate bases or by one unifying principle. We want to see it as the outcome of an improvement in the child's procedures for dealing with the variety of descriptions that he comes into possession of.

In the traditional "theories of learning" there was a tendency to ask

"How does such-and-such a "response" become connected to such-and-such a "stimulus"?

We now see that the proper questions are much more like

"How can such-and-such a procedure be added to the descriptive or deductive systems"?

4.2 LEARNING

A serious complaint about the heuristic programs of the past was their very limited ability to learn. This made them too inflexible to be useful except in very special situations. Over the years many direct attempts to construct "learning programs" led to very indifferent results. There is a close analogy, we feel, between this and the similar situation in the history of constructing psychological theories of learning.

If a child were to learn that 7+5=12 and 39+54=93 and, say, one hundred other such "responses", we would not agree he had learned to add. What is required is that he learns an appropriate procedure and how to apply it to numbers he has never used before. Another side of this "stimulus-response" problem: just as in the Analogy situation, the secret of learning often lies in the discovery of descriptions that emphasize the "essential" aspects of things or events, and omit or subjugate the "accidental" features. It would do us little good to remember that some particular thing happened in exactly a certain situation, since identical conditions never recur.

We do not need, or want, to remember the precise details of a broken chair, but we do want to remember that bad things happen when chairs have broken rungs—for that is an essential difference between this and a usable chair. Indeed, the greater our knowledge and powers of observation, the more selective must be our choice of descriptions, because of the magnified problem of becoming lost in searching through networks of irrelevant details.

Finally, one hears complaints of the form "You programmed it to do that! It didn't learn it by itself"! There is a spectrum of degrees of autonomy in learning activities, and one wonders what are the distinctive features of importance between a child learning while playing by himself, discovering things under the shrewd guidance of an attentive instructor, prying a theory out of a mediocre textbook, and having it explained directly and concisely by a superb expositor.

It is tempting to try to disentangle this messy web of different phenomena. The appearance of an impossibly refractory problem in science is often the result of fusing fundamentally different problems (each of which may be relatively simple) when there is no common solution to the whole set. We think this is true of the many different ways in which programs can be said to learn. But despite this diversity there are important common themes. Most important of these, we feel, is the need for enough descriptive structure to represent the relation between learning situations and the concepts learned from them. Another theme comes from noticing that the kinds of learning we have found most difficult to simulate are those that involve a large stock of prior knowledge and analytic abilities. This leads us to propose for study very pure forms of the problem of handling divers kinds of knowledge—prior to worrying about the problems of acquiring such knowledge. To separate out these strands we will consider at various points such not-entirely-separable ideas of "learning" as these:

Learning by development or maturation
Learning without description (by quantitative adaptation)

Learning by building and modifying description
Learning by being taught
Learning by Analogy
Learning by being told
Learning by being programmed
Learning by Understanding

4.3 Incremental Adaptation.

There is a large literature concerned with clustering methods, scaling, factor analysis, and optimal decision theories, in which one find proposals for programs that "learn" by successive modifications of numerical parameters. An outstanding example of this is seen in one of the well-known programs of A. Samuel, that plays a good game of Checkers. Other examples abound; all perceptron-like "adaptive" machines, all "hill-climbing" optimization programs, most "stochastic learning" models using reinforcement. Some details can be found in the later chapters of our book, PERCEPTRONS.

The conclusions drawn in *Perceptrons* are too technical to review here in detail, but we can describe the general picture that emerges. Within the classes of concepts that these machines can represent, that is, describe as rather literal "sums" of already programmed "parts"—the learning abilities are effective and interesting. However, the descriptive powers of these quasi-linear learning schemes have such peculiar and crippling limitations that they can be used only in special ways.

For example, we can construct, by special methods, a perceptron that could learn either to recognize squares, or to recognize circles. But the same machine would probably not be able to learn the class of "circles or squares"! It certainly could not describe (hence learn to recognize) a relational compound like "a circle inside a square".

These limitations are very confining. It is true that such methods can be useful in "decision-making" and diagnostic situations where things are understood so poorly that a "weighted decision" is better than nothing! But we think it might be useful to put this in perspective by assigning it as an example of a new concept of TERMINAL LEARNING.

The basic problem with this kind of "learning program" is that once the program has been run, we end up only with numerical values of some parameters.

The information in such an array of numbers is so homogeneous and unstructured—the "weight" of each "factor" depends so much on what other factors are also involved in the process—that each number itself has no separate meaning. We are convinced that the results of experience, to be useful to "higher level processes", must be summarized in forms that are convertible to structures that have at least some of the characteristics of computer programs—that is, something like fragments of program or descriptions of ways to modify programs.

Without such capabilities, the simple "adaptive" systems can "learn" some things, to be sure, but they cannot learn to learn better! They are confined to sharpening whatever "linear separation" or similar hypotheses they are initially set to evaluate. A terminal learning scheme can often be useful at the final stage of a performance or an application, but it is potentially crippling to use it within a system that may be expected later to develop further.

One could make similar criticisms of another aspect of the adaptive "branch and bound" procedures found in most game playing and other heuristic programs that follow the "look-ahead" and "minimax" tradition. Suppose that in analyzing a chess position we discovered that a certain square is vulnerable to a rook-queen fork by moving a knight to that square. The traditional program returns only a certain numerical value to that move. What it really should do is return a description of this feature of this position. Then the previous plausible-move generator can be given a constructive suggestion: look for moves that add a defense to that square, or a counter-attack, or moving one of the threatened pieces, etc. Then subsequent exploration will discover more such suggestions.

Eventually, these conditions may come to a logical conflict, e.g., by requiring a piece to attack two squares that cannot both lie in its range. At this point, a deductive program could see that it is necessary to think back to an earlier position. Similarly, a description of that situation, in turn, could be carried further back, so that eventually the move generator can come to work with a knowledgeable analysis of the strategic problem. Surely this is the sort of thing good players must do, but no programs yet do anything much like it.

This argument, if translated into technical specifications, would say that if a chess program were to "really" analyze positions it must first have descriptive methods to modify or "update" its state of knowledge. Then it needs ways to "understand" this knowledge in the sense of being able to make inferences or deductions that help decide what experiments next to try. Here again, we encounter the problem of "common sense" knowledge since although some of this structure will be specific to chess, much also belongs to more general principles of strategy and planning.

People working on these homogeneous "adaptive learning" schemas (either in heuristic programming or in psychology) do recognize this kind of problem. Unfortunately, most approaches to it take the form of attempting to generalize the coefficient-optimizing schema directly to multi-level structures of the same kind, such as n-layer perceptrons. In doing so, one immediately runs into mathematical problems: no one has found suitably attractive generalizations (for n levels) of the kinds of convergence theorems that, at the first level, make perceptrons (for example) seem so tempting. We are inclined to suspect that this difficulty is fundamental—that there simply do not exist algorithms for finding solutions in such spaces that operate by successive local approximations. Unfortunately we do not know how to prove anything about this or, for that matter, to formulate it in a respectably technical manner.

We could make similar remarks about most of the traditional "theories of learning" studied in Psychology courses. Almost all of these are involved with the equivalent of setting up connections with the equivalent of numerical coefficients between

"nodes" all of the same general character. Some of these models have a limited capacity to form "chains of responses", or to cause some classes of events to acquire some control over the establishment of other kinds of connections. But none of these theories, from Pavlov on, seem to have adequate ability to build up processes that can alter in interesting ways the manner in which other kinds of data are handled. These theories are therefore so inadequate, from a modern computation-theory view, that today we find it difficult to discuss them seriously.

Trial and Error

Why, then, have such theories been so persistently pursued? Their followers were certainly not naive about these difficulties. One influence, we think, has been a pervasive misconception about the role of multiple trials, and of "practice", in learning. The supposition that repeated experiences are necessary for permanent learning certainly tempts one to look for "quantitative" models in which each experience has a small but cumulative effect on some quantity, say, "strength-of-connection".

In the so-called "stimulus-sampling" theories we do see an attempt to show how certain kinds of one-trial learning processes could yield an external appearance of slow improvement.

In this kind of theory, a response can become connected with many different combinations of stimulus features or elements as a result of a sampling process. In each learning event a new combination can be tried and tested. This is certainly closer to the direction we are pointing. However, we are less interested in why it takes so many trials to train an animal to perform a simple sequence of acts, and more interested in why a child can learn what a word means (in many instances) with only a single never-repeated explanation.

What is the basis for the multiple-trial belief?
When a person is "memorizing" something he may repeat it over and over.
When he practices a piece of music he plays it over and over.
When we want him to learn to add we give him thousands of "exercises".
When he learns tennis he hits thousands of balls.

Consider two extreme views of this.

In the NUMERICAL theory he moves, in each trial, a little way toward the goal, strengthening the desired and weakening the undesired components of the behavior.

In the SYMBOLIC view, in each trial there is a qualitative change in the structure of the activity—in its program.

Many small changes are involved in debugging a new program, especially if one is not good at debugging! It is not a matter of strengthening components already weakly present so much as proposing and testing new ones.

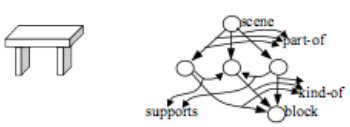
The external appearance of slow improvement, in the SYMBOLIC view, is an illusion due to our lack of discernment. Even practicing scales, we would conjecture, involves distinct changes in one's strategies or plans for linking the many motor acts to already existing sequential process-schema in different ways, or altering the internal structures of those schemas.

The improvement comes from definite, albeit many, moments of conscious or unconscious analysis, conjecture, and structural experiment. "Thoughtless" trials are essentially wasted. To be sure, this is an extreme view. There are, no doubt, physiological aspects of motor and other types of learning which really do require some repetition and/or persistence for reliable performance.

Our point is that the extent of this is really quite unknown and one should not make it the main focus of theory-making, because that path may never lead to insight into the important structural aspects of the problem. In motor-skill learning, for example, it is quite possible one needs much less practice than is popularly supposed. It takes a child perhaps fifteen minutes to learn to walk on stilts. But if you tell him to be sure to keep pulling them up, it takes only five minutes. Could we develop new linguistic skills so that we could explain the whole thing? We might conjecture that the "natural athlete" has no magical, global, coordination faculty but only (or should we say "only"!) has worked out for himself an unusually expressive abstract scheme for manipulating representations of physical activities.

4.4 Learning by building descriptions.

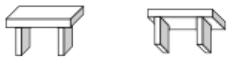
We can illustrate much more powerful concepts of learning in the context of a procedure developed by P. Winston to learn to recognize simple kinds of structures from examples. Like the SEE program of Guzman (which it uses as a sub-process) it works in the environment of children's' building blocks. When presented with a scene, it first observes relations between features and regions, then groups these to find proposed structures and objects, and then attempts to identify them (using description-matching methods and the results of earlier learning experiences). Thus, the simple scene on the left is described by a network of abstract objects, relations, and relations between relations.



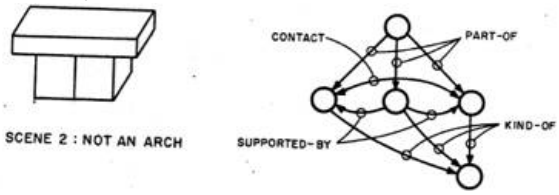
In this diagram, the heavy circles represent particular physical objects, the other circles represent other kinds of concepts, and the labels on the arrows represent relations. The program is equipped from the start to recognize certain spatial relations such as contact, support, and some other properties of relative position. We tell the machine that this is (an

example of) an ARCH, and it stores the description-network away under that title.

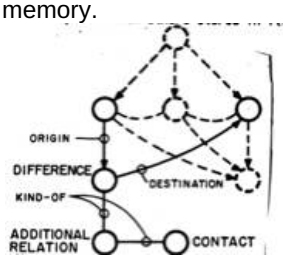
Note that since these properties describe only relative spatial relations, the very same network serves to describe both of these figures, which are visually quite different but geometrically the same.



Next we present SCENE 2, to the left below, and the machine constructs the network shown to its right.



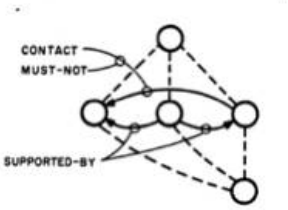
This differs from the network of scene 1 in only a few respects. If the program is asked what this structure "is," it will compare this description with others stored in its memory.



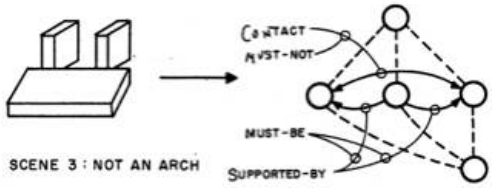
[This picture shows an arch with its two supports in contact, so that there's no hole between them.]

It already has networks for tables, towers, and a few other structures but, as one might expect, the structure it finds most similar is the ARCH description stored just a moment ago. So it tentatively identifies this as an arch. In doing this, it also builds a descriptive network that describes the difference between scene 1 and scene 2, and the difference is represented somewhat like this.

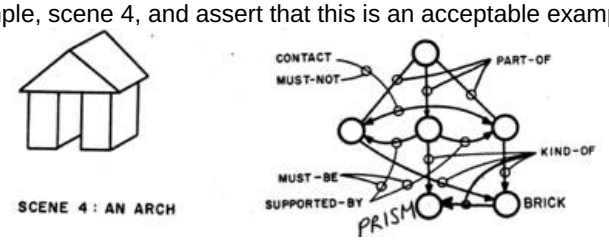
Now we tell the machine that scene 2 is NOT an example of an ARCH. It must therefore modify its description of "ARCH" so that structure 2 will no longer match the description, hence will no longer be "seen" as an ARCH. The method is to add a "rejection pointer" for the contact relation.



Now for the next example: we present scene 3 and assert that this, too, is not an ARCH. The most prominent difference, in this case, is that the new structure lacks the support relations—and the program for modifying "ARCH" now adds an "enforcement pointer" to the support relations.



Finally, we present another example, scene 4, and assert that this is an acceptable example of an ARCH.



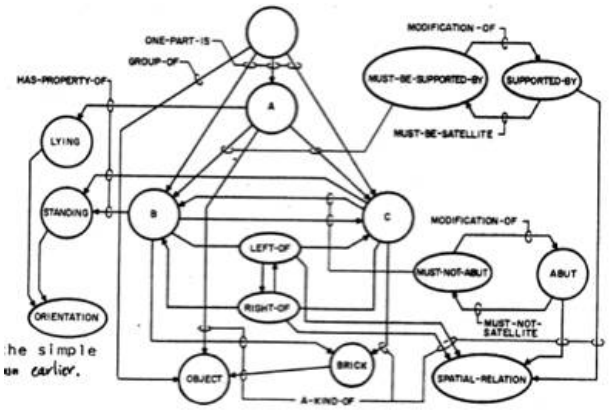
The most important difference, now, is the shape of the top block. The machine has to modify the description of "ARCH" so that the top block can be either a brick or a wedge. One strategy for this would be simply to invent a new class of objects -- "brick-or-wedge." This would be extremely "conservative", as a generalization or explanation. Winston's strategy is to look in memory for the smallest class that contains both bricks and wedges. In the machine's `present state the only existing such classes are "prism" and "object" -- the latter is the class of all bodies, and includes the "prism" category, so the new description will say that the top object is a kind of prism. If we replaced the wedge by a pyramid, and told it that this, too, is an arch, it would have to change the top object-description to "object," because this is the smallest class containing "brick" and "pyramid." Now we can summarize the program's conclusion: an arch is

"A structure in which a prismatic body is supported by two upright blocks that do not touch one another."

We have just seen how the program learns to identify correctly the membership of Scenes 1-4 as to whether they are ARCHes or not. As a consequence, it will probably "generalize" automatically to decide that these are also arches, because there are no "must-be-a..." enforcement pointers to either the supports or the top.

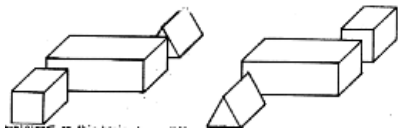
Of course this judgment really depends on the machine's entire experience, i.e., on what concepts are already learned, and upon details of the comparison programs.

We have suppressed many interesting details of the behavior of Winston's program, especially about how it decides which differences are "most important". For example, the final form of the network for "ARCH" is more like this,



than like the simple schemata shown earlier.

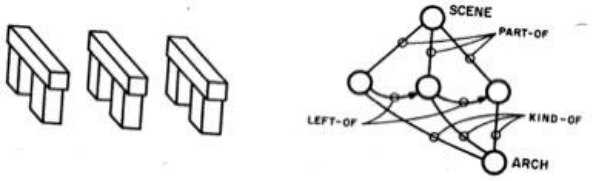
While on the subject, it should be noticed that within the network are represented relations between relations, as well as objects, properties and simple relations. There are important advantages to this when it comes to construction of the difference-descriptions. If the comparison program can be told that the difference between "IN-FRONT-OF" and "BEHIND," as well as that between "LEFT-OF" and "RIGHT-OF," can both be described in terms of "vertical axis symmetry," then it can be programmed to observe that all (high-level) differences between the two scenes below can be "explained" on the basis that they differ only in respect to a vertical axis rotation.



This is an example of a beautifully abstract form of description manipulation that, psychologically, would traditionally be attributed to something much more like an imaginary graphical rotation of the scene -- (as though there were no critically complicated problems in that reconstruction). In his thesis, Winston has only initiated such studies, and we know little about how far one can go with these methods. How much more structure would one need, to be able to learn, from examples, such concepts as symmetry? How difficult will it be to adapt such a system to learning new procedures, instead of structures? At first this might seem a huge step, but the ideas in the next section, on describing groups and repetitive structures, make the gap seem to become smaller. We shall see that the advantages of having a description for a "concept" (rather than just a competence) are absolutely crucial for further progress. These advantages include:

> The ability to compare and contrast descriptions (as we shall see in section 4.6)
> The ability to make deductions involving the concept, to adapt it to new situations.
> Combining several descriptions to make new concepts.

An example of the latter: Every structural "concept" that Winston's program acquires is automatically incorporated within its own internal descriptive mechanisms. Thus, if the machine were presented with the nine-block scene below, before learning a concept of ARCH, it would have produced an impossibly complex and almost useless network of relations between the nine blocks. But after learning ARCH, it will now describe it in a much more intelligent way,



because its descriptive mechanisms can proceed from local to global aggregates using as much available knowledge as it can apply. In doing this we encounter, now on a higher level, grouping problems very much like those we saw in our sketch of Guzman's SEE program, and in many cases one can adopt analogous strategies.

4.4 Learning by being taught.

Imagine a child playing with a toy car and his blocks. He wants to build an interesting structure to play with. If the user of Winston's program were present, he could teach the child how to make an arch by the process just described, for it is not hard to convert the above description into a procedure for building arches. In fact, in chapter 5, we shall give a sketch of exactly how this can be done! This is precisely what Winograd's program does when it translates from the semantic analysis of an object-describing noun-phrase into a robot program for building with blocks! See chapter 5.

It is not necessary for the child to have a teacher, however. In the course of "playing" he can try experiments with the blocks and the car, and he can recognize "success" in either of these cases, among others:

a) He knows how to recognize an "ARCH" once it is built—but does not know how to describe or to build it.
b) He has a functional play-goal: construct a road-problem for himself that is not too easy and not too hard—such as an obstacle that requires two hands to overcome, but cannot be negotiated trivially with one hand.

In case (a), the child knows how to tell which structures are in the class.

In case (b), while experimenting the child will indeed find that Scene 1 is good, Scene 2 is impossible, Scene 3 is too easy, and Scene 4 (discovered as the simplest variant of the successful Scene 1) is also good. Here we get the same overall effect—through the same mechanism—yet in humanistic terms the behavior would be described much more naturally in terms of "exploratory," or "play," or "undirected" activity. The result, if described in structural terms, is again

"A structure in which an object is supported by two upright bricks that do not touch one another."

This is certainly not a perfect logical equivalent of the adult's idea of an arch; nor does it contain explicitly the idea of a surrounded passage or hole. Still, for the playing child's purposes, it would represent perhaps an important step toward formulation and acquisition of such concepts.

Again we have left alone some very important loose ends. We have concealed in the catch-all expressions "play" or "exploration" some supremely important conditions that must be fulfilled—and at early stages of child development they won't be, and the things that are learned during "play" will be different!

The child must already be equipped with procedures that have a decent chance of generating plausible structures.

To do this, he must be able to describe to some extent why an experiment is unsatisfactory. If he cannot get his car between the supports, he must be able to think of moving the supports apart. This is not very hard, since pushing against the obstacle will sometimes do this.

Since most experiments not carefully planned lead to useless structures, he has to have some ability to reconstruct a usable version of earlier and better situations after a disaster.

Without the teacher, it is unlikely that he will get good results after just four trials! He must have enough persistence in his goal-structure to carry through. To do this consistently would presuppose a good assessment of the problem's difficulty. Of course, if this is missing, he will find something else to do; not all play is productive!

Winston's program seems to be a reasonable model for kinds of behavior that would be plausible in, if not typical of, a child. The "concept" the program will develop, after seeing a sequence of examples, will depend very much on the examples chosen, on the order in which they are presented, and of course on the set of concepts the program has acquired previously. In many cases the experimenter may not get the result he wants; presenting examples in the wrong order could get the program (or child) irreparably off the track, and he might have to back up—or perhaps restart at an earlier stage. We cannot expect our concept-learning programs to be foolproof any more than a teacher can expect an instructional technique always to work—without some insights into what is happening in each student's mind.

Of course there are many small but important details of how the program decides what to do at each step. Which differences should have high priorities? Which parts of the description networks should be matched? What explanations should it assign to each difference that is noticed?

Thus, in building with blocks, the relations "support" and "contact" ought to dominate properties of color, particular shapes and even other spatial relations like "in front of" or "to the right of."

In a different realm of activity, a different set of priorities might be essential, lest learning be slow or simply wrong. So, one can conclude that we must also develop intermediate structures in "learning to learn; a prerequisite to a child's (or machine's) mastery of mechanical structures will be some preparation in acquiring, grouping, and interrelating the more elementary descriptive structures to be used in assembling, comparing and modifying the representations to be used in the performance-level learning itself. This is exactly the conclusion we reached, in 4.1, about the requirements implicit in "maturation".

4.6 Analogy, again}

Now we can return to our very first topic, solving problems involving analogies. In section 1.1 we proposed that the key idea would lie in finding ways to describe changes in descriptions.
But this is exactly what happens in the program we have just described. When asked to describe a new scene situation, Winston's program makes use of the other descriptions it remembers, so that it can describe the scene in terms of already-learned concepts. Although we have not explained in detail how this is done, it is important to mention that the result of comparing two descriptions, in this system, is itself a description! Basically, the comparison works this way:

1. The two descriptions are "matched together", using various heuristic rules to decide which nodes probably correspond.

2. We create a new network, whose nodes are associated with pairs of nodes from the two descriptions that were matched.
This is the skeleton of the comparison-description.

3. We associate with each node of this skeleton, a "comparison note" describing the correspondence. If the descriptions immediately local to two "corresponding" nodes are the same, the comparison-note is trivial. But if there are differences, (e.g., if one is a brick and the other a wedge) the "comparison note" describes this difference. Since these descriptive elements have the same format as by the original scene descriptions, one can operate upon them with the same programs.

In particular, two difference-descriptions can be compared as handily as any other pair of descriptions.

Now we can apply this idea to the analogy problem. The machine must select that scene X (from a small collection of alternatives) which best completes the statement

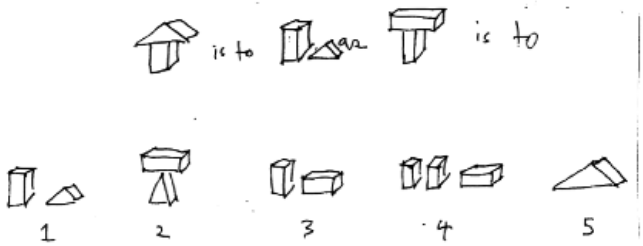
A is to B as C is to X

That is, one must find how B relates to A and find an X that relates to C in the same way. Using the expression *D[A:B]* to denote the difference-description-network resulting from comparing A with B, we simply compare the structures resulting from:

D[D[A:B]:D[C:X1]],
D[D[A:B]:D[C:X2]],
D[D[A:B]:D[C:X3]], etc.

Each of these summarizes the discrepancies within the "analogical explanations" for each corresponding possible answer. So to make the decision, we have to choose the "best" or "simplest" of these. We will not give details of how this is done; it is described in Chapter 7 of Winston's thesis. But note that some such device was needed already for the basic ability to identify a presented scene most closely with one of the descriptive models in memory. Thus the program must incorporate, in its comparison mechanism, conventions and priorities about such matters as whether the difference between Right and Left is to be considered simpler than the difference between Right and Above.

In this example

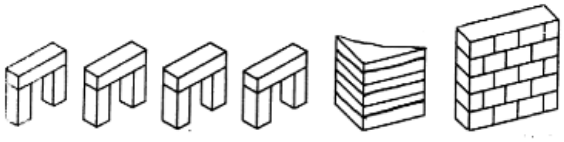


the machine chooses THREE as its answer, ONE as its second choice. In the slightly altered problem

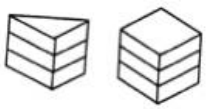
It chooses FOUR as its answer.

4.7 Grouping and Induction

The problem of recognizing or discerning grouping or clusters of related things is another recurrent concern not only in Psychology, but also in statistics, artificial intelligence, theory of inductive inference; indeed, of science and art in general. Most studies of "clustering" have centered around attempts to adapt numerical methods from the theory of multivariate statistics to group data into subsets that minimize some formula which compares selected inter- and intra- group measures of relatedness. But such theories are not easily adaptable to such important and interesting problems as discerning that this picture shows, not 12 + 6 + 20 = 38 objects, but *"a row of arches, a tower of cubes, and a brick wall."*



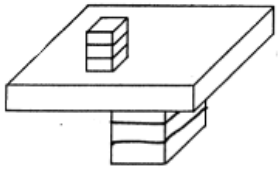
More subtly, how do we "know" that one of these is 'three wedges' while the other is 'three blocks'?



Visually, the lower objects in each tower are the same. These problems, too, can be treated by the same general methodology used in our approach to Analogy and to Learning of structures in scene-analysis.

On many occasions we have been asked why the AI Laboratory is so concerned with special problems like machine vision, rather than more general approaches and problems about intelligence. In the early stages of a new science one proceeds best by gaining a very deep and thorough understanding of a few particular problems; that way one discovers important phenomena, difficulties, and insights, without which one risks fruitless periods of speculations and generalities. If the reader can see the present discussion in terms of general problems about induction and learning, the fruitfulness of the approach should speak for itself; we cannot imagine anyone believing the usefulness of these ideas is in any important way confined to description of visual or mechanical structures!

Take the groupings in the preceding figures and ask: "What qualities of the scene-descriptions characterize the intuitively acceptable groups." In some groups, like those shown above, it seems clear that the important feature is a CHAIN, say, of supported-by or in-front-of relations. In other cases it seems obvious that several objects show a common relationship to another. But no simple rules work in all situations.

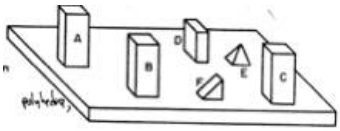


In this scene one does not usually see a single group or tower of seven blocks. Whether it is appropriate to describe this as "a seven-block stack," or as "a three-block stack supporting a plate that in turn supports a three-block stack," or as yet something else, depends on one's current purposes, orientations, or specifically on what grouping criteria are currently activated for whatever reason.

In some situations the discrepancies in the individual properties of the blocks should cause the grouping procedure to separate out the three-block stacks in spite of the fact that the support-relation chain continues through all seven blocks.

We next summarize some experiments along this line, again reporting results from P. Winston's dissertation. In Winston's grouping program, a generous hypothesis is followed by a series of criticisms and modifications.

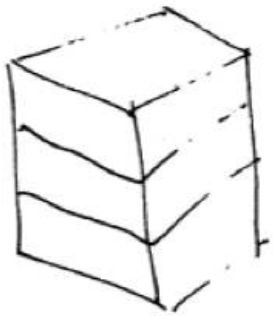
For example, when several objects have the same or very nearly the same description, they are immediately taken as candidates for a group. The blocks on this table are typical. All are blocks, all are standing, and all are supported by the board.



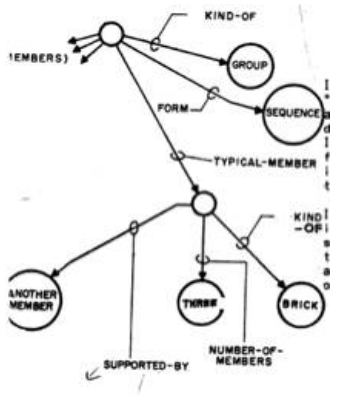
This proposal is then examined to eliminate objects which seem atypical, until a fairly homogeneous set remains. To do this, a program lists all relationships exhibited by more than half of the candidates in the set.

When the procedure operates, the first pass through the loop rejects E and F, mainly on the basis of shape. (Size is not considered in this pass because the six objects are too heterogeneous for "size" to be put on the common-relationships list.) In a second pass, however, more than half the remaining objects share the "medium" size property, and block D is rejected, mainly because it does not share this property. So, finally, the procedure accepts only A B and C into the group. Obviously this is appropriate for some goals, but not others.

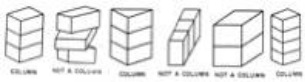
When grouping concepts are injected into the description framework, there can be unexpected and exciting consequences for other problems of induction. The figure below shows the network representation obtained when the grouping process operates on the description of this 3-block column.



Into the description is introduced a "typical member" to which is attributed the common properties discerned by the grouping procedure. In this case, chaining was used to form the group and the description includes the fact that there were three elements in the chain.



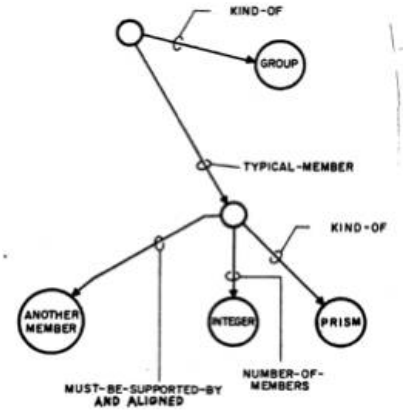
In a learning experiment, the program is presented with the depicted sequence of scenes shown below and is told that the first, third, and sixth are instances of "column" while the others are not.



The second example causes the enforcement of a new pointer, "ALIGNED," a concept already available to the program that refers to the neat parallel alignment of edges. The third example tells the system that the typical member can be a wedge or a brick; the smallest common generalization here is "PRISM" so now a "column" can be any neatly piled stack of prisms.

The fourth example changes "supported-by" to "must-be-supported-by"; the fifth, which is not seen as a group because it has only two elements, changes "one-part-is-a group" to "one-part-must-be a group".

The sixth and final example is of particular interest with respect to traditional induction questions. Comparison of it with the current concept of "column" yields a difference- description whose highest-priority feature is the occurrence of "FOUR" instead of "THREE," in the number-of-members property of the main group. What is the smallest class that contains both "THREE" and "FOUR?" In the program's present state, the only available superset is "INTEGER." Thus we obtain this description of "column", which permits a column to have any number of elements!



Is this too rash a generalization to make from so few examples? The answer depends on too many other things for the question to make much sense. If the program had already some concept of "small integer," it could call upon that. On a higher level we could imagine a program that supervised the application of any generalization about integers, and attaches an auxiliary "warning" pointer label to conclusions based on marginally weak evidence. We are still far from knowing how to design a powerful yet subtle and sensitive inductive learning program, but the schemata developed in Winston's work should take us a substantial part of the way.

Finally, we note that in describing a sequential group in terms of atypical member and its relations with the adjacent members of the chain, we have come to something not too unlike that in programming languages that use "loops," entry, and exit conditions. Again, a structure developed in the context of visual scene-analysis suggests points of contact with more widely applicable notions.

5.0 Knowledge and Generality

We now turn to another set of questions connected with our long-range goal of understanding "general intelligence". An intelligent person, even a young child, is vastly more versatile than the "toy" programs we have described. He can do many things; each program can do only one kind of thing. When one of our programs fails to do what we want, we may be able to change it, but this usually requires major revisions and redesign. An intelligent human is much more autonomous. He can often solve a new kind of problem himself, or find how to proceed by asking someone else or by reading a book.

One might try to explain this by supposing that we have "better thinking processes" than do our programs. But it is premature, we think, to propose a sharp boundary between any of these:

Having knowledge about how to solve a problem,

Having a procedure that can solve the problem,
Knowing a procedure that can solve the problem!

In any case, we think that much of what a person can do is picked up from his culture in various ways, and the "secrets" of how knowledge is organized lie largely outside the individual. Therefore, we have to find adequate models of how knowledge systems work, how individuals acquire them, and how they interact both in the culture and within the individuals. How can we build programs that need not be rebuilt whenever the problems we want to solve are slightly changed? One wants something less like ordinary computer "programming" and more like "telling" someone how to do something, by informal explanations and examples. In effect, we want larger effects while specifying less. We do not want to be bothered with "trivial" details. The missing information has to be supplied from the machine's internal knowledge. This in turn requires the machine itself to solve the kinds of easy problems we expect people to handle routinely --- even unconsciously --- in everyday life. The machine must have both the kinds of information and the kinds of reasoning abilities that we associate with the expression "common sense". There are differences of opinion about such questions, and we digress to discuss the situation.

Artificial Intelligence, as a field of inquiry has been passing through a crisis of identity. As we see it, the problem stems from the tendency for the pursuit of technical methods to become detached from their original goals so that they follow a developmental pattern of their own. This is not necessarily a bad thing; many productive areas of research were born of such splits. Every discipline has had to deal with such situations and it has happened often in the study of human intelligence. Nevertheless, if one is interested in the particular goal of building a science of Intelligence, one has to be concerned with the use of resources both on the local scale of conserving one's own time and energy and on a global scale of watching to see whether the scientific community seems to be directing itself effectively.

We suspect that there is now such a problem in connection with the studies of Mechanical Theorem Proving.

5.1 Uniform procedures vs. Heuristic Knowledge

As a first approximation to formulating the issues, consider a typical research project working on "automatic theorem proving". Schematically, the project has the form of a large computer program which can accept a body of knowledge or "data base," such as a set of axioms for group theory, or a set of statements about pencils being at desks, desks being in houses, and soon. Given this, the program is asked to prove or disprove various assertions. What normally happens is that if the problem is sufficiently simple, and if the body of knowledge is sufficiently restricted in size, or in content or in formulation, the program does a presentable job. But as the restrictions are relaxed it grinds to an exponential stop of one sort or another.

There are two kinds of strategy for how to improve the program. Although no one actually holds either policy in its extreme form and although we encounter theoretical difficulties when we try to formalize them, it nevertheless is useful to identify their extreme forms

The POWER strategy seeks a generalized increase in computational power. It may look toward new kinds of computers ("parallel" or "fuzzy" or "associative" or whatever) or it may look toward extensions of deductive generality, or information retrieval, or search algorithms—things like better "resolution" methods, better methods for exploring trees and nets, hash-coded triplets, etc. In each case the improvement sought is intended to be "uniform" --- independent of the particular database.

The KNOWLEDGE strategy sees progress as coming from better ways to express, recognize, and use diverse and particular forms of knowledge. This theory sees the problem as epistemological rather than as a matter of computational power or mathematical generality. It supposes, for example, that when a scientist solves a new problem, he engages a highly organized structure of especially appropriate facts, models, analogies, planning mechanisms, self-discipline procedures, etc., etc. To be sure, he also engages "general" problem-solving schemata but it is by no means obvious that very smart people are that way directly because of the superior power of their general methods—as compared with average people. Indirectly, perhaps, but that is another matter: a very intelligent person might be that way because of specific local features of his knowledge-organizing knowledge rather than because of global qualities of his "thinking" which, except for the effects of his self-applied knowledge, might be little different from a child's.

This distinction between procedural power and organization of knowledge is surely a caricature of a more sophisticated kind of "trade-off" that we do not yet know how to discuss. A smart person is not that way, surely, either because he has luckily got a lot of his information well organized or because he has a very efficient deductive scheme. His intelligence is surely more dynamic in that he has (somehow) acquired a body of procedures that guide the organization of more knowledge and the formation of new procedures, to permit bootstrapping. In particular, he learns many ways to keep his "general" methods from making elaborate but irrelevant deductions and inferences.

5.1.1 Successive Approximations and Plans

The mechanical theorem-proving programs fail unless provided with carefully formulated diets of data; either if given to little knowledge and asked advanced theorems, or given too much knowledge and asked easy questions. In any case, the contrast with a good mathematician's behavior is striking; the programs seem to have no "global" strategies. If a human mathematician is asked to find the volume of some object of unusual shape he will probably try to use some heuristic technique like:

Thus, one would transform:

Now, in his final "proof" the heuristic principle that was used will not appear explicitly, even though its use was crucial. The

three kinds of information in

The knowledge exhibited in the proof,
The knowledge used to find the proof, and
The knowledge required for "understanding" or explaining the proof so that one can put it to other uses,

are not necessarily the same in extent or in content. The "Theorem Prover" systems have not been oriented toward making it easy to employ the second and the third kinds of knowledge. We have just given an example of how the second type of knowledge can be used.

The third kind of knowledge is exemplified by the following story about an engineer or physicist analyzing a physical system. First, he will make up a fairy-tale:

"The system has perfectly rigid bodies, that can be treated as purely geometric. There is no friction, and the forces obey Hooke's law."

Then he solves his equations. He finds the system offers infinite resistance to disturbance at a certain frequency. He has used a standard plan—call it *ULTRASIMPLE*—which produced an absurdity. But he does not reject this absurdity, completely! Instead, he says: "I know this phenomenon! It tells me that the "real" system has an interesting resonance near this frequency".

Next, he calls upon some of his higher-order knowledge about the behavior of plan*ULTRASIMPLE*. Accordingly, this tells him next to call upon another plan, LINEAR, to help make a new model which includes certain damping and coupling terms.

Next, he studies this system near the interesting frequency that was uncovered by plan*ULTRASIMPLE*. He knows that his new model is probably very bad at other, far-away, frequencies at which he will get false phenomena because of the unaltered assumptions about rigidity; he has reason to believe these harmless in the frequency band now being studied. Then he solves the new second-order equations. This time he might obtain a pair of finite, close-together resonances of opposite phase. That "explains" the singularity in the simpler model.

We abandoned one simple "micro-world" and adopted another, slightly more complicated and better adapted to the better-understood situation. This too may serve only temporarily and then be replaced by a more specialized set of assumptions for studying how nonlinearities affect the fine structure of the resonances: a new plan, *NONLINEAR* or *INELASTIC* or *THIRD-ORDER* or *DISCRETE*, or whatever his third-type knowledge suggests.

One cannot overemphasize the importance of this kind of scenario both in technical and in everyday thinking. We are dependent on having simple but highly developed models of many phenomena. Each model—or "micro-world" as we shall call it—is very schematic; in either our first-order or second-order models, we talk about a fairyland in which things are so simplified that almost every statement about them would be literally false if asserted about the real world. Nevertheless, we feel they are so important that we plan to assign a large portion of our effort to developing a collection of these micro-worlds and finding how to embed their suggestive and predictive powers in larger systems without being misled by their incompatibility with literal truth. We see this problem—of using schematic heuristic knowledge—as a central problem in Artificial Intelligence.

5.2 Micro-worlds and Understanding}

In order to study such problems, we would like to have collections of knowledge for several "micro-worlds", ultimately to learn how to knit them together. Especially, we would like to make such a system able to extend its own knowledge base by understanding the kinds of information found in books. One might begin by studying the problems one encounters in trying to understand the stories given to young children in schoolbooks. Any six-year-old understands much more about each of such crucial and various things as

time space planning explaining
causing doing preventing allowing
failing knowing intending wanting
owning giving breaking hurrying

than do any of our current heuristic programs. Eugene Charniak, a graduate student, is now well along in developing some such models, and part of the following discussion is based on his experiences.

Although we might describe this project as concerned with "Understanding Narrative", —of comprehending a story as a sequence of statements as read from a book—that image does not quite do justice to the generality of the task. One has the same kinds of problems in:

Making sense of a sequence of events one has seen or otherwise experienced (what caused what?)
Watching something being built (why was that done first?)
Understanding a mathematical proof (what was the real point, what were mere technical details?)

Many mental activities usually considered to be non-sequential have similar qualities, as in seeing a scene: why is there a shadow here? —What is that? —Oh, it must be the bracket for that shelf.

In any case, we do not yet know enough about this problem of common sense. One can fill a small book just describing the commonsense knowledge needed to solve an ordinary problem like how to get to the airport, or how to change a tire. Each new problem area fills a new catalogue. Eventually, no doubt, after one accumulates enough knowledge, many new

problems can be understood with just a few additional pieces of information. But we have no right to expect this to happen before the system contains the kind of breadth of knowledge a young person attains in his elementary school years!

We do not believe that this knowledge can be dumped into a massive database without organization, nor do we see how embedding it in a uniformly structured network would do much good. We see competence as emerging from processes in which some kinds of knowledge direct the application of other kinds in which retrieval is not primarily the result of linked associations but rather is computed by heuristic and logical processes that embed specific knowledge about what kinds of information are usually appropriate to the particular goal that is current.

We already know some effective ways to structure logically deep but epistemologically narrow bodies of knowledge, as the result of research on special purpose heuristic programs like MACSYMA, DENDRAL, CHESS, or the Vision System discussed herein. To get experience with broader, if shallower, systems we plan to build up small models of real world situations; each should be a small but complete heuristic problem-solving system, organized so that its functions are openly represented in forms that can be understood not only by programmers but also by other programs. Then the simple-minded solutions proposed by these mini-theories may be used as plans for more sophisticated systems, and their programs can be used as starting points for learning programs that intend to improve them.

In the next section we will describe a micro-world whose subject matter has a close relation to the vision world already described. Its objects are geometric solids such as rectangular blocks, wedges, pyramids, and the like. They are moved and assembled into structures by ACTIONS, which are taken on the basis of deductions about such properties as shape, spatial relations, support, etc. These interact with a base of knowledge that is partly permanent and partly contingent on external commands and recent events.

= = = = = = = = =

The remainder of this report was a detailed discussion of Terry Winograd's dissertation,which was issued as MIT AI Technical Report 235, February 1971 with the title *Procedures as a Representation for Data in a Computer Program for Understanding Natural Language* It was published as a full issue of the journal *Cognitive Psychology* Vol. 3 No 1, 1972, and as a book, *Understanding Natural Language* (Academic Press, 1972). See also http://hci.stanford.edu/~winograd/shrdlu/

The thesis can be read at

ftp://publications.ai.mit.edu/ai-publications/0-499/AITR-235.ps or

ftp://publications.ai.mit.edu/ai-publications/pdf/AITR-235.pdf