those conversations very closely, and my imaginary objector is a combination of many real-life people. The discussion is not very structured and ranges over a variety of topics, but this seems unavoidable.

7.2 A CONVERSATION

Can we begin with the levels-of-explanation idea, since you attribute so much importance to it? How is it related to ideas about feature detectors and in particular to Horace Barlow's first dogma (1972, p. 380), which states, "A description of the activity of a single nerve cell which is transmitted to and influences other nerve cells, and of a nerve cell's response to such influences from other cells, is a complete enough description for functional understanding of the nervous system"?

Here, of course, I must disagree with Barlow's formulation, although I do agree with one of the thoughts behind this dogma, namely, that there is nothing else looking at what the cells are doing—they are the ultimate correlates of perception. However, the dogma fails to take level one analysis—the level of the computational theory—into account. You cannot understand stereopsis simply by thinking about neurons. You have to understand uniqueness, continuity, and the fundamental theorem of stereopsis. You cannot understand structure from motion without knowing a result like the structure-from-motion theorem, which shows how such a phenomenon is possible. In addition, and critically important for a researcher, the levels approach enforces a rigid intellectual discipline on one's endeavors. As long as you think in terms of mechanisms or neurons, you are liable to think too imprecisely, in similes.

Remember the moral from the early stereopsis networks discussed in section 3.3! None of them formulated the computational problem precisely at the top level, and almost all the proposed networks actually computed the wrong thing. Another example was the notion of segmentation to carve up an image into regions and objects. This wasted an enormous amount of time and led to the development of all kinds of special relaxation and hypothesize-and-test methods for agglomerating areas of the picture into useful regions (see Chapter 4). The problem again was that people became so entranced by the mechanisms for doing something that they erroneously thought they understood it well enough to build machinery for it—just as had occurred in the simpler case of stereopsis. It was only with a level-one attack—the formulation of the 2½-D sketch and its attendant and precisely stated problems—that real progress was possible.

Have I made my case strong enough yet? The levels idea is crucial, and perception cannot be understood without it—never by thinking just about synaptic vesicles or about neurons and axons, just as flight cannot be understood by studying only feathers. Aerodynamics provides the context in which to properly understand feathers. Another key point is that explanations of a given phenomenon must be sought at the appropriate level. It's no use, for example, trying to understand the fast Fourier transform in terms of transistors as it runs on an IBM 370. There's just no point—it's too difficult.

For instance, take the retina. I have argued that from a computational point of view, it signals $\nabla^2 G * I$ (the X channels) and its time derivative $\partial/\partial t\,(\nabla^2 G * I)$ (the Y channels). From a computational point of view, this is a precise specification of what the retina does. Of course, it does a lot more—it transduces the light, allows for a huge dynamic range, has a fovea with interesting characteristics, can be moved around, and so forth. What you accept as a reasonable description of what the retina does depends on your point of view. I personally accept $\nabla^2 G$ as an adequate description, though I take an unashamedly information-processing point of view. A retinal physiologist would not accept this, because he would want to know exactly *how* the retina computes this term. A receptor chemist, on the other hand, would scarcely admit that these sorts of consideration have anything at all to do with the retina! Each point of view corresponds to a different level of explanation, and all must eventually be satisfied.

Yes, I see the point. You're simply saying that, from an information-processing point of view, what is done and why assumes paramount importance—this is your top level. The implementation details don't matter so much from this perspective provided that they do the right thing.

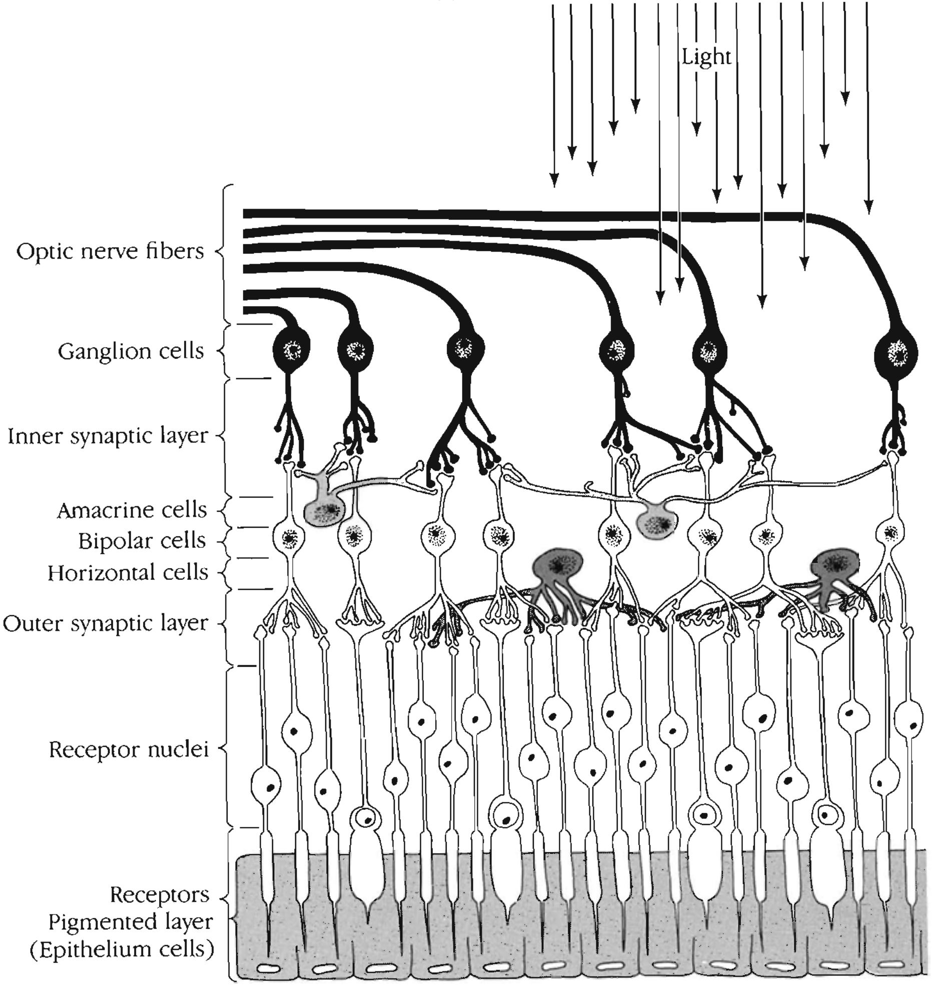
I'd like to make that point even more strongly. Figure 7–1 shows three descriptions of essentially the same thing. At the top is the mathematical description that we're so familiar with, $\nabla^2 G * I$. Figure 7–1(b) shows a piece of the retina, which we believe does roughly this, at least in part. And Figure 7–1(c) illustrates a silicon chip, built for us by Graham Nudd of the Hughes Research Laboratories in charge-coupled device technology, which carries out the $\nabla^2 G$ convolution. So, in a real sense, all these three things—the formula, the retina, and the chip—are similar at the most general level of description of their function.

Are the different levels of explanation really independent?

Not really, though the computational theory of a process is rather independent of the algorithm or implementation levels, since it is determined solely by the information-processing task to be solved. The algorithm depends heavily on the computational theory, of course, but it also depends on the characteristics of the hardware in which it is to be implemented. For instance, biological hardware might support parallel algorithms more readily than serial ones, whereas the reverse is probably true of today's digital electronic technology.

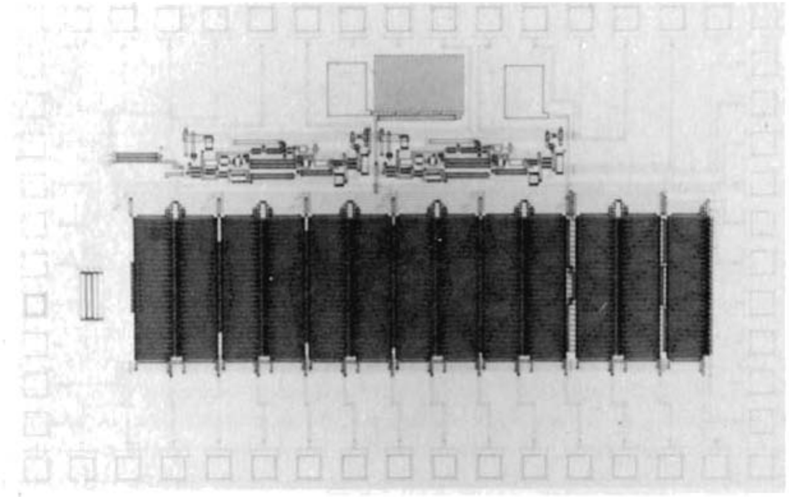
$$\nabla^2 G * I(x,y),$$
$$\text{where } \nabla^2 G(r) = -\frac{1}{\pi\sigma^4}\left(1-\frac{r^2}{2\sigma^2}\right)\exp\left(\frac{-r^2}{2\sigma^2}\right)$$

(a)



(b)

Figure 7–1. (a) The mathematical formula that describes the initial filtering of an image. ∇^2 is the Laplacian, G is a Gaussian, I (x,y) represents the image, and * the operation of convolution. (b) A cross section of the retina, part of whose function is to compute (a). (c) The circuit diagram of a silicon chip, built by Graham Nudd at Hughes Research Laboratories, which is capable of computing (a) at television rates.



(c)

Figure 7–1 (continued).

I cannot really accept that the computational theory is so independent of the other levels. To be precise, I can imagine that two quite distinct theories of a process might be possible. Theory 1 might be vastly superior to theory 2, which may be only a poor man's version in some way, but it could happen that neural nets have no easy way of implementing theory 1 but can do theory 2 very well. Effort would thus be misplaced in an elaborate development of theory 1.

Yes, this could certainly happen, and I think it already has in the case of deriving shape from shading. I would not be at all surprised if it was unreasonably difficult to solve Horn's integral equations for shape from shading with neural networks, yet the equations can be solved on a computer for simple cases. Human ability to infer shape from shading is very limited, and it may be based on simplistic assumptions that are often violated—a sort of theory 2 of the kind you mentioned. Nevertheless, I doubt that the effort put into a deep study like Horn's was misplaced, even in the circumstances. Although it will not yield direct information about human shape-from-shading strategies, it probably provides indispensable background information for discovering the particular poor man's version that we ourselves use.

What about the old feature detector ideas? How did they fit in?

Historically, I think, the notion of a feature—and I would not now care to define it at all precisely—played an important role in shifting our conceptions away from Lashley's mass-action ideas (according to which the brain was a kind of thinking porridge whose only critical factor was how much was working at the time) and toward the much more specific view of single-neuron action that we now have. This movement was initiated by Barlow (1953), Kuffler (1953), Lettvin and others (1959), and, of course, Hubel and Wiesel (1962, 1968). Essentially, these findings ultimately lead to the notion that single nerve cells can have as one of their functions the job of signaling explicitly whenever a particular, very specific configuration is present in the input, and this type of thinking was formulated in terms of features.

But there are a number of fascinating points here arising mainly from the basic question, When does a specific configuration in the image imply a specific configuration in the environment? The first point, which we met in Chapter 1, has to do with how descriptions of the environment actually get made. In a true sense, for example, the frog does not detect *flies*—it detects small, moving, black spots of about the right size. Similarly, the housefly does not really represent the visual world about it—it merely computes a couple of parameters $(\psi,\dot{\psi})$, which it inserts into a fast torque generator and which cause it to chase its mate with sufficiently frequent success. We, on the other hand, very definitely do compute explicit properties of the real visible surfaces out there, and one interesting aspect of the evolution of visual systems is the gradual movement toward the difficult task of representing progressively more objective aspects of the visual world. The payoff is more flexibility; the price, the complexity of the analysis and hence the time and size of brain required for it.

But wasn't there more to the features idea than that?

Yes, and that, too, is an interesting set of issues that harks back to some extent to the philosophers of perception, who thought in terms of "sense atoms" grouped into larger "molecules" of sensory experience, which were the things we could recognize. One can perhaps follow a tradition of attempts at feature-based recognition. This started with the Barlow (1953) ideas, involved Kruskal's (1964) multidimensional-scaling technique, Jardine and Sibson's (1971) excellent work on cluster analysis, my early ideas about the neocortex (Marr, 1970), and the mountainous literature on statistical decision theory.

What was the main idea?

The hope was that you looked at the image, detected features on it, and used the features you found to classify and hence recognize what you were looking at. The approach is based on an assumption which essentially says that useful classes of objects define convex or nearly convex regions in some multidimensional feature space where the dimensions correspond to the individual features measured.That is, the "same" objects—members of a common class—have more similar features than objects that are not the same.

That sounds perfectly reasonable. What went wrong?

It's just not true, unfortunately, because the visual world is so complex. Does feature refer to the image or the object? Different lighting conditions produce radically different images, as do different vantage points. Even in the very restricted world of isolated, two-dimensional, hand-printed characters, it is difficult to decide what a feature should be. Think of a *5* gradually changing into a *6*—a corner disappears, a gap narrows. Almost no single feature is necessary for any numeral. The visual descriptions necessary to solve this problem have to be more complex and less directly related to what we naturally think of as their representation as a string of motor strokes.

So your main argument is that the world is just too complex to yield to the types of analysis suggested by the feature detector idea?

That is correct unless, of course, the visual environment can be rigidly constrained—the lighting, the vantage point, the domain of visible elements, and so forth. If this is done, then some progress can be made. Otherwise not, and we have to look quite carefully in the literature to see this, because people do not report negative results, even though such results can be very important in deciding whether to pursue a particular line of attack.

What are the options if the domain of study cannot be so rigidly constrained?

There are basically two: Use a more complicated decision criterion or use a better representation. Using a more complicated decision criterion means abandoning the hope that classes correspond to convex clusters of features and introducing logical ideas in the decision process so that the questions asked at a given point in the classification process may depend on the answers just obtained. It is roughly true to say that artificial intelligence grew out of this approach. It leads to a view of recognition or classification as an exercise in problem solving. Decisions and routes to the solution depend sensitively on partial results found along the way, and these in turn determine the information deployed next to allow the process to continue. We saw some examples of this type of thinking in Chapter 5. The other option is to use a representation or series of representations that are better tailored to the problem at hand. In practice, this turns out to be the more important task for the particular case of vision, although for problems like medical diagnosis the problem-solving approach may be more profitable.

Are there perhaps other ways in which we might try to think about these things? What about Winograd's (1972) procedural representation of knowledge, for example, according to which terms like pick-up *or* block *are represented by programs. If you want to pick up the block, you simply run the two programs in sequence. That sounds like a very sensible approach to me. How does that relate to your two options?*

The procedural representation idea isn't really a representation at all; it is an implementation mechanism. A representation is a much more precisely defined object. For example, there was never any result defining the scope of the procedural representation or establishing any uniqueness characteristics (in the sense of Chapter 5). It is no more a representation than is a property list! In order to define a representation, as we have seen, we must define its primitives, how they may be organized, and so on. Now the primitives in these procedural representations are simply the primitives of the underlying programming language—in Winograd's case, PLANNER or LISP. Such primitives are useless for representing what the process is actually doing in any high-level description, just as the individual instructions in a machine language program for the fast Fourier transform are useless for understanding the transform. To begin to understand and manipulate the code, one has to add comments to it. At this point it is these, not the code, that in effect provide the representation of what the code is doing from the point of view of the manipulator. G. J. Sussman's (1975) program HACKER was essentially an exercise in writing useful standard comments within a particular and restricted programming domain.

Why do you say a property list is not a way of representing knowledge? Surely it is?

I did not say that, I said it wasn't a *representation*. A property list is a programming mechanism that one may use to *implement* a representation, but it is not a representation in itself. To see this, just ask the simple question, What can and what cannot be represented in a property list, or, expressed in our earlier language, what is its scope? Is each description unique? It is meaningless to ask these questions about property lists, just as it is about procedures. Both these ideas are universal from a representational point of view, because both are in fact notions at a lower level of explanation pertaining to decisions about implementation. They are *mechanisms*, not representations. Choosing one mechanism rather than the other will affect how easy it is for the programmer to make a certain piece of information explicit, but the decision about what is to be made explicit and what is not is a decision about the representation itself and is independent of the implementing mechanism.

Ah yes, and here we come back to the feature idea again. For it was surely the notion of a feature which led eventually to the idea that a representation has as its business the making of certain information explicit, wasn't it?

Very much so. But I do think that the time has now come to abandon those older ways of thinking, it being more fruitful to think instead of systems of representations that can describe as fully as desired firstly images and then other derived aspects of the visual world. And I also think it is important not to be too anxious to relate our ideas immediately to neurons. We should first be sure that our representations and algorithms are sensible, robust, and supported by psychophysical evidence. Then we can delve into the neurophysiology.

Before leaving this topic, I feel there is one other matter we should raise. This is the question of features—well, let's call them descriptions from now on—and of measurements for getting them. What exactly is the difference between a descriptive element—perhaps we could call it an assertion—and a measurement? Is this even an important point?

There are two aspects to this. One is historical—a point I felt lay in terrible confusion back in 1974—and the second is a modern question. Let us look first at the historical question. Put most simply, people confused measurements and assertions. For example, a cell with a center–surround receptive field will respond to a blob, but it will also respond to many other things—a line, an edge, two blobs, and so forth. In fact, one can often say no more than that it signals a convolution—our old friend $\nabla^2 G * I$, for instance. Nevertheless, people did call these cells blob detectors.

Now that is not so bad in the retina, but if we were to take Hubel and Wiesel's (1962) definition of a simple cell—the simplest type of receptive field—literally, it, too, would be performing a linear convolution with one excitatory and one inhibitory strip, signaling something like a first directional derivative. I do not now believe these cells are linear convolvers (see Chapter 2), but the point is that people thought of them simultaneously as linear convolvers *and* as feature detectors, and that is criminal, intellectually. Of course, you can use the output of such convolvers to find edges, but it needs extra work. You have to find peaks in the first derivatives or zero-crossings in the second. And, of course, we now think that simple cells are in fact zero-crossing detectors. But the point is that here again, just because of imprecise thinking by computer vision people as well as by physiologists, that whole rich theory of early vision had been missed (see Chapter 2).

The second aspect is the modern one, and I have already raised it in Chapter 2. It has to do with when and how vision "goes symbolic." Most would agree that an intensity array $I(x,y)$ or even its convolution $\nabla^2 G * I$ is not a very symbolic object. It is a continuous two-dimensional array with few points of manifest interest. Yet by the time we talk about people or cars or fields or trees, we are clearly being very symbolic, and I think again that most would find suggestions of symbols in Hubel and Wiesel's (1962) recordings. Our view is that vision goes symbolic almost immediately, right at the level of zero-crossings, and the beauty of this is that the transition from the analogue arraylike representation to the discrete, oriented, sloped zero-crossing segments is probably accomplished without loss of information (Marr, Poggio, and Ullman, 1979; Nishihara, 1981).

And the use of symbols does not stop there either. Almost the whole of early vision appears to be highly symbolic in character. Terminations, discontinuities, place tokens, virtual lines, groups, boundaries—all these things are very abstract constructions, and few of their neurophysiological correlates have been found, but experiments like Stevens' (1978) tell us that such things must be there (see Chapter 2).

How else might one approach these phenomena? What about some kind of transformational or grammatical approach, like the one Chomsky used?

People have tried to write picture grammars involving rules that must be obeyed by line drawings (Narasimhan, 1970), but they have been unsuccessful in general and never successful on a real image. The best of the early approaches, was, I think, the blocks-world analysis of Guzman (1968), Mackworth (1973), and Waltz (1975). Unfortunately, this did not generalize—it suffered from the wrong choice of a miniworld, as indeed has much research in artificial intelligence. The great virtue of artificial intelligence has been that it forced people to substantiate their opinions by writing programs, and in doing so, these opinions were often found to be wrong. It forced a constructive way of thinking—disallowing, for example, Bertrand Russell's definition of the percept of an object as the set of all possible appearances of the object (Russell, 1921). But in having to program things, research was too often limited to a miniworld in which very many factors appear in only simple forms. Though the programs solved none of the individual problems, on the whole they ran just well enough to get by with luck. Winograd's (1972) blocks-world program was of this genre. The underlying conceptual fault is to ignore the modularity that must be present to help decompose the problem.

I do not follow. Why must it be there? How was it being ignored?

Once again, I think the clearest examples come from vision. An early miniworld, or domain of study if you like, was the blocks world—compositions of matte white prisms against a black background. The study of such a domain led to Waltz's (1975) careful cataloguing of the legal junctions of the various types of edges (as in Figure 1–3). Allowing for shadows, Waltz found that most line drawings of such scenes could be interpreted unambiguously. But notice that not one of the general processes listed in Chapter 3 was elucidated by this approach. The reason is that the general processes that combine to make up human vision cannot be easily studied by restricting oneself to any particular miniworld except by carefully choosing it in relation to something that one already suspects of corresponding to a genuine module, like the world of random-dot stereograms.

It is critical to appreciate the difference between these two kinds of miniworlds. One is very particular, the other general. Only the second kind has been found to be of value so far, although constraints in the spirit of

Waltz's may turn out to be useful for the 2½-D sketch (see Chapter 4). The reason is that for genuine computational modules with general and not limited abilities, we can actually prove theorems that show the modules will always work in the real world.

This is the true difference between the approach described in this book and the original conception of artificial intelligence, which, in its desperate effort to pack a whole working miniworld into a program—an endeavor that requires a huge amount of work—was forced to neglect and eventually to abandon attempts at real theory, turning instead to the development of better computer tools. This endeavor has met with little success. So although the artificial intelligence approach was necessary to haul us out of our false preconceptions about the simplicity of vision, it in turn became limited and hidebound because of its failure to recognize what a true computational theory is and how it should be deployed.

Are there any rules for doing this successfully?

I don't think so, and it's perfectly natural to get it wrong first. The example of flight that came up earlier makes a number of points in a nice way. First, it's obvious that you cannot understand how a bird flies by speculating on the fine structure of a feather. So the next natural step is to try to copy how the bird behaves—what I call the mimicry phase. So people built imitation wings and flapped them. That didn't work either. This phase is essentially copying at the lower two levels or possibly only at level two. The real advance comes only when you understand that an airfoil provides lift in accordance with Bernoulli's equation. That is the level-one part—aerodynamics. It is why a bird and a 747 are similar—and why both are dissimilar from a gnat, which keeps itself aloft not by means of an airfoil but by "treading air" in an essentially turbulent regime.

But at some stage, one has to relate one's level-one ideas directly to neural machinery, surely? You talked about the eyes—the retina and $\nabla^2 G$—but what about eye movements? I understand that from your—I should say, from an information-processing and levels point of view, they are quite trivial to deal with. But that doesn't make it any easier for me to think of compensating for them in neural machinery.

Yes, I admit that this is a thorny issue. But first, I hope I made it clear in Chapter 4 that eye movements involve much more than just a subtraction. We saw there how the representation of surface orientation, for example, is quite intimately bound up with whether you choose a retinocentric polar frame (the natural one from the point of view of imaging) or a more invariant type of retinocentric frame.

The second point is that, by delaying the transition out of a retinocentric frame, the difficulty of the arithmetic that is necessary when one at last performs the transition is correspondingly eased. In the manner of Chapter

5, we can move directly to a 3-D model representation, which is located in a stable frame around the viewer; and then all we have to check is that when the eyes move, the appropriate blob moves as expected.

Lastly, I think that here, as always, it is important not to be fooled by the apparent detail and luxury of our perception. We met this earlier in connection with the immediacy and vividness of our perception. I would be surprised if we can keep track of more than a handful of objects during eye movements, and I expect our powers are quite limited in this respect.

Yes, I see the plausibility of the argument. But this doesn't need our levels, does it? It seems a rather different kind of issue.

Absolutely true, but that is mostly because the level-one theory of eye movements is so simple that we don't notice that it's even there. In fact, general ideas along these lines were in Gibson's thinking, I suspect, and were certainly being articulated by Marvin Minsky and Seymour Papert in the late 1960s and early 1970s. But the details to these general ideas were never filled in. In a cu:ious sense, this was because artificial intelligence remained decerebrate. It never realized that there was a level one theory to be discovered. It remained, and often still does, stuck fast in the mud of mechanistic explanations—where memory is held to be achieved by a neural net of some kind, or by a process in a computer, or by a set of procedures.

I don't know about this. These seem quite reasonable ways of explaining memory. Why do you find them so objectionable?

Well, in simple cases like eye movements, we can think in that rather direct fashion and get away with it. But it is very dangerous to hope that this type of thinking can ever give any real insight into the computational problems that the neural mechanisms are busy solving.

For example, to take a famous and elegantly expressed case, we might discuss Minsky's frames theory a little. A frame is essentially an item to which properties may be attached. For example, consider the following properties of an elephant considered as a frame:

Name	Clyde
Color	Pink
Weight	Large
Appetite	Large

Processes can also be attached to a frame and the contents of a frame may be interconnected or indexed in various ways. In his most stimulating article, Minsky (1975) describes how many "subjectively plausible" phenomena can be thought of in this way provided that the conceptual units involved are "large" enough. But I believe the approach is fundamentally flawed by its mechanism-based thinking. This harks back to our earlier point. If frames offered a representation and not just a mechanism, we would at once see what they are capable of representing and what they are not. This may still be done, but it has not yet been; until it has, we must be wary of ideas like frames or property lists. The reason is that it's really thinking in similes rather than about the actual thing—just as thinking in terms of different parts of the Fourier spectrum is a simile in vision for thinking about descriptions of an image at different scales. It is too imprecise to be useful. Real progress can only be made in such cases by precisely formulating the information-processing problems involved in the sense of our level one.

But your point isn't about just frames, is it? Doesn't it apply to almost the whole of artificial intelligence?

Yes, very true, and mechanism-based approaches are genuinely dangerous. The problem is that the goal of such studies is mimicry rather than true understanding, and these studies can easily degenerate into the writing of programs that do no more than mimic in an unenlightening way some small aspect of human performance. Weizenbaum (1976) now judges his program ELIZA to belong to this category, and I have never seen any reason to disagree. More controversially, I would also criticize on the same grounds Newell and Simon's (1972) work on production systems and some of Norman and Rumelhart's (1974) work on long-term memory.

Why, exactly?

The reason is this. If we believe that the aim of information-processing studies is to formulate and understand particular information-processing problems, then the structure of those problems is central, not the mechanisms through which their solutions are implemented. Therefore, in exploiting this fact, the first thing to do is to find problems that we can solve well, find out how to solve them, and examine our performance in the light of that understanding. The most fruitful source of such problems is operations that we perform well, fluently, and hence unconsciously, since it is difficult to see how reliability could be achieved if there was no sound, underlying method.

Unfortunately, problem-solving research has for obvious reasons tended to concentrate on problems that we understand well intellectually but perform poorly on, like mental arithmetic and cryptarithmetic* geometry-theorem proving, or the game of chess—all problems in which human skills are of doubtful quality and in which good performance seems to rest on a huge base of knowledge and expertise.

*For example, DONALD + GERALD = ROBERT. The object is to find the digit each letter stands for.

I argue that these are exceptionally good grounds for *not* yet studying how we carry out such tasks. I have no doubt that when we do mental arithmetic we are doing something well, but it is not arithmetic, and we seem far from understanding even one component of what that something is. I therefore feel we should concentrate on the simpler problems first, for there we have some hope of genuine advancement.

If one ignores this stricture, one is left with unlikely looking mechanisms whose only recommendation is that they cannot do something we cannot do. Production systems seem to me to fit this description quite well. Even taken on their own terms as mechanisms, they leave a lot to be desired. As programming languages, they are poorly designed and hard to use, and I cannot believe that the human brain could possibly be burdened with such poor implementation decisions at so basic a level.

This mimicry idea—is it just the business of thinking in similes that you mentioned before?

Yes, very much so. In fact, we could draw another parallel, this time between production systems for students of problem solving and Fourier analysis for visual neurophysiologists. Simple operations on a spatial-frequency representation of an image can mimic several interesting phenomena that seem to be accomplished by our visual systems. These include the detection of repetition, certain visual illusions, the notion of separate independent channels, separation of overall shape from fine local detail, and a simple expression of size invariance. The reason why the spatial-frequency domain is ignored by image analysts is that it is virtually useless for the main job of vision—building up a description of what is there from the intensity array. The intuition that visual physiologists lack, and which is so important, is for how this may be done. As a computing mechanism, a production system exhibits several interesting ideas—the absence of explicit subroutine calls, a blackboard-like communication channel, and some notion of a short-term memory.

However, just because production systems display these side effects (as a Fourier analysis "displays" some visual illusions) does not mean that they have anything to do with what is really going on. For example, I would guess that the fact that short-term memory can act as a storage register is probably the least important of its functions. I expect that there are several "intellectual reflexes" that operate on items held there about which nothing is yet known and which will eventually be held to be the crucial things about short-term memory.

Studying our performance in close relation to production systems seems to me a waste of time, because it amounts to studying a mechanism, not a problem. Once again, the mechanisms that such research is trying to penetrate will be unraveled by studying the problems that need solving, just as vision research is progressing because it is the problem of vision that is being attacked, not neural visual mechanisms.

What about human memory? You implied that the same type of misdirection was evident there. What did you mean?

I was referring to Norman and Rumelhart's work on the way information seems to be organized in long-term memory. Again the danger is that questions are not asked in relation to a clear information-processing problem. Instead, they are asked and answers proposed in terms of mechanisms—in this case the mechanism is called an "active structural network," and it is so simple and general as to be devoid of theoretical substance. Norman and Rumelhart may be able to say that such an "association" seems to exist, but they cannot say of what the association consists, nor do they say that to solve problem x (which we humans can solve) memory must be organized in a particular way; and that if this organization exists, certain apparent "associations" occur as side effects.

The phenomenological side of experimental psychology can do a valuable job in discovering facts that need explaining, including those about long-term memory, and the work of Shepard (1975), Rosch (1978), and Warrington (1975), for example, seems to me very successful at this; but like experimental neurophysiology, experimental psychology will not be able to explain those facts unless information-processing research has identified and solved the underlying information-processing problems, and I think that this is where we should be concentrating our energies.

What about Gunther Stent's work on the leech, though? Isn't that rather mechanism based, too?

Yes, but it is meant to be. It is concerned with elucidating the precise mechanism by which a leech swims. I value his work very highly, like that of the Tübingen group's on the housefly, but I think that early hopes of generalizing very far from these results have not borne fruit, and the reason is the levels story again. What higher nervous systems must do is determined by the information-processing problems that they must solve. We may have some simple leechlike oscillators inside us, and they may, to be very farfetched, eventually help us to understand some aspects of respiration. But such results will not teach us how we see.

One has a strong urge to tie explanation to structure eventually—that, of course, was the impact of molecular biology. It has to be done here, don't you think? Or do you see the endeavor as totally hopeless?

Yes, I agree it has to be done for the central nervous system, but I doubt if it can ever be done completely. The complexity barrier is just too great. But we have started to do it, don't forget! The zero-crossing detection and directional selectivity stories are very close to neurons. Don't be too impatient about the later things! As I said earlier, I bet you could never understand the fast Fourier transform as implemented in transistors on an IBM 370. I can only understand its formulas for about 10 minutes at a time—let alone understand a circuit diagram implementing them. One last word—I don't think that developmental and genetic programs will be able to be understood so directly in terms of underlying mechanisms. I would guess that some levels structure will eventually be needed to understand growth, because it is complicated.

Can we perhaps return to thinking rather specifically about visual perception and what actually happens when you see?

Well, are you happy with the primal sketch ideas?

I think so. The critical point seems to be that even very early vision is a highly symbolic activity. Assertions are actually made where lines end—yes, I've even accepted that terminology and am not too worried here about neurons!—and that objective lines and virtual lines are just as "real" as one another. Both can, for example, have their orientations detected and manipulated. Isn't this the idea?

Very much so. And if there is one more key idea, it is the idea of a place token and the ability to use crude selection criteria to group such tokens together and look for patterns, just as we saw in Figure 2–3.

I'm still a little unhappy about the representation of spatial relations—in the image, that is. I remember the discussion in Chapter 2 about coordinate systems, but was a little unconvinced. How can we be sure that important spatial information isn't lost?

Well, we have to be careful here, because I do not think much in the way of spatial relations *is* made explicit very early on. For example, certainly no intrinsic structure like the angle between two lines is. This type of information is not explicit in the full primal sketch, nor would the angle between two surfaces be in the 2½-D sketch. Such quantities do not belong to perception; their realm is that of the 3-D model representation. On the other hand, a few explicit spatial relations, like virtual lines between neighboring place tokens, often carry implicitly the entire geometry of the figure. This can be true even if the length measurements are very imprecise—perhaps only ranked by size.

A striking example of the richness of the information coming from a few clues about nearness is provided by the archaeological endeavors of Flinders Petrie. He measured the similarity of graves found along the Upper Nile by judging the number of characteristics shared by pieces of pottery found in each one. By using just this similarity information, techniques like multidimensional scaling can recover the times of burial quite accurately. The story makes fascinating reading (see Kendall, 1969), but we need note only that in two dimensions, the situation is even more constrained. I do not think there's much danger of the information being lost, but I do think only rather little spatial information is made explicit at the early stages.

So we derive the full primal sketch and then all those processes of Chapter 3 run to give us surface information? And roughly speaking, that is delivered in retinocentric polar coordinates, with perhaps slight differences for each process?

Yes, indeed, and the surface information from each process is combined in the 2½-D sketch, still in a retinocentric fashion but perhaps in a more convenient frame than the polar one. In a deep sense this is the end of pure autonomous perception. At this point the information is ready to be turned into a real 3-D model type of representation, a description that you can then remember.

I'm still unhappy about this tying-together process and the idea that from all that wealth of detail all you have left is a description. It sounds too cerebral somehow.

Well, the description can be arbitrarily rich—it's just a question of how much time and energy you spend on it. The other matter, that visual perception is just the formation of such descriptions—well, that is the conceptual leap I'm asking you to make. I personally find nothing important that this view fails to account for in general, and since we probably understand 20%–25% of the whole process already, I'm frankly ready to put my money on the rest of the process being of the same character. It's a conceptual leap, to be sure, but I think this view is worth trying to live with for a while, because thinking of visual perception in terms of the formation of particular kinds of descriptions explains so much so simply. But don't try to think about vision all the time in neurons! It's just impossible—the structure of vision is complicated enough at the top level, and outrageously so in terms of wiring.

And the result of those Chapter 3 processes, embodied in the 2½-D sketch, is the end of the immediate perception?

I think it's the right place to make the division, because up to here the processes can be influenced little or not at all by higher-order considerations. They deliver what they compute—no more, no less. The term *immediate perception* is a bit misleading, because these processes can take time—think of fusing a random-dot stereogram—but they do not involve scrutiny in Julesz's sense of an active intelligent examination of the image and comparison of its parts. This is compatible with the random-dot stereogram case, because we think that when the time to perceive one is long, most of the delay is due to random-walk-like movements of the eyes as they try to find somewhere to start fusion from.

If the 2½-D sketch changes every time you move your eyes, you lose it every time you move them (except possibly for small movements purely in depth). Isn't this a terribly wasteful thing to do?

It is wasteful, surely, but if you have the machinery there capable of recomputing the scene in real time, it doesn't matter that it's wasteful. In fact, it almost has to be this way, since the point of the 2½-D sketch is to assemble and represent incoming perceptual information, not to store it, and the alternative of economizing on computing power by using more memory is of no real use here. Just suppose, for example, that a 2½-D sketch had foveal resolution everywhere and was driven by a foveal retina in the usual way. Immediately, the memory has to contain out-of-date information (or nothing) in most of its capacity. This is not what the memory is for. Before resorting to almost any real storage, one must convert to something like the 3-D model representation, which is much more stable than the viewer-centered appearance of an object in a fleeting world. So the representation in which information from the different sources is assembled must be retinocentric and transient, it should have a foveal region where resolution is high, and it should reflect exactly and only what is coming in now.

These seem sensible distinctions, but they raise a difficulty I have in relating this to my own experience. The problem is that there seem to be so many different things going on in this model for perception, yet my perception has a unity, a oneness that I feel does not jibe with or at least is not reflected in these ideas. How is all the information tied together? How can one account for the unity of visual experience?

The basic idea is indeed that very many things are delivered through almost independent processes. At the 2½-D sketch level they are tied together, but only implicitly, whereas the next step is the creation of object-centered descriptions of the visible shapes (which is perhaps localized in a viewer-centered frame), and the description here *is* a unified object made up just by adding properties to its basic shape description, rather as a novelist adds to a description by adding qualifying adjectives.

What do you mean by being tied together "only implicitly"?

Simply that although different processes operate in different ways, there is a way of finding out when they are referring to the same visual object.

You mean if a raw primal sketch process finds an edge, and a color process finds its color, the relation between the two is implicitly available? I don't quite follow.

It's all a question of addressing. In most computers, you address information by specifying where to look for it. In some computers, you access a chunk of information by specifying pieces of the chunk. That is a content-addressable memory, and such memories are easy to build. What we might have here is a mixture of these two types of addressing—something like "the edge at roughly position (x,y) in the visual field with an orientation within, say, 30° of some given value". That would uniquely specify the edge in question both for the raw primal sketch representation and for the output for the color processes. In this way, we can tie the two things together, at least in principle.

What, dare I ask, about all those cortical areas? Isn't it natural to expect that they should each deal with a different process?

I would not be surprised.

Then what you are hinting at is, essentially, that up to this point each process runs, perhaps in a different cortical area (by now there are 10 at least, aren't there?), and that by presenting each with rough information, which could be rough position and orientation, you define precisely which visual object you are referring to.

Yes, that is the addressing problem.

And then, in addition, you get the precise information with which that particular area or process is concerned—the particular color or disparity, for example.

Exactly. And I think that the critical point about this is that the joining together of information is done symbolically.

What do you mean by that?

It's not like adding together the three impressions that a printer uses to make a printed page of color. We never see the colors of things smudged beyond their boundaries. The point is that the rough position and orientation information is used as an address. If you want the position of an item's exact boundary, you look at the raw primal sketch. If you want its color, you look at the color process.

I see. This idea means that assembling the information must be a very active process, doesn't it? Unless something specifically notices that stereo, zero-crossing x is a brown border, these two pieces of information will remain separate.

Yes, I think one has to ask for the color of x. And we must expect much of this to go on automatically as we move our eyes around. That is what the 2½-D sketch is partly for, after all—reducing information about surface geometry from many retinocentric processes to a single, more usable, viewer-centered form. At the same time, links to descriptions of other aspects of a surface are presumably made easily accessible, in preparation for the task of constructing a three-dimensional, object-centered description.

So you think it's likely that the actual combination isn't done until the 3-D model starts being constructed?

Yes.

It's as though strings are there to all the relevant information clearly marked and labeled, but you don't pull it all together unless you start making a 3-D model.

Which may be a very coarse one or parts of a very fine one. And in the same way, one might expect other properties to be coarse (for example, greenish) or quite fine (for example, a specific shade of green).

But how does this correspond to my perceptual experience? My experience appears to be complete, not at all the halfway, ill-defined, fragmented sort of thing that you describe.

Well, first remember that our visual processes can work extremely rapidly. The time between requesting information about a part of the visual field and moving the eyes there, getting it, and linking it to a 3-D model is probably usually under half a second. The second thing is, How much of a novel scene can you recall if you look at it only very briefly? Not very much! Its coarse organization, or perhaps one or two details. And once you close your eyes, the richness is gone, isn't it? I think that the richness corresponds to what is available now, at the pure perceptual level, and what you can remember immediately is much more closely related to the 3-D model description that you create for it while your eyes are open.

I begin to see more clearly the force of the idea that perception is the construction of a description.

Yes, that is the core of the thing, and a really important point to come to terms with.

But let's suppose you're right, then, that the 2½-D sketch is retinocentric and that you compute out of it little 3-D models and hang them up in a space frame centered on you. What happens when you move your eyes a lot?

One thing is that the finely detailed shape that you were just looking at—suppose it was a porcelain cat—and for which you have just built up an elaborate description is reduced to a blob in the image when you turn your eyes to study its neighbor, a porcelain dog. If the blob can be distinguished confidently in the 2½-D sketch, then I would guess that there is a process that maintains the link between it and the 3-D model you've just finished building, so that if that blob moves, you know immediately *what* has moved.

But how on earth do you do that with neurons?

Hold on there—we'll face that next. But note that basically, it's not difficult computationally.

But to tie all this up with what it feels like to see—that is difficult to swallow.

It grows on you. That first step, that vision is the computation of a description, is the crucial one. Once you have accepted that, you can go on to study exactly what description and how to make it.

And again it's not at all easy for me to allow you to talk so much about computation. The brain, after all, is made of neurons, not silicon chips. But I suppose I'll get used to it. Still, if vision is the construction of descriptions, they must be implemented neurally, mustn't they? So couldn't one hope to look for neurophysiological correlates of the $2\frac{1}{2}$-D sketch or of a piece of a 3-D model? That, I would find convincing.

It would be marvelous if the implementation were that simple—close to Barlow's neural dogma! My own guess is that it is more like that than a Hebb cell assembly.

There's another more general point that is still troubling me, and it has to do with the temporal continuity of perceptual experience. I understand very well how you think continuity can be held between eye movements and so forth, but this avoids the larger question of pure continuity over time. Why, if I look at a tree, do I see it continuously as the same tree? Presumably I could at any moment start a new 3-D model for it, in which case I ought to experience it as a new tree in the same spot as the old one. Yet I don't. Do you have any comments?

The permanence of the visual world—the continuity of objects in time—is an awfully important aspect of vision, and I think it's just part of our reflexes as adults that we assume it. In fact, whole aspects of processing are based on discovering and exploiting the continuity relations—the correspondence processes of Chapter 3, for example.

Another general point. You deal only with shape here. What about the recognition as being the same thing of two objects that have different shapes but the same function—like two different kinds of chair?

This theory has nothing to say about semantic recognition, object naming or function, though that is most certainly a path almost as useful as shape determination for recognition in the external world (Warrington and Taylor, 1978). I think that the problems of understanding what we mean by the semantics of an object are fascinating, but I also think that they are very difficult indeed and at present much less accessible than the problems of visual perception.

If the overall scheme you describe is correct, would we be able to say anything about painting and drawing using this knowledge of what the visual system does with its input? Might it help to teach these skills, for example?

Perhaps, although I would hate to commit myself to a definite view yet. Nevertheless, it is interesting to think about which representations the different artists concentrate on and sometimes disrupt. The pointillists, for example, are tampering primarily with the image; the rest of the scheme is left intact, and the picture has a conventional appearance otherwise. Picasso, on the other hand, clearly disrupts most at the 3-D model level. The three-dimensionality of his figures is not realistic. An example of someone who operates primarily at the surface representation stage is a little harder—Cezanne perhaps?

With respect to other problems such as natural language, how universal is the approach you are advocating? How far can it be taken? What kind of things would it be likely to fail at?

Systems that are not modular. Things like the process by which a chain of amino acids folds to form a protein—that is to say complex, interactive systems with many influences that cannot be neglected. A burning issue in the study of natural language understanding is, of course, How modular is it, and what are the modules?

Yes, I suppose modularity is the key, but also fluency of some kind must be important, mustn't it? If a process doesn't flow well, smoothly, unattended, and without having to be patched by conscious interference, then it may have no clean theory, and that might turn it into the protein-folding class of difficult-to-understand theories. But to return to natural language, what modules have been found there?

It's not clear, and some claim it's inherently not modular and should be viewed much more heterarchically.

Doesn't that sound a little reminiscent of the early days of vision?

Yes, I'm afraid so. But there do seem to be modules and rules for modules emerging at the early level—rules for syllable formation, prosodics, and most famously Chomsky's analysis of syntax.

But how much of a module is syntax? Don't artificial intelligence workers like Schank claim that syntax is not a separable module at all?

Yes, and it is clear that the syntactical decoding of a sentence cannot proceed entirely independently of its semantical analysis. But a good case is being built up that the *amount* of interaction necessary between the two is small, and the types of questions about syntax that must be answered seem to be of a quite simple kind—for example, Should a particular clause refer to noun phrase one or to noun phrase two? Marcus (1980) was the first to explore these problems in detail; and he has shown that a very successful module can be made out of a parsing system. Above the level of syntax, however, few hints are currently available about what the modularity is, but I'm sure it must be present.

Why has artificial intelligence shown such resistance to traditional Chomskian approaches to syntactical analysis? Only Marcus seems to have embraced it.

I think there are two reasons. First, it is easy to construct examples in which syntax cannot be analyzed without some concurrent semantical analysis. Thus, syntax is not a truly isolated module, and this fact led the artificial intelligence people to jump to the opposite conclusion, that syntax is not a module at all. This is incorrect—the true situation seems to be that syntax is almost a module, requiring some interactions with semantics but only a very small number of types of interaction.

The second reason is our old friend, the levels. Noam Chomsky's transformational grammar is a level one theory, that is in no way concerned with *how* syntactical recognition should be implemented. It merely gives rules for stating *what* the decomposition of an arbitrary sentence should be. Chomsky's description of it as a competence theory was his way of saying this.

However, the levels idea has not been properly understood by computational linguists. Indeed, one of Winograd's reasons for rejecting Chomsky was that he could not invert the transformational structure and turn it into a parser! This observation could be made only by someone who failed to understand the distinction between levels one (what and why) and two (how). Winograd is not to be singled out for this error, however; everyone in artificial intelligence made it, and now that the linguists themselves are becoming computationally aware, they are falling into the same trap. The result is, I fear, that natural language computer programs have contributed rather little to natural language understanding, with the recent exception of Marcus (1980), who has begun to construct a genuine level-two theory of the parsing algorithm we use.

What do you feel are the most promising approaches to semantics?

Probably what I call the problem of multiple descriptions of objects and the resolution of the problems of reference that multiple descriptions introduce.

Could you expand on this?

Well, like many others in the field, I expect that at the heart of our understanding of intelligence will lie at least one and probably several important principles about organizing and representing knowledge that in some sense capture what is important about the general nature of our intellectual abilities. While still somewhat vague, the ideas that seem to be emerging are as follows:

1. The chunks of reasoning, language, memory, and perception ought to be larger than most recent theories in psychology have allowed (Minsky, 1975). They must also be very flexible, and incorporating this requirement precisely will not be easy.

2. The perception of an event or of an object must include the simultaneous computation of several different descriptions of it that capture diverse aspects of the use, purpose, or circumstances of the event or object.

3. The various descriptions referred to in point 2 include coarse versions as well as fine ones. These coarse descriptions are a vital link in choosing the appropriate overall scenarios demanded by point 1 and in correctly establishing the roles played by the objects and actions that caused those scenarios to be chosen.

An example will help to make these points clear. If one reads

> The fly buzzed irritatingly on the windowpane.
> John picked up the newspaper.

the immediate inference is that John's intentions toward the fly are fundamentally malicious. If he had picked up the telephone, the inference would be less secure. It is generally agreed that an "insect-damaging" scenario is somehow deployed during the reading of these sentences, being suggested in its coarsest form by the fly buzzing irritatingly. Such a scenario will contain a reference to something that can squash an insect on a brittle surface—a description that a newspaper fits, but not a telephone. We might therefore conclude that when the newspaper is mentioned (or, in the case of vision, seen) not only is it described internally as a newspaper and some rough 3-D model description of its shape and axes set up, but it is also described as a light, flexible object with area. Because the second sentence might have continued "and sat down to read," the newspaper must also be described as reading matter; similarly, it must also be described as a combustible article, as something that rustles, and so forth. Since we do not usually know in advance what aspect of an object or action is important, it follows that most of the time a given object will give rise to several different coarse internal descriptions. Similarly for actions. It may be important to note that the description of fly swatting or reading or fire lighting does not have to be attached to the newspaper—a description of the newspaper is merely available that will match its role in each scenario.

Why do you think this must be so?

Because the importance of a primitive, coarse catalogue of events and objects lies in the role that such coarse descriptions play in the ultimate access and construction of perhaps exquisitely tailored specific scenarios, rather in the way that a general 3-D animal model can finish up as a very specific Cheshire cat after due interaction between the image and information stored in the catalogue of models. What existed as little more than a malicious intent toward the innocent fly after the first sentence becomes, with the additional information about the newspaper, a very specific case of fly squashing. Exactly how this is best done and exactly what descriptions should accompany different words or perceived objects is not yet known.

What about other types of processing that the brain does, such as the planning and execution of behavior? Might not these be simpler places to start looking for modules? After all, semantics is one of the most advanced areas of human ability, so it's not unreasonable to expect that it may be complex. I would try something simpler.

I think that may be excellent advice, and it reminds me of a fascinating experiment done some time ago by Stamm (1969). He was running what is called a delayed-response task (see Figure 7–2). In this, a scrap of food

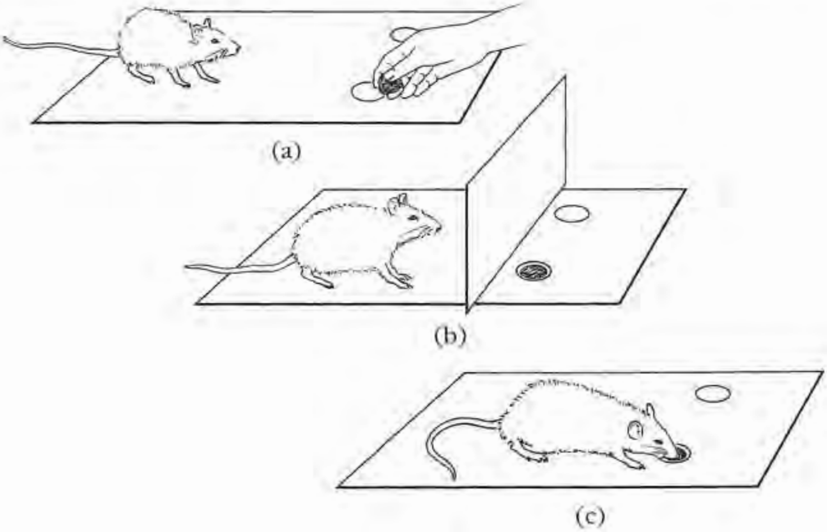


Figure 7–2. A delayed-response task. A scrap of food is placed under one of the wells in full view of the animal. Then a screen descends for a period. When the screen is raised, the animal has to choose one of the wells. If he looks under the correct one, he is rewarded with the food.

is placed in one of two wells, a screen comes down, a delay ensues, the screen lifts, and the animal is then free to choose the well in which he thinks the food is hidden. Certain portions of the prefrontal cortex are known to be involved in this task, and the animal cannot perform it if they are removed. Stamm used a technique—depolarization—whereby he could effectively disable these areas for the precise period he desired. He asked, When must the area be operating for the task to be carried out? It turned out that the animal had to have its area working as the screen came down at the beginning of the delay; if the area was knocked out at any other time, it mattered either much less or not at all!

One possible way of thinking about this experiment is this. Any real-time computer must be able to construct plans, set them up for execution under the appropriate conditions, and set the triggers for them. One cannot recompute everything afresh each time, and indeed the structure of a human personality consists in part of thousands of such little plans, all set to run a person's behavior if the appropriate conditions arise. But something must *write* these plans, and here in Stamm's experiment maybe we are seeing a simple example of this happening. As the wells are removed from view, the animal writes into its set of plans to go to the appropriate well when it can. A simple plan, but a plan nevertheless.

If we carry this idea a little further, we see that it splits the central system into what one might call the planner and the executive. The planner writes plans and their triggers to the executive, which, when the time and conditions are ripe, executes them. Is it too absurd to suggest that during hypnosis the executive becomes externally programmable and that this is why it is possible to set up plans under hypnosis that are executed later when the assigned conditions are met? The idea bears reflection, at least.

That is an interesting idea. I have not seen any previous explanation about why it should be possible to "program" someone at all, and your suggestion is certainly plausible. But what about the stereotyped nature of the programming? We are ourselves very flexible, are we not? It's a little difficult to reconcile that with a set of programmed responses.

I think that depends entirely on how large, rich, and subtle the set of responses has grown to be. If there is wide variety of responses and considerable ability to act differently in only subtly different situations, then we would be called flexible—and freer, incidentally, since we would be taking a wider range of relevant information appropriately into account. If we take no information (random response) or only one piece (compulsive response), then we are certainly not acting flexibly or freely.

That seems a sensible distinction. But as we move closer to saying the brain is a computer, I must say I do get more and more fearful about the meaning of human values.

Well, to say the brain is a computer is correct but misleading. It's really a highly specialized information-processing device—or rather, a whole lot of them. Viewing our brains as information-processing devices is not demeaning and does not negate human values. If anything, it tends to support them and may in the end help us to understand what from an information-processing view human values actually are, why they have selective value, and how they are knitted into the capacity for social mores and organization with which our genes have endowed us.