# A FORMAL THEORY OF PERCEPTION

by

William Arthur Rottmayer

TECHNICAL REPORT NO. 161

November 13, 1970

PSYCHOLOGY SERIES



INSTITUTE FOR MATHEMATICAL STUDIES IN THE SOCIAL SCIENCES

STANFORD UNIVERSITY

STANFORD, CALIFORNIA

ACKNOWLEDGMENTS

I wish to express my special gratitude to Professor Patrick Suppes, who formulated the problems I worked on, and whose advice, guidance and encouragement at all stages of this work was invaluable.

I also want to thank the other members of the reading committee, Professor Moravcsik and Professor Hintikka, whose criticisms and suggestions were extremely useful, George Huff, who made large contributions to this work, particularly in the early stages of the research, and my wife, Nola, whose encouragement and typing skill made this work easier and more enjoyable.

Finally, I wish to acknowledge the partial financial support I received from National Science Foundation grant NSFGJ-443X through the Institute for Mathematical Studies in the Social Sciences and fellowship support from Stanford University.

TABLE OF CONTENTS

# A FORMAL THEORY OF PERCEPTION

William Arthur Rottmayer

Stanford University
Stanford, California 94305

## CHAPTER 1

## BACKGROUND AND MOTIVATION

Work on this problem began as a joint effort of three people, Patrick Suppes, George Huff, and myself. The characterization of the problem is due primarily to Professor Suppes. The particular method we chose to attack it grew out of discussions among the three of us, and it is difficult to separate the contributions of each. This method consisted of constructing a particular model. Once the model was agreed upon, it was possible to independently obtain results about it, and most of the results contained herein are due to my own efforts.

The approach we took to the problems of perception concerns itself much more with scientific work than has been customary in the approach taken to these problems by recent philosophers. For this reason, it is useful to discuss the conditions that led us to this approach before turning to the details of the work we did. These considerations were not made explicit before we began, but were definitely there in the backs of our minds. This explicit account is my own creation, but it was obtained by reflecting on the common work we did. Thus it is an accurate account of my own motivation and a more or less satisfactory account of Professor Suppes and Mr. Huff's motivation. This account breaks down into three parts. First, there is a rough characterization of the dominant themes in the recent philosophical approach to perception and then our approach is compared and contrasted with this approach. Secondly, a brief account of the scientific work that influenced us is given. This is a good method of showing the main features of our work, and is also useful since many philosophers are perhaps unfamiliar with much of this material. Finally, there is a detailed discussion of why we felt our approach is advantageous

in trying to solve certain of the problems of perception. This chapter is divided into three sections, corresponding to these three topics. In the following discussion, it is understood that the entire discussion concerns perception. The things I say are meant to apply only to the philosophical discussion of perception, and are not applicable in any way to other philosophical problems, unless a claim to the contrary is made. I do not maintain that what I say applies to problems other than perception simply because I have no way of supporting such a view. Indeed, I believe that many of the things I say concerning perception are not true if applied without restriction to other philosophical problems. In any case, there is no reason to bring up the more general view in this paper, since it concerns itself entirely with the problem of perception.

## Section 1

For convenience in discussing different approaches to the problem of perception, it is instructive to think of talk about perception as occurring in one of three languages: the language of physics and physiology (PP), the language of psychology and computer science (PC), and ordinary English (OE). PP contains talk of light waves stimulating the retina and electrical impulses being transmitted to the brain along the optic nerve. PC contains talk of the inputs and outputs of information processing systems, and how these systems can be altered by learning. Another way of characterizing PC is to say PC talks of perception in the same way Chomsky talks of language. Of the three languages, PC is the newest and least developed, and thus the most unfamiliar. Hence, the above characterization is not completely satisfactory. However, it does give a rough idea and what I have in mind will become clearer as the paper progresses. OE is well known to philosophers. This threefold division of perceptual talk is not the only one possible, and it is certainly true that none of the three languages has been precisely defined and that there are significant borderline cases. This division is useful in stating my view of the philosophical problem of perception and how it should be approached, however, and that is all that is necessary.

There is no peculiarly philosophical language in the above division. The reason is that philosophical problems do not arise in a special language; they arise in a language that is already being used in a non-philosophical way. Philosophers may invent special terms for talking about the non-philosophical language in order to facilitate their discussion. The basic problems, however, are problems that are statable, perhaps in an imprecise way, in the non-philosophical language. I believe this is true of philosophical problems in general, and that the problem of perception is not exceptional in this regard. The particular threefold division into OE, PC, and PP was chosen because of its special relevance to perception, however, and would probably be unsatisfactory for most other uses. Using this threefold division as the framework for the discussion, the question arises of how does philosophy fit into this framework. Some philosophical problems deal with the interrelationship of the three languages. Issues involving questions of reduction fall in this category. If one arranges the three languages in order of complexity, OE is the simplest, PC is next, and PP is the most complex. Thus, if one were interested in the problems of reduction, there are two things that could be done: reduce OE to PC, or reduce PC to PP. Reducing OE to PP would simply be a matter of combining these two steps. However, we are not interested in reductionism, so the interrelationship between the different languages is not an important factor in our work. The remaining philosophical problems must be statable in at least one of the remaining languages. Which language is the likely candidate? PP isn't, for two reasons. First, it is not possible to state the philosophical problems of perception in PP, since in this language talk of even ordinary aspects of perception becomes unmanageably complicated. Indeed, in the present state of affairs, it is not even clear how one would go about translating philosophical problems into PP. Secondly, the conceptual framework of PP is well worked out, and once it is possible to deal with a problem in PP, there are no longer philosophical mysteries surrounding it. This preliminary discussion has thus led to the position that the interesting philosophical problems are statable in either OE or PC, or both. The real problem is which of these three possibilities is correct. My own position is that philosophical problems arise in both OE and PC, but that

the most important problems arise in PC. I do not want to dispute the claim that some of the philosophical problems of perception arise in OE, but I do disagree with the view that the problems of perception of primary philosophic interest arise in OE. Thus, I think philosophers working on perception should work both in OE and in PC, with more emphasis on the latter than the former.

This position is different from the one prevalent in twentieth century British-American philosophy, which is that philosophical problems, including the philosophical problems of perception, arise in OE. The prevalence of the view that philosophical problems arise in OE is closely related with two other beliefs which are characteristic of English philosophy in this century: namely that there is a sharp distinction between philosophy and science and the rejection of the causal theory of perception. The reason for this connection is clear. As far as the philosophically interesting problems of perception are concerned, PC is the language of science. If philosophers work in PC, then there will be no clear separation between their work and scientific work. This is not to say that the two will be identical, for presumably the philosopher's approach and goals will differ from the scientist's. If the philosophers confine their attention to OE, then there will be a sharp boundary between their work and the scientist's work. This boundary will be at least as sharp as the boundary between OE and PC, which is fairly clear. Thus, the belief that philosophers should work in OE goes hand in hand with the belief that there is a sharp distinction between philosophy and science. Secondly, it is also fairly clear that the theory of perception which is implicitly contained in OE, if there is in fact such a theory, is not a causal theory. H. P. Grice is the only modern philosopher I know of who has attempted to give a causal account of perception in OE, and by his own admission, his theory 'is very far from the spirit of the original theory.'[1] The theory implicit in PC is a causal theory with the original spirit, i.e., it is a genuine causal theory. I will have more to say concerning the causal theory later.

---

[1]H. P. Grice, "The Causal Theory of Perception," Perceiving, Sensing, and Knowing, ed. Robert J. Swartz (New York: Doubleday, 1965), p. 472.

To sum things up, scientists work in PC or PP, and accept the causal theory. Philosophers have worked in OE, rejected the causal theory, and correctly recognized that if this is correct, there is a sharp distinction between philosophy and science. My own view is that there is no such sharp distinction, that philosophers should work in PC as well as OE, and that the causal theory is correct.

The work we have been doing on the problem of perception is in the language PC. This work is not an isolated attempt to deal with some of the problems of perception, but is part of a unified approach to the whole problem. Two features of this work can be illustrated by contrasting it with the classic materialist doctrine. Materialism, when restricted to perception and stated in the present framework, is the claim that statements in both OE and PC can be reduced to statements in PP. There are two differences between our approach and the materialist program. The first difference is that our approach is not, like materialism, a reduction. It is not an attempt to reduce OE to PC. Rather, it is an attempt to state and solve classical philosophical problems within PC. Perhaps OE could be reduced to PC, but this is irrelevant to what we are trying to accomplish. Secondly, materialists have always claimed that PP was adequate for all talk of perception <u>in principle</u>, but have not tried to carry out the necessary reduction in detail. In solving specific problems of perception, it is not helpful to know whether or not a particular reduction is possible <u>in principle</u>; the only thing that would be of use would be an actual reduction. Our approach deals with specific problems and is useful when one has to deal with these problems. The fact that materialism is of no use in dealing with specific problems is perhaps the main reason for the twentieth century philosophical concentration on OE and the consequent split with science. Speaking in the present framework, at the beginning of this century there were only two languages which philosophers such as G. E. Moore could work in: PP and OE. There was no way to work in PP, so OE was the only possibility. It has proven very difficult to deal with all the problems of perception in OE, but fortunately it is not necessary to make the choice that confronted Moore, for PC is now available. This language can be applied to specific problems. Two such

problems are the problem of synthetic a priori knowledge and the problem of sense data. Both of these problems are difficult to state, let alone solve, in OE. There is nothing in OE that corresponds to the predicates synthetic and a priori in any straight forward fashion for there is no need for such concepts in ordinary discourse. OE also contains no sense-data terms. Thus, it is very difficult to discuss these problems in OE for the language does not even provide an adequate conceptual framework in which to state these problems. It is my belief that PC does provide such a framework. Later, I will give a model, drawn from PC, of the perceptual process which serves as a satisfactory framework in which to discuss these problems. Given this model, it is easy to see to what part of the perceptual process the terms 'synthetic a priori' and 'sense-data' apply to, and hence to see precisely what the problems are. Moreover, this model indicates in a general way what a satisfactory solution would look like. The outlook is not completely optimistic, however, for to actually get an explicit solution to these questions would require a much more well-developed theory. This will require a lot of work, and what we have done is only the beginnings of a complete theory.

## Section 2

In a situation from either ordinary life or a psychological experiment, it is often convenient to divide human activity into perceptual input, the processing of this input together with information stored in memory, and the resulting output. Ignoring the output device, an organism capable of such activity can be thought of as consisting of three parts, the perceptual component, memory, and the processing device. This paper deals with the perceptual component, which we believe is the least understood part. A computer provides at least a rough first approximation to the processing device, and there are also roughly adequate models for memory. There is currently no such model for the perceptual component, not even a very rough first approximation model that will provide a framework for dealing with the problems of perception. Providing such a model is much too large a problem to deal with all at once, so we have restricted ourselves to a small part of the problem.

It is natural to divide the perceptual component into five parts, corresponding to the five senses. Of these parts, the visual part occupies a place of salient importance, and it has been widely discussed in both the philosophical and the scientific literature. Thus, we decided to concentrate completely on the visual part, and this decision has guided our subsequent thinking. The fact, which now appears evident to me, that our model applies equally well to the tactile part is simply a happy coincidence. It occurred to me only after we had settled upon the approach we have taken. This fact was made possible by our decision to concentrate on geometry, which is at least intuitively based on both our visual and tactile experiences. It really results from the particular starting point we chose, as I will explain shortly. Right now, I want to give some motivation for concentrating on geometry.

Figuratively speaking, our idea is that visual perception has many factors, and that geometry is what ties them all together. More accurately, it provides the framework to which all the other factors must be attached in order to come up with a satisfactory model for the whole visual part. This conception is the basis of much of the scientific work in the area. Moreover, philosophers have long attributed central importance to vision and to geometry. This is almost self-evident, but a few remarks concerning it are in order. Locke calls vision 'the most comprehensive of the senses,' and one of Berkeley's major works is an essay concerning it. More generally, the typical example used in philosophical discussion of perception is almost always an example from visual perception, as in the Moore case below. The importance of geometry isn't quite so evident until one realizes that philosophers used to talk of 'extension' and nowadays talk of 'space' and 'spatial relations' instead of geometry. This is primarily a terminological point, however, since extension was regarded as the subject matter of geometry just as spatial relationships are now. Thus, Descartes and Kant, whom I will discuss later, are good examples of philosophers who assign a crucial role to geometry. More generally, any philosopher who uses spatial properties to individuate sense data or physical objects shares this viewpoint to some extent. G. E. Moore is a typical example. In a general discussion of what happens when we perceive an object, he confines his

attention to the particular case of what happens when we see an envelope.[2] The importance of position, size, and shape are evident throughout the discussion, but their overriding importance comes out in Moore's definition of a material object: "I propose, then, to define a material object as something which (1) does occupy space; (2) is not a sense datum of any kind whatever; and (3) is not a mind, nor act of consciousness."[3] Moore admits that this is an incomplete definition, but it is interesting that (1) is the only positive element in a definition that is supposed to be at least partially satisfactory.

The best way to characterize our particular approach is to contrast it with two other scientific approaches to the same problem. The first of these is the artificial intelligence approach. This work is done primarily by computer scientists, and it is concentrated in two places, Massachusetts Institute of Technology and the Stanford Research Institute. The goal is to write a computer program that has roughly the visual capabilities of the human brain. The computer uses a television camera for an eye, and the problem is to get a camera-program combination that can do the same kinds of tasks that the eye-brain combination can. There are two features of this work I want to mention. Our approach is the same as the artificial intelligence one in regard to the first of these, but completely different in regard to the second. The major problem encountered is getting the computer to be able to divide the scene it is presented with into regions that go together in the way humans can. This is ncessary if the computer is going to be able to distinguish physical objects by just looking at them. Geometry plays a crucial role in this problem, and this is further justification for concentrating our efforts on it. Indeed, on this approach the primary reason for investigating our other visual abilities, such as the ability to recognize colors and textures, is that these abilities provide us clues as to how to divide the visual scene into different regions and about the spatial orientation and relationships of these regions. Thus, for instance, a

---

[2] G. E. Moore, Some Main Problems of Philosophy (New York: Collier Books, 1966), p. 43 ff.

[3] Ibid., p. 148.

sudden change in color, texture, or light intensity is not of interest by itself, but is interesting because it indicates a boundary between two regions. On this analysis, it is natural to divide the computer's task into two distinct and quite independent tasks: drawing in boundary lines and then analyzing the resulting line drawing into bunches of regions that go together, i.e., are faces of the same physical object. It is true that in solving a definite problem the computer will go back and forth between these two tasks; for example, it will draw in an edge of a cube that doesn't show up in the first drawing on the basis that an edge is needed to make the analysis of the whole scene satisfactory and that a finer check of the place in the scene where this line ought to appear reveals some indication a line should be there. This sort of interaction not only works, but it is intuively very appealing, since it seems people operate in the same way, i.e., if they aren't satisfied with the picture they get from a quick glance at a scene, they go back and inspect it in detail. However, this sort of interaction doesn't alter the fact that the two tasks are conceptually quite independent. This point suggests that it would be wise to study the two tasks separately, and solve the larger problem by combining the answers to the two smaller ones. We accept the above analysis, and the course we took was to concentrate on two-dimensional line drawings and thus on the second of the two tasks. I believe this discussion is worth emphasizing, for at first glance, it is not at all clear how the specialized model we deal with, which concerns itself entirely with straight line drawings, can be regarded as part of a general theory of visual perception. We do regard it as such, and as the above discussion makes clear, have definite ideas on the place it would occupy in a complete theory. It is interesting to note that Helmholtz came to much the same viewpoint as the result of extensive optical experiments nearly a hundred years ago. He noticed that people are very attentive to visual characteristics that indicate how what they see is divided into physical objects or give clues concerning the size, shape, and distance of these objects. Indeed, adults process these clues so automatically that they can describe much more accurately the objective sizes and shapes of objects than they can the subjective visual phenomena. This habit is so engrained that it takes years of practice to even be aware, to even see,

in the ordinary sense of the word, the subjective visual phenomena. Most people are as unconscious of these phenomena as they are of the blind spot, and it is one of the main purposes of artistic education to bring these phenomena to consciousness. I will say more of Helmholtz's views in Section 3.

The second distinguishing feature of the artificial intelligence approach is that it is interested solely in building a machine that can do the tasks in question, not in building one that can learn to do them. We are interested in the latter task. It is clear that humans have to learn many of the facts they use in analyzing a visual scene, and thus only a learning device of some sort can be a completely satisfactory model. This is not an easy thing to do, however, and we have felt compelled to deal with far simpler problems than the artificial intelligence people are currently dealing with. The upshot of this is that our work is really a complement for the artificial intelligence approach, rather than a competitor for it.

The second approach I want to contrast ours with is the perceptron approach. Actually, it is much more accurate to say that Minsky and Papert's book Perceptrons[4] is what influenced us, rather than the perceptron approach. The following characterization of the perceptron approach is taken mostly from their book. The approach is like ours in that it emphasizes learning. A perceptron is in fact a simple sort of learning device. What it is supposed to do is come up with an answer to a complicated question after being given the answer to a lot of simpler questions. Suppose there are n of these simpler questions and each one is of the form 'does the predicate $F_i, 1 \leq i \leq n$, hold.' If it does, $F_i = 1$; if not, $F_i = 0$. The perceptron has n coefficients $a_i$, and it computes an answer to the complicated question, which is of the form 'does the predicate G hold,' by computing $\Sigma a_i F_i$. If the sum is greater than some number k, the perceptron answers yes ($G = 1$); if not, no ($G = 0$). It learns by changing the $a_i$'s, i.e., it is given an initial value for each $a_i$ and k and run through a number of trials, being told

[4] Marvin Minsky and Seymour Papert, Perceptrons (Cambridge: MIT Press, 1969).

the correct answer after each trial and alters the coefficients on this basis according to some preordained strategy. Machines of this type have a surprising amount of power. They can, in fact, given the appropriate predicates, learn to play championship checkers. It was widely believed a decade or so ago that they could learn just about anything. People clung to this belief even though it remained largely unsubstaniated, and this fact led Minsky and Papert to write Perceptrons, in which the inadequacy of perceptrons for certain tasks was clearly shown. This was done by showing that given a certain natural perceptual setup, which I will briefly describe, perceptrons cannot satisfactorily learn geometrical predicates.

This setup is a simplified model of the retina or a television camera. A two-dimensional plane is divided into squares (for the present purposes, the shape is inessential, and squares were chosen for convenience) and the processing device is told, for each square, that it is black or white. Given this information, it should be possible to compute the value of certain simple predicates $F_i$, and from these, the perceptron should be able to compute the value of a more complicated predicate G. The question now is how to characterize simplicity in this setup. One answer is that one predicate is simpler than another if its value depends on the color of fewer squares. Intuitively, it also seems desirable to localize these squares, e.g., requiring that they be adjacent. The first notion is sufficient, however, because Minsky and Papert showed that if G is the predicate 'connected,' then it is necessary for one of the $F_i$ to depend on the whole retina if a perceptron is going to be able to compute G correctly. This is completely unsatisfactory, since all the $F_i$'s must be simpler than G if the perceptron is going to be able to accomplish anything substantive. Thus, the setup must be altered in some way, and what we have done is replace both the model of the retina and the perceptron.

Instead of the above model of the retina, we decided to deal only with straight line figures, and to take the notions of straight line and intersection as primitive. I have already discussed the motivation for dealing with line drawings. The reason for having only straight lines is that we felt that solving this special problem would be a big step

toward solving the more general problem, and that this special problem was complicated enough. In its final form, we regard the learning device as simply being presented with all the information concerning the straight lines and their intersections. This is the reason that our model is applicable to tactile as well as visual perception. For, given a drawing with raised lines, a braille drawing, a person could gather the information we regard as being presented by touch. The fact that this would require motion, and hence take time, is not essential, since all we require is that the device at some time have all the information at its disposal, not that it gather it all at once. This will require some memory, but memory is necessary anyway. Moreover, if one accepts Helmholtz's hypothesis that movement of the eye is necessary to be able to perceive visual straightness,[5] there is no difference in the memory requirement for either type of perception. It is true that we do learn to recognize fairly accurately that some lines are straight without moving the eye, just as we can feel that some edges are straight without moving the hand. The above discussion is really about primitive visual straightness and primitive tactile straightness, i.e., the perceptual phenomena on which the idea of straightness ultimately rests.

Helmholtz's hypothesis is not universally accepted. Our work could be indirectly useful in establishing whether or not it is true. We were originally interested in the question of how people scan straight-line drawings. When presented with a drawing, people don't simply look at one point on it, but their eyes move back and forth across it. The motions used, and why they are used, are not well understood at all. It seems reasonable, if Helmholtz is correct, to expect that the primary purpose of some of the motions is to decide which lines are straight. This would require special eye movements.[6] If one knew what the other factors were which determined how a figure is scanned, it would be easy to recognize such movements. Our interest in the scanner (Chapter 2)

---

[5] Fred Roberts and Patrick Suppes, "Some Problems in the Geometry of Visual Perception," Synthese, 17 (1967), 177.

[6] Ibid., p. 178.

was motivated by a desire to know what some of these other factors were. We finally had to abandon the hope of solving this question when it became apparent that it was necessary to solve the problems we finally dealt with before there was any hope of dealing with the scanning problem. This problem is still an interesting one, both in itself and for the light it will shed on Helmholtz's hypothesis, and I believe our work will provide good background material for solving it.

We also discarded the perceptron as the model of the learning device. Instead, we took the finite state automaton (fsa) to be the model for what the learning device should be at asymptote. The justification for this move is discussed at length in Chapter 3. Once this move is made, the obvious problem is to find a way to code the information in a straight-line drawing in such a way that it can be put on the input tape of an fsa. Finding such a coding, discovering its geometrical properties and finding convenient methods for determining which predicates a given coding has are the central points covered in Chapter 2. It is necessary to have convenient methods for determining these predicates before attempting to build a device that can learn to recognize them. It is difficult to build a learning device when one knows the method and operations that it uses, it is virtually impossible otherwise. Thus, the material on this subject is an important step towards our goal. The operations we eventually used on the codings are set-theoretical in nature and would require the full power of a Turing machine, not an fsa, to execute. This is irrelevant as far as the coding problem is concerned, however, since the input tape of a Turing machine is exactly the same as that of an fsa. There were strong reasons for this switch, however, since it does have the unfortunate effect of creating a gap between the work we did on the coding and the work we did on learning. I will say more about this gap, and how it might be filled, in Chapter 4.

The last point in this section is that the effort to come up with a coding is interesting in itself, apart from the specific purpose of applying it to results in learning theory, which is what motivated us. It is obvious that the picture theory, which says that there are triangles in the brain when one is looking at triangles, is false. At least there

is no evidence to support the belief that it is true. If one believes that what is really present in the brain is different electrical states of the nerves, there is no reason to suppose there are triangles present. Indeed, it is difficult to see what such a triangle would look like, and not at all clear what would be explained by positing the existence of such a triangle. Moreover, one would really have to hold that the triangle was somehow transmitted bodily from the retina to the brain in order to make such a belief plausible; if what is transported is merely a coded electrical impulse from which the triangle is reconstructed in the brain, the brain might just as well operate directly on the coded electrical impulse, since it contains the necessary information. Besides, the motivation for believing there are triangles in the brain is that it is hard to see how to code a triangle in an electrical impulse, and such a coding must already exist unless the triangle is present at all points along the optic nerve. It will have to be able to jump across synapses, too. I'm not sure there is anybody who would actually hold such a theory, but it is a very natural way to look at the problem, so these remarks are perhaps worthwhile. Granting that there are no triangles in the brain, it is interesting to try to do geometry in strange contexts that resemble more closely what the actual coding might be. I believe the coding we use is closer to the actual coding, although it certainly isn't too close. It might have some of the same general properties, however. What I can say, though, is that thinking about our strange-looking coding has the very desirable effect of freeing one's mind from the picture theory, which has a tendency to influence one's thinking after it has been consciously rejected. This can be of philosophical value, as the discussion of abstract ideas in the next section indicates.

## Section 3

The purpose of this section is to justify the claim made in Section 1 that it is advantageous to deal with the philosophical problems of perception in PC. It consists of specific examples of problems which I believe are more appropriately dealt with in PC than in OE and a few general remarks on why I think this is the case. To provide a framework

for this discussion, the first order of business is to state the natural PC model of the perceptual process.

Speaking in a common sense way, perception is a process, at one end of which there is a physical object, e.g., a table (which I call the object), and at the other end what a person seeing the table is conscious, or aware, of (which I call the percept). In PP, this process is a continuous one, and hence extremely complicated and difficult to work with. In PC, however, the process can be broken down into four parts, as in the following diagram:

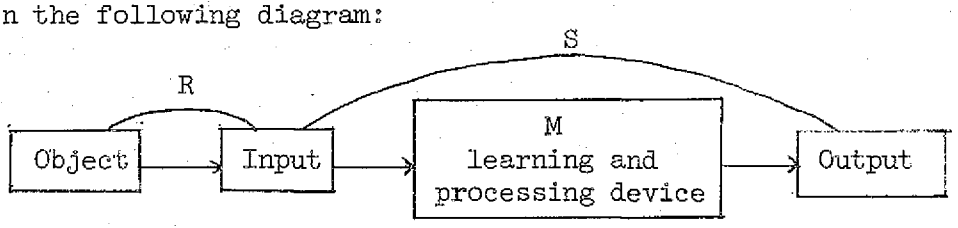



Figure 1

Components of the Perceptual Process

The way to understand the input to M is this: Imagine the description of a person looking at a table that would occur in PP. Light waves emanate from the table, are refracted at the crystaline lens, strike the retina, from which certain electrical impulses are transmitted to the brain, and a certain state of the brain results which corresponds to the percept. The percept is dependent on the person's previous learning (Hume and Kant would say 'experience', but it seems to me that the word 'learning' is more accurate), for it is a well-known fact that people with different backgrounds and training are aware of different things when looking at the same object. Thus, somewhere in the perceptual process, learning has to take place. It can't take place before the light waves hit the retina. Thus, somewhere in the retina, the optic nerve, or the brain, there is the first place at which learning can take place. The state of the electrical impulse right before it reaches this point is what corresponds to the input to M, for after this point what happens occurs in M, which is the learning device. The relation R between the object and the input is primarily subject matter for physicists and physiologists, and hence it should be studied in PP. Many of the ordinary philosophical examples of illusions, such as a stick in

water appearing bent and railroad tracks appearing to converge, concern themselves with R. Illusions that do arise because there is not a perfect correspondence between object and input do not cause any conceptual difficulties, as far as I can see, and thus do not seem to be philosophically interesting. The output of M corresponds to the percept, as it is the last thing in the process. The relation S between the input and the percept is dealt with in many of the psychological examples of illusions, such as the figure-ground distinction. This relation seems to me to be the philosophically significant one. The way to study it is to study M, and this is the main thrust of our work. Figure 1 represents all of what I called the perceptual component in the discussion at the beginning of Section 2. The learning and processing that occurs in M is primarily unconscious. The output of M is the input for the conscious processing device, which I referred to simply as the processing device in the earlier discussion. As far as the classical philosophical theories are concerned, this picture is closest to representative realism, since the input 'represents' the object.

The first problem I want to consider is the question of the perceptual given and in particular the question of sense-data. My sources are the first chapter in H. H. Price's Perception,[7] and Helmholtz's Physiological Optics.[8] Price is representative of the dominant themes in recent philosophy, while Helmholtz holds a more scientifically oriented view. They can be taken as arguing for opposing theses concerning perception, and hence data. This seems to be Price's position, for he mentions Helmholtz by name and purports to refute Helmholtz's theory. I don't believe that this is an accurate description of what actually occurs, however, I will give the reasons for this belief after discussing Price's argument.

The first order of business is to tell what a sense-datum is. Price gives some examples, and then says, "This peculiar and ultimate manner of being present to consciousness is called being given, and

---

[7] H. H. Price, Perception (London: Methuen & Company, 1954).

[8] H. von Helmholtz, Physiological Optics, trans. James P. C. Southhall (Menasha, Wisconsin: "The Optical Society of America", 1924-25).

that which is thus present is called a datum."[9] He then says he can't give a positive argument for the belief that there are sense data, but can answer arguments against this belief. There are two theses, an *a priori* one, and an empirical one, of which he says, "Either of these theses, if established, would be very damaging."[10] I will not discuss the *a priori* thesis, which is uninteresting. Price characterizes the empirical thesis this way: "This (thesis) maintains that it is in fact impossible to discover any data. For if we try to point at an instance, it is said, we shall have to confess that the so-called datum is not really given at all, but is the product of interpretation."[11] He attributes such an argument to idealists, but I think, in fact, it is easier to understand it if one is not an idealist, e.g., from Helmholtz's point of view. He then gives three arguments, and says "So far, we have been attacking the critics of the Given upon their own ground. And that ground is this. They begin by assuming that there is a distinction between 'the real given' or the given-as-it-is-in-itself on the one hand, and 'what the given seems to be' on the other."[12] He then gives his most important argument, which is that "The distinction between the Given as it really is and what the Given seems to be is altogether untenable."[13] I must confess that when the argument is put in these terms, I have difficulty in seeing how to resolve the issue one way or the other. However, it seems to me that the essential point of the anti-sense data argument is that there are two things in perception that must be kept clearly separate, and this is something which I believe is true. In the above model, the two things are the input to M and the percept. In more familiar terms, the two things are the actual stimulation of the retina, what Quine calls the "ocular irradiation pattern,"[14] and what we are actually conscious of perceiving. I note that the retinal

---

[9] Price, p. 3.

[10] *Ibid.*, p. 6.

[11] *Ibid.*, p. 7f.

[12] *Ibid.*, p. 9.

[13] *Ibid.*, p. 10.

[14] Willard Quine, *Word & Object* (Cambridge: MIT Press, 1965), p. 31.

stimulation doesn't necessarily correspond to the input, but using it makes the necessity for having two elements more obvious. That these two things are distinct, I think, is undeniable, but I will give some material from Helmholtz in support of it. This is the reason why I believe the anti-sense data position is easier to hold if one is not an idealist, since an idealist would have trouble making sense of the phrase 'retinal stimulation.' In a later chapter, Price mentions Helmholtz specifically while discussing causal theories of perception, i.e., theories that say we must infer what we are finally aware of. He says "The theory may say, with Helmholtz and others, 'You do infer but you are not conscious of inferring, because you do it so quickly and without any effort.' This will not do. If we are not conscious of inferring, what evidence is there that we do infer at all? And if it be replied 'Of course you do, for all consciousness of matter must be inferential,' we must point out that this begs the question."[15] The conclusion one is supposed to draw from the above paragraph is that there is no evidence that we do infer. The only thing I can imagine that Price had in mind when he wrote this is that it is obviously impossible to get any direct introspective evidence that there is an inference since the inference is, by hypothesis, unconscious. To conclude from this that there is no evidence is clearly mistaken, however. All it shows is introspective evidence is impossible, and thus that the evidence one has to adduce must be of a different, and in a way indirect, nature. There is a whole body of such evidence in favor of Helmholtz's view, much of it contained in Physiological Optics, which Price simply ignores. For example, consider the phenomenon of the blind spot. People simply fill in this hole in the visual field to look like the surrounding area. This certainly takes place unconsciously, and if Price's argument against Helmholtz is correct, it follows that we could have no evidence that this occurs. This is manifestly false. The only thing we can't have is direct introspective evidence.

It therefore appears at first sight that Price has put forward an extremely bad argument. If one takes him to being arguing against Helmholtz on Helmholtz's own ground, this is certainly the case.

---

[15] Price, p. 67.

Helmholtz was working in a scientific context, his language being a combination of PP and PC. If one puts Price's argument in this context, it is immediately obvious that it is a bad argument. The example of the blind spot was, after all, drawn from the scientific realm. There is a better explanation of what has occurred here, however. This explanation is plausible, puts things in context, and gives deeper insight into what was really going on. Moreover, it doesn't imply the highly improbable conclusion that Price is guilty of such an obvious blunder. In reading Price, one gets the unmistakable impression that Price's arguments are simply irrelevant to Helmholtz's position, not that they are bad arguments against Helmholtz. The reason is simply, to go back to what was said in Section 1, that Price is working on OE, while Helmholtz is working in PP and PC. Thus, Price's argument about unconscious inferences makes perfect sense in OE, which is where he is working, but is manifestly unsound in PC or PP, which is where Helmholtz is working. Thus, Price's error is not that he gives a bad argument against Helmholtz, but that he believes he is offering an argument against Helmholtz at all. In talking about perception, it is very easy to forget what context one is talking in and to assume that everyone is talking in the same context that oneself is unless one explicitly takes notice of the context in which the talk is occurring. This is the reason for the emphasis placed on PP, PC, and OE in this paper. I have found that such a framework is necessary if I am going to be able to keep things in their proper contexts. I believe Price's error consisted in thinking that Helmholtz was talking in OE. If what is said above is correct, it is understandable that Price should hold this view, and if it were true, his argument against Helmholtz would, in fact, be a reasonable one. Given that Helmholtz was talking in a scientific context, however, all arguments in OE are going to be irrelevant to his position. The only way to refute Helmholtz would be to argue in PP or PC, and Price has not done this. The upshot of all this is that Price's view is a reasonable one in OE, and that Helmholtz's one is reasonable in PP and PC. The crucial issue is where the philosophical problems lie, and Price simply assumes that they lie in OE. Thus, he fails to offer any arguments against someone who holds that philosophical problems lie in PC or PP.

In Section 24, Volume II of Physiological Optics, Helmholtz lists the results of experiments with color contrast. One particularly interesting feature is that contrast phenomena (which are illusions, since in them a uniform surface appears to have different colors) disappear if distinct boundaries are drawn between the two differently colored areas. Helmholtz says, "Incidently, it comes out plainly in the capricious result of these experiments how hard it is for us to make accurate comparisons of luminosity and colour of two surfaces that are not directly in contact with each other and have no border between them."[16] It is not surprising that not being directly in contact would have an adverse effect, but it is surprising that a sharply defined border would be so important. The reason is that people pay attention to color differences that aid them in dividing what appears in the visual field into different objects and ignore color changes that are no help in this.

In Section 26, Helmholtz gives the following as one of his basic principles in explaining the results of optical experiments: "We are not in the habit of observing our sensations accurately, except as they are useful in enabling us to recognize external objects."[17] This confirms the role assigned to color vision by the artificial intelligence people that was mentioned in Section 1. It is also, as Helmholtz remarks, one of the main goals of an artistic education to make people aware of these things they usually don't see. The surprising thing isn't that habit has led people to ignore some color differences, but that this habit can actually lead people to see different colors where only one really exists. As Helmholtz says, regarding contrast experiments, "If the inducing field is supposed to be an independent body, usually the contrast colour does not come out so as to be perceived."[18] If the two fields are not regarded as being independent bodies, then the phenomena appears. At the end of the section, he makes an interesting remark: "To those readers who as yet know little about the influence of psychic activities on our sense perception, it may perhaps seem incredible that through psychic activity

---

[16] Helmholtz, Volume II, p. 291.

[17] Helmholtz, Volume III, p. 6.

[18] Helmholtz, Volume II, p. 295.

a colour can appear in the visual field where there is none. The author must beg them to suspend judgment until they have become acquainted with the facts in Part III of this work, which will deal with sense-perception."[19] In particular, Section 26, the first section in Part III, is a very good discussion of the philosophical issues involved. In this section, there is a long discussion justifying the claim that what a person is conscious of is the result of an unconscious inference. It is an inductive type of inference, but even taking this into account, Helmholtz calls what happens in perception an inference only because what happens resembles an argument; there is a premise, the retinal stimulation, and a conclusion, what we are conscious of seeing. Thus, this view is essentially the view that there are two different things that must be distinguished in perception, and what one calls the connection between them is a terminological question. Calling it an inference seems as appropriate as anything else, and Helmholtz states this is the only reason he uses the term. Hence, Price's procedure of dealing with the view that there are two elements in perception and the causal theory separately is not very illuminating. To give a telling argument against Helmholtz's position would therefore require an argument against the claim that there are two distinct elements in the perceptual process. To do this would require very ingenious explanations of phenomena that seemingly can only be explained by making such a distinction and would be a very difficult task to accomplish. Price has not even attempted to do this.

A serious discussion of whether or not there are sense data requires some criterion for recognizing data. Such a criterion is given by Helmholtz, for after a long discussion, he says, "My conclusion is that nothing in our sense-perceptions can be recognized as sensation which can be overcome in the perceptual image and converted into its opposite by factors that are demonstrably due to experience."[20] In the terminology I have been using, this translates to "Nothing in our sense-perception can

---

[19] Helmholtz, Volume II, p. 295.

[20] Helmholtz, Volume III, p. 13.

be recognized as data which can be altered in the percept by learning." I have changed 'overcome and be converted into its opposite' to 'altered,' but, since it is the nature of a datum to be unalterable, the two statements have the same meaning. There are two types of illusions, those that disappear once we are aware of them, and those that don't. It is difficult to regard the former as being data, but there is a question about the latter, and it was to help answer this question that Helmholtz formulated the above criterion. His idea is that even illusions that don't disappear are not necessarily part of the perceptual data, for the effects of that which is the result of years of experience may not be negated by simply becoming aware of the fact that the habit does lead to illusion. The experiment where people were fitted with glasses that inverted the retinal image is a good example of the distinction that Helmholtz has in mind, and it also serves to support his position. When people first put on the glasses, they encountered all sorts of difficulty, and there was simply no way to overcome these difficulties by consciously inverting the visual field. After a few days, these difficulties disappeared, and things appeared upright. When the glasses were removed, the same difficulties reappeared, and again disappeared after a few days. This shows that what we are conscious of can be changed by experience, and hence Helmholtz's criterion, which may seem innocent at first, would probably rule out colors and a lot of other things as being perceptual data once the appropriate experiments are performed. For example, experiments of the above type with contrast phenomena, if possible to perform, would probably show that color perception is also due to learning. Certainly, if Helmholtz's explanations are correct, this would be the result.

Helmholtz's criterion seems to me to be the best one for determining what data are. If one applies it to the PC model of perception, the input to M is what should be called the data. The percept cannot be the data, for it can be altered by learning. Thus, instead of having the object, input and output as in Figure 1, the terminology should be object, sense data, and percept.

It is now easy to state Price's position in this context. His position is that the sense-data and percept are identical, for he identifies the data with what we are conscious of, which is the percept. In OE, these things are not clearly distinguished, so Price's arguments have force, but in PC, they are obviously distinct, so that the most ingenious arguments lack force.

In terms of Figure 1, we now have a clear picture of what a sense-datum is. This gives a clear framework for talk about sense-data, and allows one to talk precisely concerning them with a minimum of effort. Disputes about sense-data are thus easy to state, and time is not wasted in preliminary skirmishing whose main outcome is to make the issue precise. This framework also provides a touchstone for easily evaluating arguments concerning sense-data, where otherwise it is difficult to evaluate such arguments. Finally, this framework provides a context in which the disputes concerning sense-data might be solved to the satisfaction of everyone. It should make it easier to gain knowledge concerning specific problems by saving time that might otherwise be wasted in dealing with ill-defined problems.

The next problem is the problem of synthetic a priori knowledge. I am not going to discuss the general question of whether or not there is synthetic a priori knowledge, but limit myself to the particular question of whether or not geometrical knowledge is synthetic a priori. The answer to the particular question will influence to a great extent the answer to the more general question, since geometry is one of the most likely candidates for the status of synthetic a priori knowledge. Moreover, much of what is said concerning geometry will be applicable to other areas of knowledge. Besides limiting the discussion to geometrical knowledge, I will consider from all that has been written by philosophers on this question only the view of Kant. His view is the most important, however, since what he said influenced all the subsequent developments. Thus, this isn't really a severe limitation.

To show that geometrical knowledge is synthetic a priori, one has to establish the two independent claims that it is synthetic and that it is a priori. Thus, the discussion naturally breaks down into two parts. At the time Kant wrote, it was generally believed that geometrical

knowledge was not synthetic, but that it was a priori. It seems to me that the claim that geometrical knowledge is synthetic is less disputable than the claim that it is a priori, however. Thus, I believe Kant is correct in holding this knowledge to be synthetic, but that what he says about its being a priori needs some clarification and modification. It is for this latter task that our work is pecularily suited.

Kant believes that it is obvious that mathematical judgments are synthetic if one thinks about them. He thinks that previous thinkers had simply overlooked this fact. They were led to believe that mathematical judgments were analytic because of the prominent role deductive inference plays in mathematics. However, "This was a great mistake, for a synthetical proposition can indeed be established by the law of contradiction, but only by presupposing another synthetical proposition from which it follows, but never by that law alone."[21] Kant divides mathematical judgments into two classes, arithmetical and geometrical. He argues that arithmetical judgments are synthetic, and then says, "Just as little is any principle of geometry analytical. That a straight line is the shortest path between two points is a synthetical proposition. For my concept, a straight line contains nothing of quantity, but only of quality. The concept 'shortest' is therefore altogether additional, and cannot be obtained by any analysis of the concept 'straight line.' Here, too, intuition must come to aid us. It alone makes the synthesis possible."[22]

I believe Kant is entirely correct in believing that geometrical judgments are synthetic. Certainly, they are not logical truths, and they are not true by definition. My own view is that they have exactly the same status as the basic principles of any theoretical science, e.g., Newton's three laws of motion. Certainly they were discovered empirically, being based on Egyptian surveying techniques. Granting that Kant is correct on this point, there is one point of difference I have with him. Kant believes that the term 'synthetic' applies to individual

---

[21] Immanuel Kant, Prolegomena to any Future Metaphysic (Indianapolis: The Liberal Arts Press, 1950), p. 15.

[22] Ibid., p. 16.

judgments. I think Quine is correct in saying that this is not a good way to use the term, but that it should be applied to much larger units than individual judgments.[23] Quine says it is the whole of science, but for the present purposes, all I want to maintain is that the term should apply to all of a geometry rather than its individual propositions. For instance, given the present state of affairs, where there is more than one geometry, I think it would be wise to use the term 'synthetic' in the sentence

"Euclidean geometry is synthetic."

but not in the sentence

"The proposition that the sum of the three angles of a triangle equal 180° is synthetic."

It seems to me that this point is quite unobjectionable, for Kant and subsequent philosophers, even though they speak of individual judgments, or sentences, as being synthetic, believe that all geometrical judgments go together; if one is synthetic, then they all are, and vice versa. Thus, I believe that it is simply an unfortunate oversight that Kant applies 'synthetic' to individual judgments, for there is really nothing in his system to lead him to do this. It is unfortunate because it focuses one's attention on the wrong thing, and thus is very misleading. The same point applies to the term '*a priori*' as well, and it is important for my discussion of this term.

The prevailing attitude at Kant's time was that geometrical judgments are *a priori*. Thus, Kant's main concern is to show that they are synthetic, for then he will have examples of synthetic *a priori* knowledge. The criterion for deciding if a judgment is *a priori* is to see if it is necessarily true. Thus, speaking of two principles of physics, Kant says "Both propositions are not only necessary, and therefore in their origin *a priori*, but also synthetic."[24] Kant follows Hume in believing that any proposition that is known from experience, i.e., empirical knowledge, cannot be necessarily true. Kant's main concern is not in showing that

---

[23] Willard Quine, *From a Logical Point of View* (New York: Harper & Row, 1963), p. 42.

[24] Immanuel Kant, *Critique of Pure Reason* (trans. Norman Kemp Smith) unabridged edition (New York: St. Martin's Press, 1965), p. 54.

geometrical judgments are synthetic a priori, but in answering the question of how this is possible.[25]

The fact that geometrical judgments are synthetic a priori plays an important role in Kant's system. If one is interested simply in the question of whether or not there is a priori knowledge, then the question of the status of arithmetic is independent of the question of the status of geometry, and hence it might be thought superfluous that Kant mentions geometry specifically. This is not true for two reasons. First, for the actual developments in the Critique of Pure Reason, Kant needs to assume that both kinds of knowledge are synthetic a priori, for he thinks we have two types of intuition, inner (time) and outer (space). Moreover, arithmetical judgments are supposed to be based on inner intuition, geometrical judgments on outer intuition. The connection between inner intuition and arithmetical judgments is very nebulous indeed, while the connection between outer intuition and geometry is completely straightforward. Secondly, it intuitively seems more plausible to believe that arithmetical judgments are analytic than that geometrical judgments are. To convince someone who believes that arithmetical judgments are analytic that there is synthetic a priori knowledge depends entirely on convincing him that geometrical judgments enjoy this status.

I don't believe Kant is entirely correct in believing that geometrical judgments are a priori. First, as remarked above, I think the term should be applied to all of geometry, not to individual judgments. Apart from this, the fact that non-Euclidean geometries have been discovered and used in physics indicates that Kant is wrong. Another indication is given by the nineteenth-century debate between Hering and Helmholtz. Hering, among others, tried to construct a scientific theory of visual perception on the basis of Kant's theory. This theory took the way we perceive space as being given, rather than learned, and hence is called by Helmholtz the intuition theory. Opposed to this was the empirical theory, whose chief proponent was Helmholtz, which held that we must

---

[25] Ibid., p. 55.

learn to preceive space. Helmholtz was a clear winner in this debate, for it is very difficult to explain some illusions on the basis of the intuition theory, while the empirical theory explains them nicely. These two developments show that Kant's particular theory is wrong, but they don't show that his approach is wrong, i.e., that a theory similar to Kant's is wrong. I think this latter question is the important one.

In terms of Figure 1, geometrical knowledge is indicated by the appropriate outputs of the learning and processing device, M.[26] This output depends on the input and the internal structure of M. Since knowledge consists of the ability to give the appropriate response to any input, geometrical knowledge is a property of the state of M. M will change as learning takes place, but its knowledge at any point in its experience is a property of M at that point. The structure of M at any given point is determined, perhaps only probabilistically, by the original structure of M before any learning has taken place, and by the history of the inputs M has received. Perhaps in a growing organism it will not be easy to separate the original structure from the history of inputs; it is possible that some learning might take place before the processing device, perhaps area 17 of the cortex, is fully developed. I believe that such a factor is epistemologically irrelevant, and that there is no need to take it into account in the present discussion. Certainly, Kant doesn't consider such a factor.

In this context, the best way to interpret a priori is not as an all-or-none predicate, but as a matter of degree. I am suggesting, in other words, that it is best to treat a priori the same way that Quine treats synthetic. Thus, geometry will be more a priori the more the state of M, once it has acquired geometrical knowledge, is determined more by the original structure than it is by the history of inputs; it will be more a posteriori the more the history of inputs determines the state. Given this interpretation of a priori, Kant's theory is that the state of M is determined entirely by the original structure. Kant can

---

[26] The following discussion is due to a suggestion of Professor Julius Moravcsik, who pointed out that the claim that the problem of synthetic a priori knowledge is similar to the problem of innate ideas needs some justification.

allow for some sort of 'learning' by saying that it requires some experience to actualize geometrical knowledge. On my interpretation, this amounts to saying that M changes, and that these changes require inputs, but that how M will change is determined entirely by the original structure. This sort of 'learning' theory is popular with all kinds of intuition theories, be they theories concerning morality or causality, and is not peculiar to Kant. I personally can't see any justification for such theories, but this is largely irrelevant to the present point, since, as mentioned above, Kant's theory has quite conclusively been refuted. The opposite extreme from Kant's view is that the state of M is determined entirely by the history of inputs. This view has been stated historically by saying that the mind is a tabula rasa. This view is completely impossible, as M has to have some structure in order to learn from the inputs it receives. The true view is located somewhere between these two extremes. If the original structure actually does determine geometry to a great extent, which certainly seems plausible, then I believe it is fair to say that Kant was essentially correct, i.e., that he had the right approach. I don't know whether or not this is the case, but it certainly is a possibility.

The problem now is to decide what the correct mixture of original structure and history of inputs really is. One way would be to alter the history of inputs to different devices, and see how different the resulting geometries are. Psychological experiments in weird perceptual conditions can be regarded as attempts to do this. This is not the correct way to approach the problem, in my opinion. Such results will be quite fragmentary, unless a person's perceptual conditions are completely altered all the time. Otherwise, these results show only what happens when a device that has already learned ordinary geometry is briefly exposed to differing inputs, and this is a much different situation than actually learning geometry entirely under different conditions. Moreover, the experiments I am thinking of are like the one with glasses that inverted the retinal images, which do not naturally lead a subject to different geometrical assumptions. It would be interesting to see experiments that could have this outcome. Moreover, these results can at best give only a partial answer until an at least approximate

characterization of the original structure of M is available. It is my belief, once such a characterization is available, that such experiments will not be necessary. My suggestion is that given the characterization, it should be possible to determine what constraints the original structure puts on the geometries that could possibly be learned. I think this could be done in a way similar to how logicians treat the problem of how categorical a set of axioms is. Roughly speaking, an original structure will be less categorical the fewer different possible geometries that it allows. My proposal is to say that a geometry is more a priori the more categorical the original structure that learned it is.

There is one refinement of this view that seems desirable. Rather than saying that all the geometries that a certain device can learn are equally a priori, it seems plausible to order these geometries according to the ease with which they are learned. Thus, the geometries that are learned more easily, which are more natural for the device, would be regarded as being more a priori than geometries that are difficult to learn, that are unnatural. For people, this would result in saying that Euclidean geometry is more a priori than non-Euclidean geometries, which is intuitively appealing.

Note that now a priori is not used as an absolute term, but must be relativized to a particular learning device. Thus, one has to say 'a priori for M,' for example, rather than simply 'a priori.' This is how the term should be used, for it is clear that two devices could have exactly the same geometry, and that it would be almost completely a priori for the one and not very a priori at all for the other.

This relativized use of the term may seem a little strange, but actually, there is a good explanation of why it hasn't been used in this way. When the term is used, it normally means 'a priori for people,' and it is tacitly assumed that all people closely resemble one another in the way they learn geometry. This use of the term is simply a special case of the more general use that I advocate, and it seems to me that the great concentration of attention on this one case is what led people to overlook the fact that a priori is actually a relative term.

This way of looking at the term 'a priori' is very similar to the way Chomsky looks at the term 'innate.' Innate, as it was used, e.g., in Descartes, applied originally to specific ideas, such as the idea of God.[27] Chomsky changes this use completely, and talks of the innate abilities of the "acquisition device," which is the device that learns language. Thus, he applies the term to the original structure of the acquisition device which is how I have used the term a priori, since it depends on the categoricalness of the original structure. This should not be too surprising, for both 'innate' and 'a priori' were used as opposites for 'empirical.' It seems to me that the only reason for two terms was that they were thought to apply to different things. If one agrees with Quine, then it is wrong to apply such a term to either ideas or individual judgments. Both should be applied to whole theories, so that the differences between them collapse, and they become synonymous.

The first two problems I have considered dealt with learning and their solutions depend on a more fully developed theory. The third problem I will discuss has neither of these features; it doesn't deal with learning, and I believe a solution (at least for the admittedly limited context we are working in) does not require any further theoretical developments. In fact, I propose such a solution at the end of Chapter 2 after the necessary preliminary work has been discussed. This problem arose in the British empiricists' discussion of abstract ideas.

Locke, Berkeley and Hume thought that whenever one thought about a proposition concerning the abstract idea of a triangle, what one was actually doing was considering the image of a particular idea that one had in his head. The problem with this was that this particular idea had to have definite properties, such as being a definite size, while the abstract idea should have no such definite properties. They were unable to solve this problem. This is not surprising, for their theory was a form of the picture theory discussed in Section 2, and it seems to me that there is no way to solve their problem if one accepts the picture

---

[27] Rene Descartes, "Meditations," Descartes' Philosophical Writings, trans. Norman Kemp Smith (London: Macmillan & Co., 1952), p. 215ff.

theory. The solution I propose rejects the picture theory, and identifies the abstract idea of triangle with a procedure for recognizing triangles. The philosopher who is closest to this conception is Kant.[28] His notion of a schema of a concept is very close to what I have in mind, as is evident from his definition: "This representation of a universal procedure of imagination providing an image for a concept, I entitle the schema of this concept."[29]

These three examples show the utility of working in PC. I now want to offer some general considerations that support the same point. The first consideration is implicit in what has been said before; namely, that the really difficult problems do not arise if one confines oneself to OE. I think Wittgenstein is correct in believing that philosophical problems are the product of confusion, if one restricts one's attention entirely to OE. Moreover, he is also correct in believing that philosophers have contributed more to creating these problems than they have to solving them. Certainly, the ordinary user of OE sees no serious difficulties, and I am also unable to locate them. The problem as far as perception is concerned, I think, is that philosophers have taken a problem that is essentially scientific in nature, e.g., the problem of sense data, and tried to find a solution in OE rather than in the scientific context in which the problem arose. This is why the issue of sense data, for example, immediately becomes clearer when one puts it back in PC. The first step in this philosophical approach is to reject the scientific solution to the problem. This is a necessary step in order to get the 'philosophical,' as opposed to the scientific, inquiry underway, and philosophers have been aware of this fact. Price felt compelled to refute the scientific theory, Helmholtz's causal theory, before giving his own analysis. We have already seen the inadequacy of his argument against Helmholtz. G. E. Moore is another example of a philosopher who felt this step was necessary. Towards the end of his

---

[28] This was also pointed out to me by Professor Moravcsik.

[29] Kant, Critique, p. 182.

paper Some Judgments of Perception,[30] Moore mentions two different ways of characterizing the problem of perception. I will give these in the opposite order he gives them, mentioning the one he takes, which is the one that has been widely discussed by philosophers, first, since it is a more self-contained statement. He says, "The only other suggestion I can make is that there may be some ultimate, not further definable relation of 'being a manifestation of,' such that we might conceivably be judging: "There is one and only one thing of which this presented object is a manifestation, and that thing is part of the surface of an inkstand.""[31] The first possibility he mentions is, "It might, no doubt, be possible to define some kind of causal relations, such that it might be plausibly held that it and it alone causes this presented object in that particular way. But any such definition would, so far as I can see, be necessarily very complicated."[32] This is all he says on the matter. Thus, the relation between object and percept can be taken to be a complicated causal one or a simple, unanalyzable one of some unspecified sort. Moore gives no explicit reasons for rejecting the former view, but what he says implies he does it simply because the view is complicated. This is understandable, since as mentioned above, the scientific theory Moore is thinking of is a theory in PP, which is too complicated to deal with effectively. This is no longer true, since PC is available. Thus, neither philosopher has a convincing argument on this point, and it seems to me that until such an argument is given, there is no reason to believe that the philosophical problems of perception are problems of OE. There is good reason to believe that some of them aren't, as I have indicated above.

The belief that philosophy should be done in OE and that there is a sharp distinction between philosophy and science is a twentieth century phenomenon. It arose earlier in this century and is due in large part to the great influence of Moore. In work before this time, no such sharp distinction was drawn. Thus, in the works of philosophers like Descartes, Locke and Berkeley, and scientists like Helmholtz, one finds no distinction

---

[30] G. E. Moore, "Some Judgments of Perception," Perceiving, Sensing, & Knowing, ed. Robert J. Swartz (New York: Doubleday & Company, 1965).

[31] Ibid., p. 26.

[32] Ibid., p. 26.

between philosophic questions and scientific questions. More importantly, all four treat some questions that a philosopher like Moore would regard as scientific next to and in the same way as they treat questions he would regard as philosophic. For example, Locke mentions Molyneux's problem concerning a man born blind, and Berkeley discusses this problem, Dr. Barrow's problem and why the moon appears larger on the horizon, in an Essay Towards a New Theory of Vision, which I think is the most interesting work written by a philosopher in the area of visual perception. The philosophical nature of some of Helmholtz's remarks has already been discussed. The fact that this earlier work was much more fruitful than the twentieth century work is a powerful reason for accepting the earlier view. Thus, when things are put in their proper historical perspective, it is seen that the position I take in regard to this question is much closer to the classic traditional position than is the view that philosophy should be done in OE. I therefore feel that a person who holds this latter view is actually under stronger obligation to defend his approach than I am to defend mine. I have pointed out above that our approach resembles in many ways the approach to linguistics which was initiated by Chomsky. It is interesting, as Chomsky himself points out in Cartesian Linguistics, that his approach is a return to ideas that were prevalent before this century but that had been rejected in the early part of this century. Moreover, the close analogy between perception and linguistics, which has been mentioned several times, itself has a long tradition, going back to Berkeley's conception of what is perceived as being the language of nature.

There is one final point concerning the philosophical significance of the present work that should be made: it is not necessary that one should share my view that the main problems of philosophic interest are problems of PC in order to maintain that work such as ours is philosophically significant. It seems to me that the view that not all philosophical problems are problems of OE is sufficient. Once one accepts this view, he is immediately struck by the fact that almost all the philosophers working on the problem of perception are working in OE. This seems like a misallocation of effort, particularly since there is no

really convincing argument that philosophers should confine their attention to OE. It is usually not very fruitful for everyone to adopt the same approach, even if the approach is basically correct. It is worthwhile to have people espousing the opposite view, since this will at least serve to keep those who hold the majority view from lapsing into dogmatic slumbers. I hope the present work at least has this minimal effect.

This concludes the remarks concerning the motivation and background for the work we did. The rest is an account of the specific model we worked on, except for the section at the end of Chapter 2 concerning abstract ideas.

# CHAPTER 2

## CODINGS

As mentioned in Chapter 1, we worked on two specific problems, the coding problem and learning. However, it took us some time to accomplish this division. Even after making the division, we went back and forth between the two problems, but for ease of presentation, I am going to treat them separately. The work on each problem will be presented roughly in the order it happened, and the things we tried and found wanting will be included. This is probably the best way of explaining what we finally did, and in any case it will provide the background and motivation for our eventual problem. This chapter deals primarily with the coding problem, the next with learning.

We decided at the outset to restrict our attention to two-dimensional straight-line figures. To get things started, we limited ourselves to figures in which at most two lines intersect at each point, figuring it would be wise to try to solve this simpler problem first, and later on try to remove this somewhat artificial restriction. ('Figure' will henceforth mean a figure of this type.) What we wanted was a device that could learn a geometrical predicate applicable to such figures by going through a series of trials, where on each trial it is presented with a figure, responds yes or no, and then is told the correct answer (i.e., whether or not the figure presented on that trial did in fact have the specified predicate). It would be nice if the device could learn several predicates this way and then use them to learn more complicated predicates, but the basic situation is when the device has no geometrical predicates at its disposal. Using the eye-brain combination as our model, we came

up with a device with three components: scanner, memory, and processing device, the first corresponding to the eye, the last two to the brain. For convenience in discussing the following problems, we lumped the last two together and called the result 'automaton.' This was due to the fact that, in the present context, the internal structure of the memory and processing device isn't important, only its input/output, and this resembles that of a finite state automaton. Intuitively, the automaton needs three abilities: on each trial it has to be able to use the scanner to acquire information, use this information to get an answer, and have the ability to learn from trial to trial so that it will eventually get all right answers.

Keeping this rather vague idea of the automaton in mind, the next task was to explicitly characterize the scanner it had at its disposal. The scanner has to be able to receive and execute instructions from the automaton and to report back what it sees. We decided to work with polar rather than rectangular coordinates since it is more natural to think of the eye moving a certain distance in a specified direction rather than moving up a certain distance and then over another distance. The retina of the scanner is a small circle with a special point X in the middle, which corresponds to the fovea. One of our main concerns in writing it out was to allow for perceptual error and indeterminancy, e.g., forming only a rough idea of the size of an angle, and it would be fairly easy to put these things in the following device:

1. Await instruction: Search $(\theta,r)$ go to 2
   Follow $(\theta)$ go to 6
   Automaton cannot order Follow unless special point X is on line.
2. Move in direction $\theta$ until: line appears on retina go to 4
   distance r is covered go to 3
3. Report Miss to automaton go to 1
4. Move so that special point is on the nearest vertex if one appears in retina, or on the nearest line if not go to 5
5. Report distance moved since last report, type of vertex special point is on according to pictures and angles lines make with horizontal

   a ——  b ∠  c —/  d /  e —  go to 1
6. Move special point along line nearest to angle $\theta$ until vertex is reached go to 5

The advantage in writing out the scanner explicitly is that it makes it possible to state presuppositions and distinguish separate problems. First, it makes clear the sense of what I said in Chapter 1 about how we took straight as primitive. We present the scanner with straight lines only and simply give it the ability to follow them. Thus, its output is the same as that of an organism that recognizes and follows straight lines. Moreover, this makes it clear that the scanner doesn't correspond exactly to the eye, since the eye can't follow straight lines by itself. Secondly, it is clear that the automaton is going to have to be able to recognize when the scanner returns to a vertex that it has already reported on. The main reason for having the scanner report all the distances and angles was to give the automaton enough information to do this. It now occurred to us that it would be wise to abstract from this problem, and simply assume that the automaton has the ability to recognize the same point every time the scanner is on it without worrying how it accomplishes this. This simplifies the automaton, but of even more importance in the present context is that it allows one to greatly simplify the information that the scanner gives the automaton. Many geometrical predicates, e.g., closed, connected, triangle, do not depend on the lengths of the particular line segments or sizes of the particular angles involved. We decided to study these predicates, and hence dropped the reports of distances and angles from the output of the scanner. Finally, it allows one to formulate the scanning problem (how people scan figures) that was mentioned in Chapter 1. In this context, it simply becomes the question of how the automaton decides what instruction to give the scanner each time it reports. We were originally very interested in this problem for several reasons. It is not clear what a good method of scanning would be, or how people do it, let alone how people learn to do it. Moreover, it seems clear that the method would vary according to what is being looked for and what has already been found: e.g., one would look differently if one wanted to know whether a figure contained a triangle than if one were interested in whether it was connected, and if one were interested only in whether the figure contained a triangle, he would stop scanning it after he found one. Moreover, it seemed plausible to believe, in our particular case, that a good method of scanning would simplify the learning. For instance, it would

be easy to recognize polygons if the scanner went around the perimeters of polygons. We failed completely to make any headway in our attempt to solve the scanning problem (about the only thing we could agree on was that it would be good to use Follow rather than Search wherever possible), and a little reflection convinced us that it was mistaken to tackle it in the first place; it became apparent that as long as the whole figure was scanned and the appropriate information stored, it made little difference how this information was obtained as far as processing it was concerned. Thus, it seemed prudent to keep the problems of how the automaton used the scanner to acquire information and how it processed this information completely distinct, and concentrate our efforts on only one of them. From consideration of examples like those mentioned above, we concluded that processing the information was the first problem to be solved, since efficient scanning depends on knowing what to look for and hence having some geometrical predicates already at hand. Besides, processing the information seemed like the more interesting problem, and ignoring how it was acquired allowed us to forget about the scanner altogether. We now simply regarded all the information, excluding distances and angles, which could be gotten from the scanner as simply given to the automaton. The problem now was to decide in what form this information should be given, i.e., how it should be encoded.

The following suggestion and elementary results are due to George Huff. Before giving the formal definition, let me first give an example. Take the figure  . The idea is to label each vertex with capital letters, say  , and put one element in the coding for each line, i.e., {AG, GCH, ABCDE, BF, HD} would be a coding for the figure. One could substitute GA for AG and still have a coding, but it is important to have only one element of the coding for each line in the figure. To exclude the possibility of putting two elements for each line into the coding is the reason for ordering the vertices in the original figure. Any other method of achieving this would also be satisfactory. Morover, it would also be possible, in the above example, to label the vertices a different way, as long as one uses A-H, and to get different codings this way. All such codings would be equivalent, as the following theorem shows.

Suppose X is a figure with vertices $v_1 \cdot v_2 \ldots v_n$ simply ordered in some way.

Definition: $\chi$ is a coding for a figure X if $\chi$ is a set, and there is a 1-1 function $C_{(\chi,X)}$ which maps the vertices of X onto an initial segment of the Roman alphabet of capital letters (with its usual order, subscripts if necessary) such that $\chi = \{\overline{C(v_{i_1})C(v_{i_2})\ldots C(v_{i_m})}: \overline{v_{i_1}, v_{i_2} \ldots v_{i_m}}$ denotes a line in the figure X with endpoints such that $v_{i_1}$ precedes $v_{i_m}$ in the ordering of the vertices and central vertices $v_{i_2} \ldots v_{i_{m-1}}$ in that order from $v_{i_1}\}$.

Definition: $\chi RX$ iff $\chi$ is a coding for X.

Definition: $\chi \equiv \gamma$ iff there is a figure X such that $\chi RX$ & $\gamma RX$.

Note that "$\equiv$" is reflexive and symmetric.

Lemma: $\chi \equiv \gamma$ iff there is a permutation p of the vertices of $\chi$ such that $p\chi = \gamma$, where $p\chi = \{\overline{p(A_{i_1})\ldots p(A_{i_m})}: \overline{A_{i_1} \ldots A_{i_m}} \in \chi\}$.

Proof: $\Rightarrow$ Denote the vertices of $\chi$ by $A_1 \ldots A_n$, those of $\gamma$ by $B_1 \ldots B_n$. $\{A_1 \ldots A_n\} = \{B_1 \ldots B_n\}$. Suppose there is a figure X such that $\chi RX$ & $\gamma RX$. Define p by $p(A_i)=B_j$ iff there is a vertex of X such that $C_\chi(v)=A_i$ & $C_\gamma(v)=B_j$. Clearly p is a permutation of $A_i \ldots A_n$. Suppose $\overline{p(A_{i_1})\ldots p(A_{i_m})} \in p\chi$. Then $\overline{A_{i_1} \ldots A_{i_m}} = \overline{C_\chi(v_{i_1})\ldots C_\chi(v_{i_m})} \in \chi$, so $\overline{v_{i_1} \ldots v_{i_m}}$ is a line in X. Therefore $\overline{C_\gamma(v_{i_1})\ldots C_\gamma(v_{i_m})} = \overline{B_{j_1} \ldots B_{j_m}} \in \gamma$. But $\overline{p(A_{i_1})\ldots p(A_{i_m})} = \overline{B_{j_1} \ldots B_{j_m}}$. Hence, $p\chi \subseteq \gamma$. Similarly for $\gamma \subseteq p\chi$.

$\Leftarrow$ Suppose there is such a permutation p. Let X be such that $\chi RX$. Want to show $\gamma = p\chi$ is such that $\gamma RX$. Define $C_\gamma$ as follows: $C_\gamma(v)=p(C_\chi v))$. Let $\gamma_1 = \{C_\gamma(v_{i_1})\ldots C_\gamma(v_{i_m}): v_{i_1} \ldots v_{i_m}$ is a line in X$\}$. Show $\gamma = \gamma_1$.

If $\overline{C_{\mathcal{Y}}(v_{i_1})\ldots C_{\mathcal{Y}}(v_{i_m})} \in \mathcal{Y}_1$, then $\overline{v_{i_1}\ldots v_{i_m}}$ is a line in X, so $\overline{C_{\mathcal{X}}(v_{i_1})\ldots C_{\mathcal{X}}(v_{i_m})} \in \mathcal{X}$, hence $\overline{p(C_{\mathcal{X}}(v_{i_1}))\ldots p(C_{\mathcal{X}}(v_{i_m}))} = \overline{C_{\mathcal{Y}}(v_{i_1})\ldots C_{\mathcal{Y}}(v_{i_m})} \in p\mathcal{X}=\mathcal{Y}$. Thus $\mathcal{Y}_1 \subseteq \mathcal{Y}$. Similarly $\mathcal{Y} \subseteq \mathcal{Y}_1$.

Corollary: "$\equiv$" is an equivalence relation.

Proof: Need to show transitivity. If $\mathcal{X}\equiv\mathcal{Y}$ & $\mathcal{Y}\equiv\mathcal{Z}$, then there are permutations $p_1, p_2$ such that $p_1\mathcal{X}=\mathcal{Y}$, $p_2\mathcal{Y}=\mathcal{Z}$. Thus, $p_2p_1\mathcal{X}=\mathcal{Y}$ Q.E.D.

It is now possible to define a mapping S from the set of figures into the equivalence classes of codings by $S(X)=[\mathcal{X}]$, where $\mathcal{X}$ is such that $\mathcal{X}RX$. S induces an equivalence relation on the set of figures defined by $X\equiv Y$ iff $S(X)=S(Y)$, i.e., $X\equiv Y$ iff any pair of codings for X and Y are equivalent. What these equivalence classes look like and what group of transformations the above equivalence relation remains invariant under we do not know. Certainly the group of transformations is not one of the ordinarily geometrically significant groups.

We decided to take this coding as our final formulation and work with it. At this point, we spent some time working on learning, the results of which are in Chapter 3, before returning to the coding.

The next question we asked was what would be a good way to recognize geometrical predicates given only the coding. The answer would determine to a large extent what we wanted our learning device to look like at asymptote, and hence this is a crucial, and, as it turns out, quite interesting, question. We concentrated on the predicate 'triangle,' or 'triangle in context,' which is true if and only if the figure contains a triangle. Given a figure that consisted only of a triangle, the three vertices will be labeled A, B, and C. The natural thing to do in this case would be simply to teach the device to recognize the pattern AB,BC,CA. This could be done by making the coding the input tape for a fairly simple finite state automaton, which was highly desirable because we had been dealing with finite state automata in our work on learning theory.

There is a problem even in this simple case, however. It is possible that the figure is coded AB,BC,AC, or in some similar way. It occurred to us that a good way to avoid this problem would be to put the coding

into some canonical form, but there is no natural way to do this for even moderately complex figures and our efforts produced no way at all. Thus, the device has to be able to recognize when different codings are equivalent, and so already it is no longer a simple pattern-recognizing device. Moreover, in order to deal with figures containing central vertices, the device needs the ability to break a line with one or more central vertices into all its possible segments. For example, in the coding [ABCD,AE,CE,BF] it would have to find the segment AC before it could find the triangle ACE. This example also illustrates the point that in complex figures the vertices of a triangle could be labeled by any three letters, so that the device has to be able to recognize the same pattern regardless of the actual letters. Finally, it is clear that if one simply gives the coding to a finite state automaton as its input tape, the automata will have to have more states the more complicated the figure becomes. Thus, the automaton will have to grow as the figures do, and hence it will be impossible to have an automaton with any fixed number of states that will be able to recognize triangles in all contexts. Hence, automata have much the same deficiency as perceptrons. Thus, this natural approach has many difficulties and its main attraction, that it connects naturally to the work we did in learning theory by way of finite state automata, turns out to in fact lead to the very difficulty we set out to solve, which is that no perceptrons of fixed complexity can recognize predicates such as 'connected.'

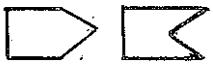
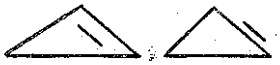
In the above example, it is clear that BF and CD are not on any triangle merely from the fact that F and D occur in only one line. We called such segments 'legs,' and noted that in cases besides the present one legs are irrelevant in the sense that they can simply be deleted without changing the value of the predicate. Thus, we thought that the device should have the ability to delete these legs. We were still at this point thinking of the device as sort of running through all possible combinations of three line segments looking for triangles, and allowing it to ignore some line segments would result in a great saving of effort. This rather simple idea led to a rather fundamental change in our thinking: instead of thinking of the device as a finite state automaton and worrying how it could learn to make the correct transitions, we thought of the

device as being able to perform simple operations and learning consisted in combining these simple operations into more complex operations that would be able to recognize the appropriate predicates.

At this point, we came up with a much better way of recognizing triangles, which confirmed this change in our thinking. In final form, this method consisted of picking a line in the figure, taking the set of lines that crossed this line, and checking to see if any vertex occurred on two of these lines: if one did, there was a triangle, if not, the original line could be deleted and the process repeated to see if there were any triangles in the whole figure. Moreover, if a vertex is found that does occur in two lines in the set, it is easy to find the other two vertices of the triangle, and hence if the device has an output mechanism, it can in fact list all the triangles in the figure. This method indicated to us that it would be fruitful to think of all the basic operations as being set-theoretical in nature, since what is required in the above method is only the ability to form sets and make deletions.

We next discussed the types of simple operations in general terms. I will give the final result of this discussion, although it didn't come out until later. Originally we had in mind three types of operation: set operations, deletions (erasing line segments), and constructions (adding line segments). The first two can be done entirely within the coding, i.e., it is not necessary to go back to the figure from which the coding was obtained to do these operations. This is not true of the constructions, for one can't tell from the coding for a figure if a line segment that is added between points on two lines will intersect other lines in the figure or not. A machine that can make constructions is more powerful than one that can't; in particular, there are two obvious things it can do that a machine without this ability couldn't: it can recognize whether a polygon is convex or concave (by connecting all its vertices and seeing whether or not all the added lines intersect) and the inside/outside of a polygon (for a convex polygon it is possible to draw lines to each of the sides without intersecting the polygon from a point on a line segment if and only if that segment is inside, and combining this with the first construction takes care of the concave case.) A machine without this ability can't do this since the figures in each

of the following pairs have the same, i.e., equivalent codings:

1)  2)  . We decided to ignore the construction operations, which has the effect of ignoring the figures and seeing what can be done with the coding alone. We next discussed whether we should have generalized or parameterized deletion operations. To take the legs example, a generalized deletion operation would simply delete all legs at once, and repeat this until all legs were removed, while a parameterized operation would remove one leg at a time, the particular leg being removed having to be specified by its endpoints (the parameters). We chose the parameterized version since it is more powerful (it can do things the general operations can't), simpler (everything can be done with one operation), and less arbitrary (just which general operations to allow would be to a certain extent an arbitrary choice). Furthermore, this type of deletion allows for more learning, since less is built into the machine to start with, which is good since it is intuitively the more natural approach, but bad since the learning is complicated. The choice was confirmed when we discovered that the parameterized deletion rules could be easily formulated using the set operations.

The above work was a joint effort of the three of us. Now definite problems had been defined and a definite framework for solving them set up. The rest of this chapter is concerned with my attempt to solve these problems and is my independent contribution to this problem.

The problem now was to see how much could be done within the framework we had agreed on. A formally nice approach would be to write down the basic operations and the ways they could be combined to get more complicated routines and then prove that a device that could do these things could recognize in a reasonable way the predicates that it intuitively ought to be able to. Such an approach proved to be unfruitful, however. First of all, a coding is essentially a set of n-tuples, so that any set-theoretical operation that can be performed on a set of n-tuples can be used on a coding. Specifying particular operations adds very little to an understanding of the present problem. Secondly, the selection of particular operations is primarily a question for the learning theory;

operations will be chosen not for their mathematical elegance, but because they facilitate the learning of geometrical predicates. Moreover, to choose them intelligently requires that one knows what routines are needed to recognize the desired predicates, and this was lacking at the time. Finally, allowing all set-theoretical operations is adequate for determining which predicates can be recognized from the coding, for it is obvious that whatever can be recognized can be recognized using them. This is an interesting problem, and solving the other problems satisfactorily depends on its solution. For these reasons, I allowed myself to use any set-theoretical operations which seemed useful in dealing with codings.

The first thing to notice is that since we have deletion rules, the definition of a coding requires a slight alteration; it is no longer desirable to require that the coding be an initial segment of the alphabet. The reason is that even though the coding of the original figure satisfies this requirement, figures obtained from it by deleting segments don't necessarily satisfy it, since it is possible to delete all the segments containing a certain vertex and still have vertices with labels from later in the alphabet left. Relaxing this requirement doesn't effect Huff's results, except that p cannot be taken to be a permutation, but instead, must just be a 1-1, onto map from one subset of the labels for vertices to another. For notational convenience, I am going to restrict the set of labels for vertices to capitals from A-H (with subscripts), and call this set I.

One can regard a coding as either a set of words of I or as a set of n-tuples of elements of I. The first method is more natural when one is dealing with automata, but the second way is more natural in the present context, since the operations are set operations. Thus, what I will now call a coding is the set of n-tuples obtained in the natural way from the original coding. I will write these n-tuples the same way as they were written in the original coding, e.g., ABC.

Definition: An n-tuple b of elements of I is a <u>line</u> iff $n \geq 2$ and no element of I occurs more than once in b.

Thus, every coding for a figure is a set of lines, but the converse is false. A trivial way a set of lines could fail to be a coding for a figure would be to have a vertex occur on only one line in the set, and

on that line as a central vertex, e.g., (ABC) is not a coding for a figure. I shall call a set of lines with no such vertices a good set of lines. This example shows that a subset of a coding is not necessarily a coding. There are, in fact, good sets which aren't codings, but I will return to this question later. I use small letters b-h as both names and variables for lines, U to denote a coding and V to denote an arbitrary set of lines. Thus, operations that can be performed on V can also be performed on U, and thus definitions that apply to V apply also to U.

There are a couple of obvious things that apply to any coding. Given any line in a coding, it is easy to tell the labels for its endpoints from the labels for its central vertices, since the former are the first and last elements of the n-tuple, while the latter are the remaining elements. It is also easy to determine the number of lines a vertex is on; simply count how many times it occurs in the coding. A formally better way would be to form the set of all lines on which it occurs, and take the cardinality of the set. I now want to restrict my attention to codings for figures that have at most two lines intersecting at each vertex, i.e., in which no vertex in the coding occurs on more than two lines. Thus, each vertex can be classified into one of four categories, depending on how many lines it occurs on and whether or not it is an end or central vertex on these lines. The four categories are single-end (occurs on one line), double-end (occurs as endpoint on two lines), double-central (occurs as central point on two lines), and end-central (occurs as endpoint on one line and central point on another). Notice that the remark in the preceding paragraph amounts to saying, in this terminology, that single central vertices cannot occur in a coding for a figure. These four categories are exactly the categories (see p. 36) a,b,d and c, respectively, that the original scanner sent to the automaton. Thus, presenting the automaton with a coding is in fact equivalent to presenting it with the information that it could get from the scanner, minus the distances and angles, which is what we wanted to do. For convenience, I will henceforth say "A is in V" instead of "A is on a line in V," and use the phrase "remove A from V" to denote the operation of replacing all n-tuples of the form qAr by the n-tuples qr.

The parameterized deletion operation can now be stated precisely. This operation on the coding corresponds to erasing a simple line segment, one with no central vertices, on a figure for which it is a coding.

Definition: AB is a simple line segment in V if there is a $b \in V$ such that $b = qABr$, where q,r are n-tuples of elements of I, $n \geq 0$.

Intuitively, there are three cases: q,r both empty (AB is the whole line), only one empty (AB is the last segment on a longer line), and both non-empty (AB is in the middle of a longer line). Originally, I wrote four operations to cover these three cases (two rules for the second case), but the following rule covers all the cases.

Deletion Operation: To delete simple line segment AB from B, form the set $V'$ by replacing qABr with the two elements qA,Br; form $V''$ by deleting all 1-tuples from $V'$; remove all single central vertices from $V''$.

$V'$ is not necessarily a set of lines. $V''$ is a set of lines, but not necessarily a good set, but the final result is a good set. More importantly, assuming that one started with a coding, the final result is a coding. Indeed, the result is a coding for the figure obtained by erasing the segment AB in any figure which U is a coding for. The converse is not true; it is possible to obtain a coding by deleting a segment from a good set of lines that is not a coding, as I will show when I take up this question later.

This parameterized deletion rule can do whatever any generalized deletion operation could do. A generalized operation deletes all simple segments of a certain type. Obviously, some restriction on the classification of simple segments is necessary to make these operations meaningful basic operations, e.g., one wouldn't want a basic operation that said delete all segments on a hexagon. The natural restriction to place on the classification is that it can depend only on the configurations (see p. 45) at each endpoint. Since it is possible to recognize the four types of vertices from the coding, it is possible to recognize the 16 types of simple segment. Hence, any of the 16 possible generalized deletion operations can be performed by deleting all segments of a certain type one by one.

I now want to state formally the informal method of recognizing triangles that was mentioned above. By saying that there is an effective procedure for recognizing triangles from the coding I mean that there is a way of actually listing the three vertices of each triangle. This is a slight departure from the learning set-up we originally envisioned, which consisted simply of yes and no answers, but there are three reasons for it. First, only slight additions are needed to a procedure that can answer yes or no correctly to the question "Does U contain a triangle?" to get a procedure that can list the vertices of each triangle. Secondly, it is possible to ask a question like "Is there a point which is the vertex of 7 triangles?" that is answered most naturally by listing the triangles. Finally, as is obvious in the case of connectedness, being able to list the information about simple predicates is a big help in being able to give yes and no answers about more complicated predicates. The actual names of the vertices that the device uses are internal to it, but it could identify the vertices by location so that it would be possible to directly check to see if it was actually recognizing the predicate correctly. Otherwise, this could be checked indirectly by asking questions like the one involving seven triangles.

Theorem 1. There is an effective procedure for listing all the triangles in a figure from the coding for the figure.

Proof. Pick a line b from U. Form the set U-{b}, and take all the lines in this set which have a vertex in common with b, getting a set W. Every vertex A which occurs on two lines in W is the vertex of some triangle. The other two vertices are the vertices which the lines A is on have in common with b. All triangles which have a segment of b for a side are found by this procedure, so now it is possible to delete b and repeat the procedure, and thus get all the traingles in U. This process terminates since U contains only finitely many lines. Q.E.D.

This proof is given for the case where at most two lines meet in a point, but it generalizes easily to the general case (where any number of lines can meet at a point). Simply form the set of lines crossing a given line as above, and for each vertex which occurs more than twice in this set each pair of lines which meet at this vertex are the sides of a triangle. The next theorem is important for two reasons: it plays a

crucial role in later work, and it was this predicate that perceptrons failed on. For these reasons I give a detailed proof. First, the required definition.

Definition: A coding is connected if there is a sequence $\langle b_0, b_1 \ldots b_n \rangle$ for every pair of lines c,d such that $b_0=c$, $b_n=d$ and for all i, $1<i<n$, $b_i$ has a vertex in common with $b_{i-1}$ and $b_{i+1}$.

Since the figures we are dealing with contain only straight lines, a figure is connected if and only if the coding for the figure is connected.

Theorem 2. There is an effective procedure for listing the components of a coding U, and hence for recognizing connectedness.

Proof. Pick a line b in U. Define the sets $M_i$, $N_i$ recursively as follows: $M_1=\{b\}$, $N_i=U-M_1$. $M_{i+1}=\{c \in N_i$: c has a vertex in common with a line in $M_i\}$, $N_{i+1}=N_i-M_{i+1}$. These sets can be found in an effective way from the coding. The $N_i$ are a decreasing sequence, i.e., for all i, $N_{i+1} \subseteq N_i$. Since U has only finitely many elements, there is some number m such that $N_{m+1}=N_m$. Let p be the least such number. I claim that $U-N_p$ is a component of U. To establish this I must show that $U-N_p$ is connected and that no line not in $U-N_p$ is connected to a line in $U-N_p$. This requires three simple lemmas.

Lemma 1. For all n, $U-N_n = \bigcup_{i \leq n} M_i$. Proof is by induction on n, but I omit it.

Lemma 2. $\bigcup_{i \leq n} M_i$ is connected. Proof by induction on n.

a. n=1. $\bigcup_{i \leq 1} M_i = \{b\}$, and is connected.

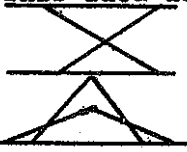
b. Suppose $\bigcup_{i \leq n} M_i$ is connected. Show $\bigcup_{i \leq n+1} M_i$ is connected.

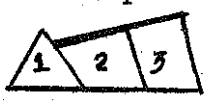
I will show that any two lines in $M_{n+1}$ are connected, which is the hardest case. If $b_0, b_m$ are in $M_{n+1}$, then they each cross at least one line in $M_n$, say $b_1$, $b_{m-1}$, respectively. By assumption there is a sequence $\langle b_1 \ldots n_{m-1} \rangle$ which connects $b_1$ to $b_{m-1}$. Thus, the sequence $\langle b_0, b_1 \ldots b_{m-1}, b_m \rangle$ connects $b_0$ to $b_m$.

Lemma 3. $\bigcup_i M_i = \bigcup_{i \leq p} M_i$. $M_{p+1}=0$ since $N_{p+1}=N_p-M_{p+1}=N_p$. But then $M_{p+2}=0$, since no line crosses an element of the empty set. Thus, for all $k>1$, $M_{p+k}=0$.

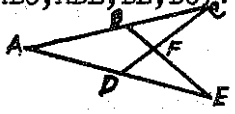
That $U-N_p$ is connected follows from lemmas 1 & 2. Suppose that a line c is connected to a line in $U-N_p$. Then it is connected to b, so there is a sequence $b_0 \ldots b_m$ of lines that cross, where $c=b_m$. Hence, $c \in \bigcup_{i \leq m} M_i$, and thus, by lemmas 1 & 3, $c \in U-N_p$. This process can be repeated on $N_p$, and then again, until all components of U are found. U is connected if and only if there are 0 or 1 components. Q.E.D.

Notice that this procedure makes no use of the fact that only two lines intersect in a point, and hence is good for the general case. Also, given this theorem, it is easy to tell if a line segment AB is on a polygon in U, since AB is on a polygon if and only if the component of U containing AB is non-empty and connected after AB is deleted. Moreover, as will be useful later, it really makes no difference whether or not U is a coding, but the same definition of connectedness and the same procedure will work for an arbitrary set of lines. Incidentally, this theorem shows that our particular approach and the perceptron approach are incomparable, i.e., that neither can do everything that the other can. Our approach can recognize connectedness, while the perceptron approach can't; but the latter can recognize the predicate 'rectangle,' while our approach can't.

I tried to extend the procedure for recognizing triangles to a method for recognizing all types of polygons by adding lines to a given line as in Theorem 2. This makes it necessary to treat polygons with an even number of sides separately from those with an odd number. To get quadrilaterals, for instance, one would take all the lines that cross two lines in the set of lines that cross the original line b. One would get pentagons by taking all lines, except b, that cross a line in this set, and seeing if a vertex occurs twice in it. This idea doesn't work out since for quadrilaterals the following case  would appear to be a quadrilateral. The following pentagon  would not get recognized since four of its sides are added at once. Things get worse as the number of sides increases, so this method proved to be infeasible.

I then thought maybe it would be possible to break the figure down into simple regions, i.e., in 

1, 2 & 3 are simple regions,

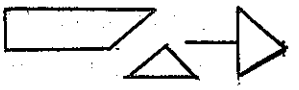
and combine these to get all the polygons. This attempt led to the discovery that it is impossible to recognize simple regions from the coding, since, for example, these two figures have the same coding but different simple regions:   . At this point I discovered a simple example of a good set of lines that isn't a coding {ABC,ADE,BE,DC}. Intuitively, this would be the coding for the figure 

with F omitted. This example also shows how a good set of lines can be converted into a coding by deletion, since {ABC,AE,BE} would be the result of deleting DC, and it is a coding. Characterizing necessary and sufficient conditions for a good set to be a coding is a very natural problem that turns out to be quite difficult.

It is obvious that a coding contains enough information to enable one to list all the polygons in a figure for which it is a coding. Since it is impossible to recognize concave/convex, inside/outside, and simple regions it seems that this is about all one can hope for, and so it seems to be a good test of the adequacy of any learning device to see if it could learn to recognize all the polygons. Before trying to build a device that could learn to do this, however, it would be nice to know how to do it for oneself, and thus know what sort of things the device is going to have to be able to learn.

After several fruitless attempts, I finally came up with a method for breaking a figure into simpler figures. Take the figure  and look at vertex A. It occurred to me that it would be possible to replace this figure by three figures, in each of which a different simple line segment containing A had been deleted, i.e., by  . Now each polygon in the original figure is in one of the new figures, and each can be gotten easily by deleting legs. In a more complicated case, a polygon might be in more than one of the resulting figures, but this duplication presents no problem. Notice that this breaking a figure into several simpler figures allows for the possibility of parallel computation, which is desirable. Moreover, it seems that this method is fairly close to the method people would use in solving this

problem, and certainly it is much more intuitively desirable than the first method I tried.

I finally realized that the restriction to two lines meeting at a point was unnecessary, since this procedure, like the triangle and connectedness procedures, does not depend on the restriction. For example, it is easy to break a figure like  into  Thus, dealing with the general case is really not much more difficult, and in a way it is easier, since it led me to concentrate on general features rather than on ad hoc devices for the special case where at most two lines meet at a point. I henceforth dealt entirely with the general case, and this led to a surprisingly simple procedure for recognizing polygons.

Before stating the theorem, it is necessary to introduce some notation and definitions.

Definition: AB is a segment in U if there is a line b = aArBt, where q,r and t may be either empty or non-empty.

Definition: P is a polygon in U if P is connected and P = {segments: each vertex in P occurs in P twice and only twice}.

If U is a coding and P is a polygon in U, then the segments labeled by elements of P form the perimeter of a polygon in every figure for which U is a coding.

Definition: A broken line is a set of simple segments in which two vertices occur once and the rest occur twice.

Definition: A broken line is a leg if it is such that no segment on it is on the perimeter of any polygon or part of a broken line between two polygons.

The legs are the stray lines attached to one polygon. Thus, the terminology is appropriate. It is now possible to state formally the 'delete legs operation' I mentioned informally.

Delete Legs Operation: Delete all simple segments in the coding that have an endpoint that is on only one line. Repeat until all such segments are removed.

This operation can be effectively performed given only the coding for a figure, and it deletes all and only legs in the figure.

Definition: For each vertex A in U, A* is the set of all simple segments containing A.

Definition: U-A* is the set of lines obtained from U by performing the deletion operation on each element of A*, except that single central vertices are not removed.

Definition: A is a breaking point of U if U-A* is disconnected.

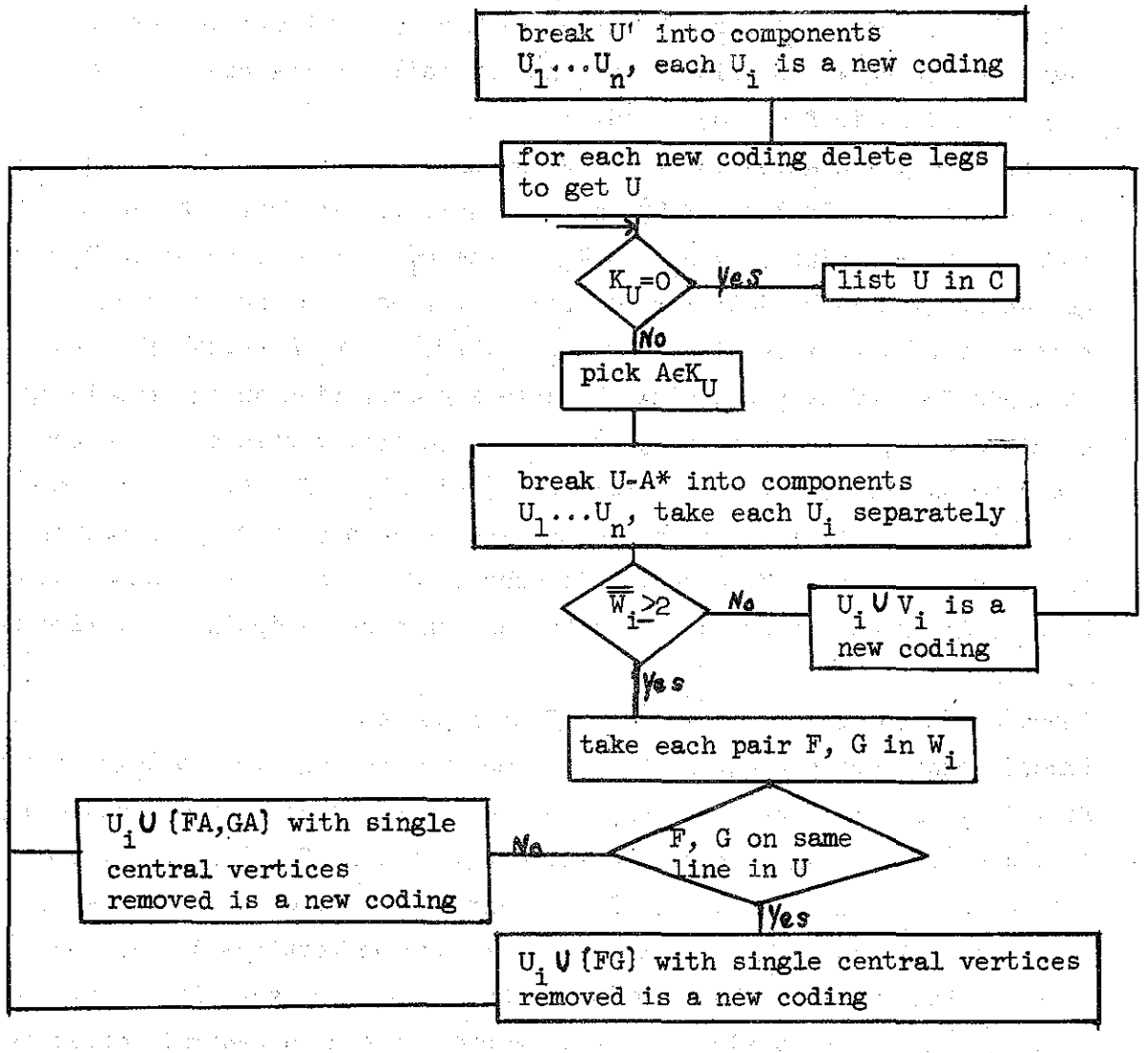
Notice that U-A* is not necessarily a coding, or even a good set of lines. However, it is simpler to leave the extra vertices in, and as remarked after Theorem 2, connectedness applies to any set of lines.

Definition: If A is a breaking point of U and $U_1 \ldots U_n$ are the components of U-A*, then each set $U_i^A = U_i \cup \{FA \in A^*: F \in U_i\}$ is called a component of A.

Thus, all the components of A are codings since each $U_i^A$ can be obtained from U by deleting all segments of A* that don't have an endpoint in $U_i$, and then deleting the components of the resulting figure that don't contain A. For notational convenience, let $\bar{\bar{A}}$ be the cordinality of A, and given any figure U let $K_U = \{A \text{ in } U: \bar{\bar{A}}^* \geq 3\}$. Also, for each component $U_i$ of U-A*, let $W_i = \{F \in U_i: FA \in A^*\}$ and $V_i = \{FA \in A^*: F \in U_i\}$.

Theorem 3. There is an effective procedure for listing all the polygons in a figure U'.

Proof. I claim the following procedure will work. Continue this procedure until no new codings are formed, and at this point C is a list of all the polygons in U'.




In talking about this procedure it is convenient to think of things happening in stages. The arrow in the flow chart makes the beginning of a new stage. Stage 0 is before the arrow is reached the first time, stage 1 between the first and second times, etc. At each stage all the new codings from the previous stage are processed simultaneously. The proof that the procedure works depends on proving three lemmas about what happens at each stage.

Lemma 1. If one begins at the arrow with a coding U and vertex A, then everything in the flow chart that is called a new coding is a coding, and is in fact connected.

Proof. Every X claimed to be a new coding is of the form $U_i$ union some subset Y of $V_i$. If $Y=V_i$, $X=U_i^A$ and is a coding. If $Y \subset V_i$, X can be obtained by removing elements in $W_i-Y$ from $U_i^A$. This is so since $U_i$ was gotten originally by deleting A* from U, but not erasing the single central vertices as required. Thus, $U_i \cup \{FG\}$ or $U_i \cup \{FA,GA\}$ is what is obtained from $U_i^A$ by performing the deletion operation on the remaining elements of $V_i$ without erasing the single central vertices. When these are removed the result is the same as $U_i^A$ with $V_i - \{FA,GA\}$ deleted, and hence is a coding since it can be obtained from a coding by using the deletion operation. Moreover, X is connected since it is a component $U_i$ of U-A* union one or two segments that have one endpoint on a line in $U_i$.

Lemma 2. Each element X listed in C is a polygon.

Proof. That each element of C is a connected coding with no legs follows from Lemma 1, the fact that components of U' are codings and because all legs are deleted right before an element is put in C. Thus, no vertex occurs on only one line in X. But no vertex occurs on 3 or more simple segments in X since $K_X=0$ means that there are no vertices A such that $\overline{\overline{A^*}} \geq 3$. Thus, there are no central vertices in X, because if there were it would have to be a single central vertex because a central vertex is on 2 simple segments in every line on which it is a central vertex, which contradicts the fact that X is a coding. Thus, each vertex occurs on exactly 2 segments. Thus, X is a set of segments and each vertex occurs exactly twice in X, hence X is a polygon.

Lemma 3. For each coding U at the beginning of stage n, P is a polygon in U if and only if P is put in C at stage n or there is a new coding X formed at stage n such that P is a polygon in X.

Proof. $\Rightarrow$ Suppose P is a polygon in U. U has no legs and hence it is either a polygon and goes in C, or it has a vertex A such that $\overline{\overline{A^*}} \geq 3$. There are two possibilities.

i) A is a breaking point of U. In this case P is a polygon in $U_i^A$ for some i. Parts of P cannot be in two different components of A, since if they were these components would not be disconnected when A* is deleted, and hence would actually be only one component. There are two possibilities: no segment in A* is on P, in which case P is in each of the new codings formed from $U_i$, or two segments in A* are on P, in which case P is on the new coding formed when this pair of segments is chosen.
ii) A is not a breaking point of U. Then there is only one component, which implies that P must be in one of the new codings by the above argument.

If P is put in C, then P=U and hence P is a polygon in U. If P is a polygon in some $U_i^A$, as it is if it is in a new coding X, then it is a polygon in U.

By repeated application of Lemma 3, it follows that the set of polygons in some new coding formed at stage n union the set of polygons put in C on or before stage n equals the set of polygons in U'. The only thing left to show is that at some stage m no polygons are in new codings formed at stage m, i.e., at some stage m no new figures are formed. This follows since for each new figure $U_i$ formed from U at stage n $\overline{\overline{K_{U_i}}} < \overline{\overline{K_U}}$, since $A \in K_U$ but $A \notin K_{U_i}$. So if $\overline{\overline{K_{U'}}} = \mathbb{N}$, the process will terminate on or before stage n. Indeed, the routine is set up so that this will happen. Thus, at the first stage no new codings are formed, all the polygons in U' are in C. Q.E.D.

There are several comments I would like to make about this procedure. The memory requirement is much greater than it was for the connectedness procedure. It is now necessary to store the original coding, the new codings and the list of polygons. The polygons could be put in the output as they are formed, but because of duplication it is still necessary that the device knows what has already been printed. The memory requirement for the new codings could be reduced by processing one new coding at a time, but this would greatly increase the computation time. It would be very desirable to eliminate the duplication that occurs in this procedure to save time and cut down the memory requirement. It seems to me that the resulting procedure would be very close to what people actually do, but I was unable to come up with a good way of achieving this economy.

Also, the idea of an irrevocable deletion has been lost. Now the coding is divided into parts and different things are done to the parts and then they are recombined. This is a more powerful procedure, and it is intuitively plausible since it is possible to ignore part of a figure and analyze the rest of it, and then analyze a different part if this doesn't lead to satisfactory results. An interesting problem concerning this procedure was suggested by Professor Jaakko Hintikka: In the procedure given, the vertex A around which the coding is decomposed is chosen at random. It seems reasonable that certain strategies for choosing A would lead to a more efficient procedure than random selection. In particular, it seems as if it might be wise to choose A so that it is on the greatest number of simple segments. I have no concrete results on this problem, however.

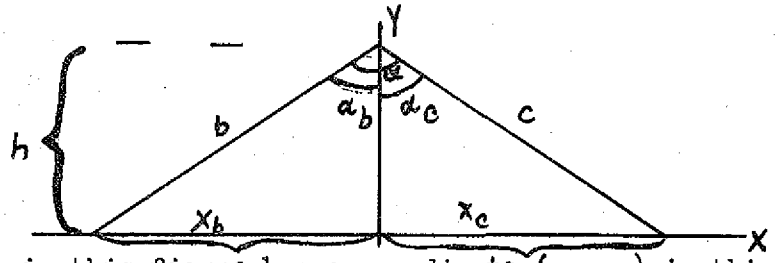
I next tried to solve the representation problem, i.e., find necessary and sufficient conditions for a good set of lines to be a coding. The attempt has led to many interesting results, but as of now it has not produced the desired theorem. I will now cover some of the work I did for three reasons: many of the results are of independent interest, listing some of them will serve to indicate the complexity of the problem and perhaps be of use to others who might be interested in this problem.

The first thing to notice is that the size of the figure makes no difference since the figure can be expanded or shrunken without changing the coding. Also, there is no problem in constructing a figure that contains no polygons, for one can just start anywhere and draw the lines and run into no problem of lines that aren't supposed to intersect intersecting. Some of the lines may get pretty small, but this is of no theoretical significance. Thus, it would be nice to have a list of all the polygons for any good set of lines to aid in determining if it is in fact a coding. The procedure of Theorem 3 will produce such a list, however, since on closer examination of Theorem 3 it is clear that it is not essential that U' or all the new codings are in fact codings. It is also true that legs can always be added to any figure, even 'legs' like  , for the square can be shrunken arbitrarily small and so this case is really no different from the case of an ordinary leg.

It seems to me that the best approach to this problem is to take the coding apart at all its breaking points and try to draw a figure for each component separately, and then try to fit them together to get the desired figure. There are two difficulties that could arise: the breaking points could be inside a polygon in both components, e.g., if in the coding for a figure containing   there were a line between A and B not crossing either square, or if A and B were the same vertex, then the alleged coding would not actually be a coding for any figure. Also, two concave polygons can't always be joined, e.g.,  can't be joined at the marked points. Trying to see which angles could be fitted together led to the following result, which shows there is no problem for angles less than $180^{\circ}$.

Theorem 4. If U is a coding for a figure F in which angle $\alpha$ is less than a straight angle, then U is a coding for a figure G in which $\alpha < \epsilon^{\circ}$ and a figure G' in which $\alpha > 180^{\circ} - \epsilon^{\circ}$, for any $\epsilon > 0$.

Proof. Let b,c be the sides of $\alpha$ and A be their point of intersection. Draw a line through the endpoints of these sides and take this to be the x-axis and the perpendicular to this line from A to be the y-axis.




Every vertex in this figure has a coordinate $(x_i, y_i)$ in this coordinate system. For any positive number r, if we set up a similar coodinate system and give each vertex coordinates $(x_i, ry_i)$, and connect them the same as in F, the result will be a figure F' with the same coding. The only thing to check is that points that lie in a straight line in the first coordinate system have their images in the second coordinate system lying on a straight line. I omit the verification of this point. If r is small enough, $\alpha$ will be very close to a straight angle. Specifically, take the shorter of the two lines, say b, and choose r so that $rh < x_b \sin(\epsilon/2)$. Then the resulting figure is an appropriate G'. By

taking r very large one can get an appropriate G. Q.E.D.
As an immediate corollary, we get that a similar result holds for angles greater than 180°.

The intuitive meaning of this theorem is that all angles in a figure less than (greater than) a straight angle are indistinguishable in a coding for the figure. Thus, for each vertex of a polygon the most one can determine is whether or not the polygon is concave or convex at that vertex. The case where all polygons are convex is simpler than the general case, so it seemed to me that it would be good to consider a figure in which as many vertices as possible were convex. If a vertex is inside a polygon not much can be done in this regard, for the vertex will then be on at least two polygons (assuming the coding is connected and has no legs or breaking points) and will be concave on one and convex on the other. Hence the following definition was framed with the idea of applying it to polygons which have no part of their perimeter enclosed by other polygons. I call such polygons outside polygons.

> Definition: A is <u>concave</u> in U if for every figure of which U is the coding every polygon P of which A is a vertex is concave at A.

Thus, if A is a central point in one of the sides of P, it is still regarded as being concave, since the following figures cannot be joined at the marked point:  . It may seem that there would be no other concave angles, but this is not so. If U is a coding for this figure 

then A is a concave angle. For angles on an outside polygon, the idea was to choose a figure in which that angle is convex, if possible.

Other examples like the above one convinced me that an angle on an outside polygon is concave if and only if there is a straight line passing through it. If there is no such line, it seems one could bend the figure out without altering the coding. A proof of this is currently lacking, however. For vertices inside a polygon, I say that vertex is acute if there are two adjacent simple line segments coming from that vertex in some figure for which U is the coding which are more than 180° apart.

The idea is that it would be possible to attach a component with that vertex concave to such a point, but not otherwise. I think it could be shown in a way similar to the method of establishing the first result in this paragraph that A is acute if and only if there is no polygon P such that there is a line from A to each of the vertices of P.

If we now call an angle acute also that is convex but on an outside polygon, the following condition is necessary and sufficient for the result of combining all the components to be a coding: if A is a breaking point in U, $U_1 \ldots U_n$ the components of A, then U is a coding if and only if each of the $U_i$ is a coding and i) at least n-1 of the $U_i$ are codings in which A is on the outside of $U_i$, and ii) at least n-1 of the $U_i$ are codings in which A is acute. Each $U_i$ is here regarded as having no legs, and hence each figure for which $U_i$ is a coding will be enclosed by some polygon, and A is on the outside of $U_i$ if there is a figure in which A is on the outside polygon. The above unestablished results would give a way of determining these conditions from the coding, and hence this would be a good way of decomposing a coding with breaking points into simpler codings.

The problem now is to determine which of the components are codings. These have no legs and no breaking points, and thus they are enclosed by a polygon. In the simple example of a good set that isn't a coding, it is impossible to draw a figure for it without having to have some point both inside and outside some polygon. Other examples convinced me this was a general phenomena, and so I thought it would be wise to see what restrictions the coding places on inside/outside relationships. Given this, there is a natural starting point for a coding that contains no breaking points or legs, since the polygon which encloses a figure for which it is a coding will have to be such that it is possible to have everything inside it. The simple example of a good set which isn't a coding has no such polygon.

This led me to define P* analogously to A* and component of P analogously to component of A.

Definition: $\underline{P^*}$ = {simple segments with at least one endpoint on P}. Notice P itself is included in P*.

Definition: U-P* is the coding obtained from U by deleting all the elements of P* except that single central vertices aren't removed.

Definition: $U_i^P$ is a component of P if $U_i^P = U_i$ (a component of U-P*) $\cup$ {XY: XY $\in$ P*, X in $U_i$}, or $U_i^P$ = {XA: A is not on P or any of the components of U-P*}.

It is now possible to state the following restrictions on outside/inside which can be determined directly from the coding.

Definition: An assignment of inside/outside to a polygon P in U is consistent if and only if for A,C not on P, B on P and Q any other polygon in U the following conditions are satisfied:

1. If B is not a vertex of P and ABC $\in$ U, then A is inside P if and only if C is outside P.
2. If B is a vertex of P, ABC $\in$ U, then P is acute at B implies either A or C is outside P and P is concave at B implies either A or C is inside P.
3. If AB is an extension of a side of P, then P is acute at B if and only if A is outside P and B is concave if and only if A is inside P.
4. Points on the same component of P are on the same side of P.
5. If P is inside Q, then Q is outside P.
6. If P and Q don't intersect, then either P is outside Q or Q is outside P.
7. If two components of P have three points in common, then they are on different sides of P.
8. If $A_1$, $A_2$, $A_3$ and $A_4$ occur in order on P, and $A_1$ and $A_3 \in U_1$ and $A_2$ and $A_4 \in U_2$, then $U_1$ and $U_2$ are on opposite sides of P.
9. P is concave at most n-3 places if P is an n-gon.

These are necessary conditions to be able to draw a figure for U without violating inside/outside. I don't believe they are sufficient, and in any case I can't prove that they are, which is the hard part. I do not see any obvious way of proving that a set of conditions is

sufficient, and this is the main problem. The thing to do would be to show that it is possible to draw a figure if the conditions are met, but there is no clear way of going about this. It probably wouldn't be too difficult to make a list of necessary and sufficient conditions once one had some idea how to do this.

A different approach to the representation problem would be to start from one of the well-known axiom systems for Euclidean geometry and see what conditions a good set of lines must fulfill in order to satisfy these axioms. However, the axioms apply to particular figures, while the problem under consideration is to see if there is any possible figure of which a given coding is a coding. Thus, if one attempted to draw a figure that had a given coding, the axiom system could tell if and when this particular attempt went wrong. The only way this would be of help in solving the representation problem would be if one had a way of listing a finite number of 'possible figures' for each coding. Given such a list, it would be possible to determine if there was a figure of which a given coding was a coding by simply running through all the possibilities. It seems plausible that for any given coding there is a method for listing a finite number of figures. Indeed, results like Theorem 4 should be useful in the attempt to find such a method. However, there is no natural way of constructing such a list, and it seemed to me that the approach I tried was more likely to be successful.

The basic idea behind my proposal for solving the problem of abstract ideas, for the limited context I have considered, is quite simple. Essentially, it consists of identifying the abstract idea of triangle with the procedure given in Theorem 1 for recognizing triangles and the abstract idea of polygon with the procedure for recognizing polygons given in Theorem 3. The problem with this is that it is obviously possible to give procedures that are variants of the procedures of those theorems which have the same end result. It would be completely arbitrary to single out any of the particular variants as the abstract idea.

In the following discussion, I will restrict myself to the abstract idea of triangle, but the same things are true for the abstract idea of polygon. There are two ways out of the above difficulty. The first way

is to identify the abstract idea with the class of all procedures that recognize triangles. This has the result that there is only one abstract idea of triangle, which is the view of Locke, Berkeley and Hume and many other philosophers. Indeed, it is the belief that there is only one procedure which justifies the terminology 'the abstract idea.'

It seems to me very unsatisfactory to identify the abstract idea with a class of any type. Moreover, upon reflection it seems that it is mistaken to believe that there is only one abstract idea of triangle, and hence I think the terminology 'the abstract idea' is misleading. As far as I can see, there is no compelling reason for holding that different people who can recognize the same property have the same abstract idea. The property of being a triangle is an objective property of figures, while the abstract idea is a subjective mental disposition of a device that can recognize triangles. Thus, the natural thing to do is to speak of a particular device's abstract idea, and hence to relativize the notion of abstract idea to particular devices. The correct terminology would be 'the abstract idea of device A,' not simply 'the abstract idea.' To justify introducing the phrase 'the abstract idea,' would require conclusive evidence that all devices have the same abstract idea. This is seemingly going to be impossible to obtain, as there certainly are different ways to recognize the same property and nothing to indicate that there is a particular one that all devices happen to use.

This use of the term 'abstract idea' may seem a little strange. However, since the traditional use presupposes that it is part of the meaning of 'abstract idea' that for every property there is only one abstract idea and yet wants to maintain that an abstract idea is mental, it seems that this use is mistaken. It seems to me that the crucial notion that must be saved is that an abstract idea is mental (a property of a device), and hence I have chosen to use 'abstract idea' relative to different devices. To be technically correct, one should really speak of 'the abstract idea of device A at time t,' but this is a detail I shall ignore.

Stated precisely, my proposal is to identify the abstract idea of a device A with the procedure that A uses to recognize the appropriate property. For recognizing triangles, such a procedure would not have to

take into consideration the specific properties of any particular triangle. Certainly the procedure of Theorem 1 doesn't, and for any other procedure that is similar to it in this regard the problem of the abstract idea having specific properties, which bothered Berkeley and Hume, does not even arise.

I will now show that a device B that uses the procedure of Theorems 1 and 3 acts in an intuitively appealing way. It seems to me that people act in roughly the same way, but this is pure conjecture. First of all, Hume had great difficulty with the abstract idea of triangle, but the idea of polygon is even more abstract, and there is no way to account for such an idea in his theory. The procedure of Theorem 3 solves this problem, since it is B's abstract idea of polygon. Moreover, this procedure has the very desirable property that it recognizes that a figure is a polygon before it recognizes how many sides it has. The desirability of this isn't obvious when one considers triangles, but it is if one considers Hume's example of a chiliagon, which is a thousand-sided polygon. Hume points out that people could recognize a chiliagon, but not by comparing it directly to some picture in their heads. Rather, they would first recognize that the figure was a polygon, and then count the sides to see that it had a thousand. If beside the procedure of Theorem 3, B had the ability to count, then it would proceed in the same way. This is another reason I believe that the procedure of Theorem 3 is close to the way people actually operate.

Another gap in Hume's theory is that he has no way of accounting for the fact that 'triangle' is a special case of the more general predicate 'polygon.' This is also solved in the case of B. Ordinarily, B uses the procedure of Theorem 1 to recognize triangles, since it is much quicker than using the procedure of Theorem 3 to recognize that the figure is a polygon and then counting to see if it has three sides. However, both procedures will in fact recognize triangles. The latter procedure shows that 'triangle' is a special case of 'polygon,' and since the former is equivalent in that it recognizes the same property, it too is a special case.

Thus, it seems to me that these procedures completely solve the problems that worried Berkeley and Hume as far as the limited context of the codings is concerned. Insofar as the codings resemble the codings people actually use, it solves the actual problem that Berkeley and Hume addressed themselves to.

## CHAPTER 3

## LEARNING THEORY

The purpose of our work in learning theory was to modify and extend the results Suppes obtained in Stimulus-Response Theory of Finite Automata (SRTFA).[1] I will give the main results of that paper and sketch the set-up used in its derivation, emphasizing the points of particular relevance to the present work. I will indicate the alterations we thought desirable, give our reasons for thinking this, and then give an account of the work we did.

As the title suggests, Suppes' main concern is to show the connection between stimulus-response (S-R) theories and finite state automata (fsa). In S-R theory, an organism learns in a series of trials. Following Suppes' formulation, what happens at each trial can be described intuitively in the following way: the organism is in state of conditioning C at the beginning of the trial, is presented with stimulus T, samples stimuli s, makes response r, receives reinforcement e, and goes to the state of conditioning C'. As in SRTFA, I use S to denote the set of possible stimuli, R to denote the set of responses and E to denote the set of reinforcements. An individual element of S is denoted by $\sigma$. T and s are in general subsets of S, where $s \subseteq T$. I will not go into the details of the general S-R model, for later in this chapter I give a modified version of this S-R model, and all the details are spelled out there.

---

[1] Patrick Suppes, "Stimulus-Response Theory of Finite Automata," Journal of Mathematical Psychology, 6, 3, October, 1969.

The intuitively correct way to make the connection between an S-R model with a finite number of stimuli and responses and an fsa is to think of the set of stimuli S as being the input alphabet A of the fsa and the set of responses R as being its set of states Q. Technically, this is not quite correct, for the axioms of S-R theory require a special stimulus $\sigma_0$ to put the fsa in its initial state $r_0$, and $\sigma_0$ is thus not in A. The rest of the stimuli are. On each trial, the S-R model gives one response, which is equivalent to the corresponding fsa making one transition. Since the transition an fsa makes is a function both of the state it's in and the letter of A it is looking at, the presented set T on a trial can't consist simply of elements of S, for elements of S correspond to letters of A. Rather, in order to account for the fact that a transition depends on the state of the fsa, T must consist of pairs $(r,\sigma)$, since elements of R correspond to elements of Q. Things are still very messy if T has more than one element, so T is restricted to having one element. If T has one element, the sampling axioms require that $s = T$, so one can say simply that a pair $(r,\sigma)$ is the stimulus on trial n without worrying whether it is s or T. The intuitive meaning of the pair $(r,\sigma)$ is that r is the organism's previous response and $\sigma$ is the present stimulus.

There are two other important features of the set up of SRTFA. The set of reinforcements must contain a reinforcement $e_r$ for each element r of R. I will discuss this feature later. Secondly, the S-R model in SRTFA is an all-or-none conditioning model, i.e., there are only two possible states of conditioning for each pair $(r,\sigma)$; either $(r,\sigma)$ is unconditioned, in which case there is a positive probability of giving each response, or it is conditioned to some response $r'$, in which case $r'$ is given with probability 1. If $(r,\sigma)$ occurs and is unconditioned and $e_{r'}$ occurs, there is a positive probability c of $(r,\sigma)$ becoming conditioned to $r'$, and if $(r,\sigma)$ is already conditioned, it remains conditioned. Notice that in this set up no states are ever conditioned incorrectly, since $e_{r'}$ always occurs after $(r,\sigma)$ if $r'$ is the correct response. Once all the pairs (except those containing $s_o$) are conditioned, the organism will behave exactly like an fsa. Indeed, one can use the

conditioning table to construct the transition matrix of the fsa. On this intuitive basis, it is possible to formally define what it means for an S-R model to become an fsa. Suppes does this, and using the ordinary notion of isomorphism between automata, he proves that for any connected fsa there is an S-R model with all its states initially unconditioned that asymptotically becomes isomorphic to it. The key to the proof is to show that for some number n there is a positive probability that each unconditioned state will occur in each sequence of n trials. Since, on each occurrence, there is a positive probability that it will be conditioned, it will eventually get conditioned. The details are similar to those that are given later.

It is particularly interesting in view of the present work that the result remains essentially unchanged if a linear learning model is used, i.e., if the probability of responding $r'$ when presented with $(r,\sigma)$ is $p_{r'}$ and $e_{r'}$ occurs, then the probability of responding $r'$ the next time $(r,\sigma)$ occurs is $p_{r'}(1-\Theta)+\Theta$, $0<\Theta<1$, and for $r''\neq r'$, the probability is $(1-\Theta)p_{r''}$. Thus, the probability of giving the correct response increases each time $(r,\sigma)$ occurs. In fact, it approaches 1 as the number of times $(r,\sigma)$ occurs approaches infinity, and hence a model of this type, though messier, will also at asymptote become the correct fsa.

The main reason we felt that modification was desirable is that the learning that takes place in this set up is of a rather simple nature. The organism learns each appropriate response independently, and hence doesn't have to learn the task as a whole. Such a model is not adequate to account for most human learning, since in the typical experimental case with a human subject, the subject discovers a method of doing the task on his own (which may remain unknown to the experimenter) and it is not necessary that the experimenter should provide a particular method of doing this, which in effect is what he is doing if he reinforces the responses of a particular fsa. In particular, it is not adequate for the learning of geometrical predicates in the way we had in mind. As mentioned in Chapter 2, we were thinking of a device learning a predicate on a series of trials where on each trial a figure

is presented, a yes or no answer elicited, and reinforcement given according to whether or not the answer is correct, which is determined by whether or not the drawing has the predicate in question. Thus, there are only two different reinforcements, not as many different reinforcements as responses. Clearly, it will take an fsa with many states to recognize interesting geometric predicates, and hence just as many responses must be possible in the S-R model. It is easy to regard a model with many responses as giving only yes and no answers, since all that is required is to partition the set of responses into two sets, $R_y$ (for yes), and $R_n$ (for no); just as in automata theory, the set of states is partitioned into the set of final states and its complement. The problem is that only one response is made, while the fsa obviously requires several transitions to get an answer. Thus, we have to regard the organism as making internal responses. There is in general nothing wrong with this, since some such activity is obviously going on, but it is a real problem in the framework of SRTFA, which requires that every response be reinforced, since it is impossible to reinforce an internal response. Finally, there is one other reason it is undesirable to have to single out a particular fsa beforehand. What one is really interested in is the S-R model eventually learning to recognize the predicate in question, and one is indifferent as to how this is done as long as the method is reasonably efficient. Since there are many fsa's that can do any given task, more than one fsa is in general acceptable.

To get an adequate model for the type of learning we wanted requires changes in the set up of SRTFA. First, the notion of trial has to be altered to allow more than one stimulus to be sampled and more than one response given on each trial. Intuitively, a trial now consists of the fsa processing a whole tape, rather than making a single response. In the case of a rat running a maze, for example, a trial now consists of the rat going through the entire maze, instead of making a choice at a certain branching point. The way to accomplish this formally is to introduce the notion of subtrial, and to consider a trial to consist of a series of subtrials. A subtrial corresponds to the trial of SRTFA as far as sampling the stimuli and responding are concerned. This will be evident from the axioms, for the sampling and response axioms are exactly

the same as those of SRTFA except that $s_n$, $T_n$, and $r_n$ (the sampled stimuli, presented set, and response, respectively, on trial n) are replaced by the corresponding notions $s_{n,m}$, $T_{n,m}$, and $r_{n,m}$ (the sampled stimuli, presented set, and response on trial n, subtrial m). In this set-up, it is convenient to think of the organism as having given initial response $r_o$ at the beginning of each trial, and thus ignore $s_o$.

The organism's final response on each trial is regarded as its answer. The intuitive idea is that the final response corresponds to the answer that the subject gives, while the previous responses are internal. There are only two reinforcements, correct and incorrect, and whether the last responseis in $R_y$ or $R_n$ is the only factor which determines which reinforcement is given. The previous responses are necessary for the organism to know what the final response should be, but do not effect the reinforcement. Formally, let $m_n$ be the last subtrial on trial n, and $e_1$ and $e_2$, respectively, be the positive and negative reinforcements. Then $r_{n,m_n}$, the response on trial n, subtrial $m_n$, is the final response on trial n. Reinforcement depends only on $r_{n,m_n}$; $e_1$ occurs if and only if response $r_{n,m_n}$ is correct, i.e., if a yes answer is correct, then $r_{n,m_n} \in R_y$, and if a no answer is correct, then $r_{n,m_n} \in R_n$.

This set-up has the property that all fsa's that can do the given task are equally acceptable. It does not require that a particular fsa be singled out as does SRTFA. What we do require is that the alphabet, set of states, initial state and set of final states be given, and that there is in fact an automaton with these four components that can do the task that is to be learned. This is accomplished formally by introducing the concept of a signature. The only thing we don't require is a particular transition table. In SRTFA, the things we require are needed to choose an appropriate S-R model, while the transition table is needed to choose the appropriate reinforcement schedule. We have completely changed the method of reinforcement so that we don't need the transition table, but we need the other things. We still think of learning as being, in a sense, the construction of a correct transition table, but this construction must be accomplished with less information. Moreover, we don't have to worry whether or not the asymptotic fsa is connected.

Indeed, we don't even require that all states be conditioned at asymptote. If at asymptote the S-R model is an fsa with inaccessible or unconditioned states but that can do the task, we are perfectly happy.

There is a basic problem with this set up; namely, how many states should be available at the beginning of the learning sequence? There is no problem with the alphabet, since it consists of the stimuli, and once the number of states is determined, it is fairly easy to choose a particular set of states, initial state and set of final states. However, there are no general results in automata theory that are helpful in deciding how many states are needed to do a particular task, and we made no progress in this direction. Even if such results were available, it is not clear how to use them, since one wants the organism to decide on the set of states itself (particularly since most of them are internal), but it is not plausible that the organism would have these results available to help it. Thus, the best approach might be to have the automata start out with a small number of states and add new ones if these don't prove sufficient. The course we took was to sidestep this problem and simply assume that enough states are present. It is clear that having unnecessary states available will greatly reduce the rate of learning, but they don't effect the asymptotic results we were concerned with.

Our first step in approaching this problem was to concentrate our attention on the simplest possible non-trivial automata, since we thought (correctly, as it turns out) that all the conceptual problems would show up even in this case. These automata have two-letter alphabets ($\sigma_1$ and $\sigma_2$), two states ($r_1$ and $r_2$), and one final (acceptance) state ($r_1$). We further required that all input tapes have length two. There are thus four different input tapes, and 16 ways to partition these into acceptable and unacceptable inputs. There are some interesting features in this set up. If one takes $r_1$ as the starting state, as we did originally, and takes $\{\sigma_1\sigma_1, \sigma_2\sigma_2\}$ to be the set of acceptable tapes, then there are two automata that will accept it, i.e., whose final response will be $r_1$ when presented with $\sigma_1\sigma_1$ or $\sigma_2\sigma_2$, and $r_2$, otherwise. These are

| | $r_1$ | $r_2$ |
|---|---|---|
| $r_1\sigma_1$ | 1 | 0 |
| $r_1\sigma_2$ | 0 | 1 |
| $r_2\sigma_1$ | 0 | 1 |
| $r_2\sigma_2$ | 1 | 0 |

and

| | $r_1$ | $r_2$ |
|---|---|---|
| $r_1\sigma_1$ | 0 | 1 |
| $r_1\sigma_2$ | 1 | 0 |
| $r_2\sigma_1$ | 1 | 0 |
| $r_2\sigma_2$ | 0 | 1 |

Notice that each connection is different, but that the final result is the same. Secondly, it is not satisfactory to simply take $r_1$ as the starting state, since there are acceptance sets, e.g., $\{\sigma_1\sigma_2, \sigma_2\sigma_1, \sigma_2\sigma_2\}$, which are accepted by no automaton (of the type we are dealing with) with initial state $r_1$, but are accepted by one with initial state $r_2$. It would be possible to add another state, but this is a complication it is best to avoid. We decided that the best way to meet this problem would be to first try to find an automaton with initial state $r_1$ that would work, and if this fails, look for one with initial state $r_2$. This method requires the organism to only be trying to construct one transition table at a time, which seems desirable. Finally, there are two sets, $\{\sigma_1\sigma_1, \sigma_1\sigma_2\}$ and $\{\sigma_2\sigma_1, \sigma_2\sigma_2\}$, which are not accepted by any two-state automaton. Intuitively, these sets are very easy to recognize, since the second element on each tape is irrelevant. The obvious thing to do to solve this difficulty would be to try to get a method of recognizing irrelevant information, and have the organism apply this to the stimuli first before trying to construct a transition table. We wanted to concentrate on automaton learning, however, so we did not use these two sets as acceptable sets.

It is easy to see the difference between the learning procedure of SRTFA and the one we want in terms of this simple case. In both cases, the object is to construct a transition table, but in SRTFA particular transitions are reinforced, while in our case the organism is only told which tapes are acceptable. This does not necessarily give information about any particular transitions even in the two-state case, as my first example shows, and things, of course, get worse as the number of states increases. The learning procedure will have to be such that either

of the two transition tables in my example could result from reinforcing $r_1$ responses to $\sigma_1\sigma_1$, and $\sigma_2\sigma_2$, and it is not clear what natural learning procedure could have this result.

One method of doing this would be to simply list all the possible transition tables, try each (or some) of them on each tape, and discard the ones that don't work. If more than one were left after this process were completed, one could be chosen arbitrarily. This method has two very nice features: it learns quickly if there is a transition table that will work, and it has a method of determining when there isn't one that will work, which is very good for adding new states or trying a different starting state. The problem is that listing all the possible transition tables is a very sophisticated procedure and contrary to the intuitive notions of learning. In particular, there is no direct way to formulate such a procedure in S-R theory. We came up with variations of this procedure that don't seem as counter intuitive as a list of all the possibilities, but finally decided that any type of enumeration procedure was undesirable. Another method, which is much closer to our final approach, is the following, which we called the brute force method. Each state is either conditioned or unconditioned, as in SRTFA. If an unconditioned state is entered in processing an input tape, there is a positive probability of responding $r_1$ and also of responding $r_2$. If the response to the tape is correct, the state becomes conditioned to the response it actually gave. There is a problem with tapes $\sigma_1\sigma_1$ and $\sigma_2\sigma_2$, since, if these are the input tapes, the same state may be entered twice. If the organism acts as a probabilistic fsa, it could respond $r_1$ one time, and $r_2$ the other, and still get the correct answer. In this case, by the above conditioning rule, the same state would have to be conditioned to different responses. Since this can't happen, either the conditioning rule must be changed, or the organism cannot be regarded as acting like a probabilistic fsa. For brute force, we didn't want to change the conditioning rule, so we decided that on a given trial responses when a given state is reentered will be determined by the response given the first time the state was entered. Even with this conditioning rule, it is obvious that some states could be conditioned incorrectly. Thus, it was necessary to introduce a deconditioning rule.

For brute force, we decided that whenever a wrong answer is given, all states will be deconditioned. Thus, the organism will have to start over. This method will eventually learn, since once all the states are correctly conditioned, they will never become unconditioned. Brute force lacks the good features of enumeration, but it does have a simple learning procedure.

The next thing we did was to reformulate two features of brute-force learning to make it more like an S-R model. We reversed our previous decision and decided that it would be better to change the conditioning rule and regard the organism as acting like a probabilistic fsa. The way to do this is to simply introduce an order in which the way states are conditioned. We took the natural course of specifying that, after reinforcement, unconditioned states are conditioned in the order they were entered on the trial, the first such state being conditioned first. The second thing was to say that only conditioned states that were used on the given trial are subject to deconditioning. These changes make it possible to write axioms very similar to those of SRTFA which lead to the desired asymptotic result.

We considered two different kinds of conditioning in this framework: all-or-none and linear. In the case of linear conditioning where there are only two responses, the conditioning and deconditioning procedures are similar. Indeed, deconditioning looks exactly like conditioning for us, since we took the deconditioning parameter to be equal to the learning parameter $\Theta$. The linear method works in the following way: Suppose at the beginning of a trial the probability of responding $r_k$ is $P_k$ when in state $(r_i,\sigma_j)$, i, j and k = 1 or 2. If $(r_i,\sigma_j)$ is entered and $r_k$ is given, then if the final response is correct, the probability of responding $r_k$ the next time $(r_i,s_j)$ is entered is $p_k(1-\Theta)+\Theta$, while if the response is incorrect, this probability is $p_k(1-\Theta)$. Once the original probabilities are specified, this is sufficient to determine all the probabilities, since there are only two possible responses, the probabilities of which must sum to one. It is possible for a certain response to be both incremented and decremented on the same trial, e.g., if $\sigma_1\sigma_1$ is the presented tape and it should be accepted, and the initial state is $r_1$, then it is possible for the responses to be

$r_1$ and then $r_2$. This answer is incorrect, so the transition from $(r_1,\sigma_1)$ to $r_1$ is decremented because of the first response and incremented because of the second one. In this set up, incorrect responses get reinforced, and it is not clear whether or not the probabilities of all the transitions will converge to 0 or 1 as the number of trials increases. It seems as though they might not in the case where $\{\sigma_1\sigma_1,\sigma_2\sigma_2\}$ is the acceptable set, since there are two possibilities that have all connections different, so we chose to concentrate on a case where this doesn't happen. The case we chose is where the acceptable set is $\{\sigma_1\sigma_1\}$. In this case, there is only one possibility (which is given by the following transition table), for which we came up with special names for the transition probabilities (given in the second table):

| | $r_1$ | $r_2$ |
|---|---|---|
| $r_1\sigma_1$ | 1 | 0 |
| $r_1\sigma_2$ | 0 | 1 |
| $r_2\sigma_1$ | 0 | 1 |
| $r_2\sigma_2$ | 0 | 1 |

| | $r_1$ | $r_2$ |
|---|---|---|
| $r_1\sigma_1$ | $a_1$ | $b_1$ |
| $r_1\sigma_2$ | $b_2$ | $a_2$ |
| $r_2\sigma_1$ | $b_3$ | $a_3$ |
| $r_2\sigma_2$ | $b_4$ | $a_4$ |

The states are numbered from 1-4, and $a_i$ is the probability of giving a correct response when in state i. With this notation, it is easy to construct the following table, which tells both which combinations of responses result in correct answers, and what the probabilities of such combinations are:

| Tape | Correct | | Incorrect | |
|---|---|---|---|---|
| $\sigma_1\sigma_1$ | $a_1a_1$ | $b_1b_3$ | $a_1b_1$ | $b_1a_3$ |
| $\sigma_1\sigma_2$ | $a_1a_2$ | $b_1a_4$ | $a_1b_2$ | $b_1b_4$ |
| $\sigma_2\sigma_1$ | $a_2a_3$ | $b_2b_1$ | $a_2b_3$ | $b_2a_1$ |
| $\sigma_2\sigma_2$ | $a_2a_4$ | $b_2a_2$ | $a_2b_4$ | $b_2b_2$ |

From this table, it is easy to see that $a_4$ will converge to 1, since it is always correctly reinforced ($a_4$ occurs only in correct column, $b_4$ only in incorrect column). Assuming that each tape occurs with probability $\frac{1}{4}$, the occurrence of $b_1 a_3$ in the incorrect column means $a_3$ can't converge to 1 unless $a_1$ does. This can easily be checked by computing the expectations. Similarly, $a_1$ can't converge unless $a_2$ does, and $a_2$ can't unless $a_3$ does. Thus, they all converge together, or none do. Moreover, they tend to cluster together, high $a_i$'s pulling low ones up and vice versa. What I tried to do was show that if all three got within some distance $\epsilon$ of 1, they would converge to 1. It seemed that some such procedure would be necessary to take care of the cases where there are more than one possible transition table. I couldn't come up with anything, and, in fact, I soon became convinced they didn't converge. I knew no good way to prove this, and since it became apparent that the all-or-none model would converge, we simply dropped the linear conditioning model. Moreover, it seems that even a model with stronger tendencies to converge, such as Luce's beta model, won't help, since there is just about as strong a tendency for incorrect responses to be reinforced as correct responses. In retrospect, it seems that the reason the brute-force method works is that sooner or later the correct responses get conditioned and once this happens only correct answers are given; thus, no negative reinforcements occur, and hence no states are deconditioned. This can't occur in models that have to have their probabilities converge to 1.

In the all-or-none conditioning model, each state is in one of two situations: conditioned, in which case it is conditioned to some response $r_k$, and unconditioned. If it is conditioned, then it responds with the response it is conditioned to with probability 1, and if it is unconditioned, there is a constant positive probability of responding any of the possible responses. Thus, the only difference between the revised brute-force method and the all-or-none conditioning model is that the latter has conditioning and deconditioning parameters c and d, respectively, $0 < c, d < 1$. Instead of all states that are entered on a given trial being conditioned with probability 1 when a correct answer is given, as in brute force, they are conditioned with probability c. Similarly, when an incorrect answer is given, the entered states are deconditioned with probability d.

The problem now is to formalize the correct S-R model and prove the desired asymptotic result. In the present S-R model, the stimuli and responses are treated the same as in SRTFA, the set S of stimuli and set R of responses both being primitive concepts. In the present set-up, however, we need an added primitive concept, that of $R_y$, which is a specified subset of R. Intuitively, if the final response on a trial is in $R_y$, then the model is regarded as having responded yes, and if the response is in $R_n = R - R_y$, then the model is regarded as having responded no. The set E of reinforcements is also primitive, but it contains only two elements, $e_1$ and $e_2$, rather than having an element corresponding to each response as in SRTFA. The fifth primitive concept is a measure $\mu$ on the set of stimuli, and is exactly the same as in SRTFA. The concept of subtrial requires the introduction of a new primitive concept M, which is a sequence of positive integers $m_n$. Each $m_n$ indicates the number of subtrials on trial n. This notion is necessary in defining the next primitive concept, that of the sample space X. Each element of X represents a possible experiment, i.e., an infinite sequence of trials, where each trial n has $m_n$ subtrials. Each trial is an $(m_n + 2)$-tuple, consisting of three things: 1) the conditioning function at the beginning of the trial which is a partial function from S into R, where $C(\sigma) = r$ means $\sigma$ is conditioned to r and $C(\sigma)$ undefined means $\sigma$ is unconditioned; 2) $m_n$ triples of the form $(T,s,r)$ each of which represents the presented set, sampled stimuli and response on a subtrial; and 3) the reinforcement which occurred. The seventh and final primitive concept is the probability measure P on the appropriate Borel field of cylinder sets of X, which is easily defined since there are only finite number of stimuli and responses. All probabilities must be defined in terms of P.

Some notation is needed to take us back and forth between elements or subsets of the sets of stimuli, responses, and reinforcements to events of the sample space X. I will follow the notation of SRTFA as closely as possible. $T_{n,m}$ is the event of set T being presented on trial n, subtrial m, i.e., it is the set of all elements of X that have T as the presented set on trial n, subtrial m. When this notation is used, I always suppose that $1 \leq m \leq m_n$. $s_{n,m}$ and $r_{n,m}$ are defined analogously.

There is no need to mention subtrial when speaking of reinforcement and conditioning. Thus, $e_{1,n}$ is the subset of elements of X in which $e_1$ occurs on trial n. $C_n$ is the event of conditioning function C occurring on trial n. I will write $\sigma \in C$ to mean $\sigma \in \text{domain}(C)$, and $\sigma \in C^r$ to mean $C(\sigma) = r$.

For each possible experiment X, and each element z of a trial of X (either a conditioning function, a triple of the form (T,s,r) or a reinforcement), Y(z) is the pattern of events preceding (and including z), i.e., Y(z) is the set of all elements of X that are the same as x up to and including z. I will write $Y(T_{n,m}, s_{n,m}, r_{n,m})$ simply as Y(n,m).

Finally, conditioning takes place all at once in this model, but it is necessary to think of states as being (possibly) conditioned in the order they occur on a trial. This is most convenient to state if we introduce the notation $C_n^m$ for the conditioning function on trial n after response m has (possibly) been conditioned. I use superscripts, since $C_n^m$ is not explicitly a part of the sample space X, unlike $T_{n,m}$, $s_{n,m}$, $r_{n,m}$ and $C_n$. Also, $C_n^{m_n} = C_{n+1}^0$.

In the following axioms, it is assumed that all events on which probabilities are conditioned have positive probability. For example, the tacit hypothesis of S2 is that $P(T_{n,m})$ and $P(T_{n',m'}) > 0$. There are three kinds of axioms: sampling axioms; conditioning axioms; and response axioms. A verbal formulation of each axiom is given together with its formal statement.

Definition: A structure $\mathscr{A} = (S,R,R_y,E,\mu,M,X,P)$ is an S-R model if and only if the following axioms are satisfied:

Sampling Axioms.

S1. $P(\mu(s_{n,m}) > 0) = 1$.
(On every subtrial a set of stimuli of positive measure is sampled with probability 1.)

S2. $P(s_{n,m} | T_{n,m}) = P(s_{n',m'} | T_{n',m'})$.
(If the same presentation set occurs on two different subtrials, then the probability of a given sample is independent of the subtrial number.)

S3. If $s \cup s' \subseteq T$ and $\mu(s) = \mu(s')$, then $P(s_{n,m} | T_{n,m}) = P(s'_{n,m} | T_{n,m})$.
(Samples of equal measure that are subsets of the presentation set have an equal probability of being sampled on a given subtrial.)

S4. $P(s_{n,m} | T_{n,m}, Y(n,m)) = P(s_{n,m} | T_{n,m})$.
(The probability of a particular sample on trial n, subtrial m, given the presentation set of stimuli, is independent of any preceding pattern Y(n,m) of events.)

Conditioning Axioms.

C1. If $r, r' \in R, r \neq r'$ and $C^r \cap C^{r'} \neq 0$, then $P(C_n) = 0$.
(On every trial with probability 1 each stimulus element is conditioned to at most one response.)

C2. $P(\sigma \in (C_n^{m+1})^r \mid \sigma \in s_{n,m}, \sigma \notin C_n^m, r_{n,m+1} = r, e_{1,n}, Y(n,m)) = c$.
(If $e_1$ occurs on trial n, the probability is c of any previously unconditioned stimulus that is sampled on a subtrial becoming conditioned to the response given on that subtrial and this probability is independent of the particular subtrial and any preceding pattern of events Y(n,m).)

C3. $P(\sigma \in (C_n^{m+1})^r \mid \sigma \in s_{n,m}, \sigma \notin C_n^m, r_{n,m+1} \neq r, e_{1,n}, Y(n,m)) = 0$.
(If $e_1$ occurs on trial n, the probability is 0 of any previously unconditioned stimulus that is sampled on a subtrial becoming conditioned to a response different from the one given on that subtrial and this probability is independent of the particular subtrial and any preceding pattern of events Y(n,m).)

C4. $P(\sigma \in (C_n^{m+1})^r \mid \sigma \in s_{n,m}, \sigma \in (C_n^m)^r, e_{1,n}, Y(n,m)) = 1$.
(If $e_1$ occurs on trial n, the conditioning of previously conditioned sampled states remains unchanged.)

C5. $P(\sigma \in C_n^{m+1} \mid \sigma \in s_{n,m}, \sigma \notin C_n^m, e_{2,n}, Y(n,m)) = 0$.
(If $e_2$ occurs on trial n, the probability is 0 of a previously unconditioned stimuli that is sampled on a subtrial becoming conditioned.)

C6. $P(\sigma \notin C_n^{m+1} | \sigma \in s_{n,m}, \sigma \in C_n^m, e_{2,n}, Y(n,m)) = d.$
(If $e_2$ occurs on trial n, the probability is d of any previously conditioned stimulus that is sampled on a subtrial becoming unconditioned and this probability is independent of the particular subtrial and any preceding pattern of events Y(n,m).)

C7. $P(\sigma \in (C_n^{m+1})^r | \sigma \notin s_{n,m}, \sigma \in (C_n^m)^r) = 1.$
(With probability 1, the conditioning of unsampled stimuli does not change.)

Response Axioms.

R1. If $\bigcup_{r \in R} C^r \cap s \neq 0$ then $P(r_{n,m} | C_n, s_{n,m}, Y(n,m)) = \frac{\mu(s \cap C^r)}{\mu(s \cap \cup C^r)}$.
(If at least one sampled stimulus is conditioned to some response, then the probability of any response is the ratio of the measure of sampled stimuli conditioned to this response to the measure of all the sampled conditioned stimuli, and this probability is independent of any preceding pattern Y(n,m) of events.)

R2. If $\bigcup_{r \in R} C^r \cap s = 0$ then there is a number $\rho_r$ such that $P(r_{n,m} | C_n, s_{n,m}, Y(n,m)) = \rho_r$.
(If no sampled stimulus is conditioned to any response, then the probability of any response r is a constant guessing probability $\rho_r$ that is independent of n and any preceding pattern Y(n,m) of events.)

As indicated earlier, the sampling and response axioms are exactly the same as in SRTFA, except that the concept of trial has been replaced by that of subtrial. Only the conditioning axioms have had to be changed to ensure the desired learning.

I will use only a very special kind of S-R model, one that has a natural relationship to fsa's. Before specifying the restrictions that are necessary, something must be said about fsa's. In the following, i, k and $\ell$ are used as subscripts for states of an fsa and responses, hence $1 \leq i, k, \ell \leq h$, and j is used as a subscript for letters of the alphabet of an fsa and stimuli, hence $1 \leq j \leq g$.

Definition: The quadruple $v = (g,h,p,H)$ is a signature if $g$, $h$ and $p$ are positive integers, $1 \leq p \leq h$, and $H \subseteq \{1,2...h\}$.

Definition: If $v = (g,h,p,H)$ is a signature, then $\mathcal{A}(v) = \{D$: D is a probabilistic fsa with alphabet A containing g elements (denoted by $a_1...a_g$), set of states Q containing h elements (denoted by $q_1...q_h$), initial state $q_p$ and set of final states F, where $q_i \in F \Leftrightarrow i \in H$, such that for all i and j, when D is in state $q_i$ and scanning $a_j$, it makes the transition to some $q_k$ with probability 1 (in which case $(q_i,a_j)$ is said to be conditioned), or for all k, it makes the transition to $q_k$ with positive probability (in which case $(q_i,a_j)$ is said to be unconditioned)$\}$.

Definition: $\mathcal{A}_1(v) = \{$deterministic fsa's with alphabet A containing g elements (denoted by $a_1...a_g$), set of states Q containing h elements (denoted by $q_1...q_h$), initial state $q_p$ and set of final states F, where $q_i \in F \Leftrightarrow i \in H\}$.[1]

If $D \in \mathcal{A}_1(v)$, I will say all states in D are conditioned.

Definition: If $D \in \mathcal{A}_1(v)$, then $(q_i,a_j)$ is said to be indifferent in D if $\forall w \in A^*$, D accepts w independently of the state of conditioning of $(q_i,a_j)$.

All states that are inaccessible in D are indifferent in D. The converse is false. For example, if F = 0 or F = Q, all states are indifferent, but the initial state, in particular, is not inaccessible. A non-trivial example would be the case of an fsa that ignores the first letter in each of the words it is presented with and never reenters the initial state. In such a case, the initial state is indifferent, but it is necessary in the sense that if the fsa has the minimum number of states possible (which can occur), it is impossible to delete the initial state from the set of states and still get an fsa that accepts the same words.

Let $A^*$ be the set of all words in the alphabet A, and $G \subseteq A^*$ the set of words we want to be accepted.

Definition: If $A' \subseteq A^*$, then $\mathcal{A}_0(v,A') = \{D \in \mathcal{A}_1(v)$: D accepts all and only elements of $A'\}$.

The only case of interest is $\mathcal{A}_0(v,G)$.

Definition: If $D \in \mathcal{A}_1(v)$, then $A_D = \{w \in A^*$: D accepts $w\}$.

Let $\overline{A'}$ and $\overline{A_D}$ be the complements of $A'$ and $A_D$ (relative to $A^*$).

---

[1] This is my original definition. I altered the definition in the final draft to make the proof a little slicker. Unfortunately this change made the proof invalid. Fortunately Nancy Moler called this (and sundry minor errors) to my attention before printing.

Definition: If $D \in \mathcal{D}_1(v)$ and $A' \subseteq A^*$, then $\Delta_{D,A'} = (A_D \cap \overline{A'}) \cup (\overline{A_D} \cap A')$. Again, the only case of interest is $\Delta_{D,G}$, which I shall simply write as $\Delta_D$.

Definition: If $D \in \mathcal{D}_1(v)$, $G \subseteq A^*$ and t is a positive integer, then $\Delta_D^t = \{w \in \Delta_D : \text{length } (w) \leq t\}$.

The purpose in defining $\Delta_D^t$ will be evident shortly.

The problem of choosing the number of states, which was discussed earlier, is the same as choosing an appropriate h. g is determined by the task to be performed, but h, p, and H must be chosen by the organism. The crucial step is choosing h, since once h is determined, p and H can be obtained fairly easily. As mentioned earlier, we found no way to determine h given the task, and simply assumed that enough states were present, which is expressed in the final theorem by the requirement that $\mathcal{D}_0(v,G)$ is non-empty.

If $S = \{\sigma_1 \ldots \sigma_g\}$ is a set of stimuli, S* is the set of words in S, and $Z \subseteq S^*$. S corresponds to A, S* to A* and Z to G. I now want to define a class of S-R models $\underline{S}(v,Z)$ and show how this class corresponds to $\mathcal{D}(v)$ and $\mathcal{D}_0(v,G)$. Let $\mathcal{S}$ be an element of $\underline{S}(v,Z)$. S and R are still taken to be primitive, but in the definition of $\mathcal{S}$, the role of S is taken by RxS. The reason for this was indicated in the discussion at the beginning of this chapter.

Definition: If $\mathcal{S} = (RxS, R, Ry, E, \mu, M, XP)$ is an S-R model, and on each subtrial $T_{n,m} = \{(r_i, \sigma_j)\}$ for some i and j, then $\sigma_{n,m}$ is the element of S occurring in $T_{n,m}$, and $\underline{\sigma_n^*} = \sigma_{n,0}, \sigma_{n,1} \ldots \sigma_{n,m_n}$.

Thus, $\sigma_n^* \in S^*$, and corresponds to word in A*.

Let f be the natural map from $A \cup Q$ onto $S \cup R$, i.e., $f(a_j) = s_j$ and $f(g_i) = r_i$. f maps A* onto S* and pairs $(q_i, a_j)$ onto pairs $(r_i, s_j)$. The relationship between $\underline{S}(v,Z)$ and $\mathcal{B}(v)$ is that for each $\mathcal{S}$ in $\underline{S}(v,Z)$, f maps the set $\mathcal{C}_{\mathcal{S}}$ of possible conditioning functions of $\mathcal{S}$ onto $\mathcal{B}(v)$. For $C \in \mathcal{C}_{\mathcal{S}}$, conditioned states in C correspond to conditioned states in the corresponding element of D of $\mathcal{D}(v)$, and unconditioned states of C correspond to unconditioned states of D. Indeed, this fact is the

reason for introducing the conditioning terminology into the definition of $\mathcal{D}(v)$. If $f(G) = Z$, then $\mathcal{D}_0(v,G)$ corresponds to the set of possible correct values of the asymptotic conditioning function of all $\mathcal{A} \in \underline{S}(v,Z)$, i.e., if $C \in \mathcal{C}_{\mathcal{A}}$ is a possible asymptotic conditioning function of $\mathcal{A}$ ($\mathcal{A}$ always responds correctly when the conditioning function is C) then $\exists D \in \mathcal{D}_0(v,G)$ s.t. all states in C are either conditioned the same as in D or are indifferent in D.

Finally, corresponding to $\Delta_D^t$, is $\Delta_C^t$.

Definition: If $\mathcal{A} = (RxS,R,R_y,E,\mu,M,X,P)$ is an S-R model, $Z \subseteq S^*$, and C a conditioning function of $\mathcal{A}$ s.t. $\forall x \in S^*$, if $\sigma_n^* = x$, then $P(r_{n,m_n} \in R_y) = 0$ or 1, then $\underline{S_C} = \{x \in S^*: \sigma_n^* = x \Rightarrow r_{n,m_n} \in R_y\}$ and $\underline{\Delta_C} = \{x \in S^*: x \in (\overline{S_C} \cap Z) \cup (\overline{Z} \cap S_C)\}$ and $\underline{\Delta_C^t} = \{x \in \Delta_C: \text{length } (x) \leq t\}$.

Whenever the notation $\Delta_C$ or $\Delta_C^t$ is used, it will be assumed that C satisfies the condition that $\forall x \in S^*$, if $C = C_n$, then $P(r_{n,m_n} \in R_y) = 0$ or 1.

Definition: If $v = (g,h,p,H)$ is a signature and $Z \subseteq S^*$, then $\underline{S}(v,Z)$ is the set of all S-R models $\mathcal{A} = (RxS,R,R_y,E,\mu,M,X,P)$ satisfying the following conditions:

i) S has g elements, denoted by $\sigma_1 \ldots \sigma_g$.

ii) R has h elements, denoted by $r_1 \ldots r_h$.

iii) $r_i \in R_y \Leftrightarrow i \in H$.

iv) $\forall n, T_{n,1} = (r_p,\sigma_j)$ for some j.

v) $\forall n$, $\forall m$ s.t. $1 < m \leq m_n, T_{n,m} = (r_{n,m-1},\sigma_j)$ for some j.

vi) $\mu(S')$ is the cardinality of $S'$ for $S' \subseteq RxS$.

vii) $\rho_i$, the probability of responding $r_i$ when no sampled stimuli are conditioned, is $> 0$.

viii) $e_1$ occurs on trial n if and only if $\sigma_n^* \in Z$ & $r_{n,m_n} \in R_y$ or $\sigma_n^* \notin Z$ & $r_{n,m_n} \in R_n$.

ix) $\forall C$ s.t. $\forall x \in S^*$, if $C = C_n$ and $\sigma_n^* = x$, then $P(r_{n,m_n} \in R_y) = 0$ or 1, $\Delta_C \neq 0 \Rightarrow (\exists \epsilon > 0)$ s.t. $(\forall n) P(\sigma_n^* \in \Delta_C^t) > \epsilon$.

Conditions i, ii, and iii guarantee a natural correspondence between A and S, Q and R, and F and $R_y$. Conditions iv and v guarantee that each $T_{n,m}$ is a singleton, and since $\mu$ is the cardinality of a set, Axiom S1 guarantees that its single element will be sampled. Thus, $T_{n,m} = s_{n,m}$, for all m and n. I will henceforth say simply that a pair $(r_i,\sigma_j)$ is the stimulus on the subtrial on which it occurs. Since each $s_{n,m}$ is a singleton, Axiom R1 guarantees that if $(r_i,\sigma_l)$ is conditioned, the response to which it is conditioned will be given with probability 1, while condition vii strengthens Axiom R2 so that if $(r_i,\sigma_i)$ is unconditioned, each response has a positive probability of being given. Condition iv guarantees that the response in the first stimulus pair is $r_p$, which corresponds to the requirement that the initial state of the fsa's is $q_p$. Condition v guarantees that the stimulus on each succeeding subtrial consists of the previous response and an element $\sigma_j$ of S. Altogether, this has the result that if $\sigma_n^* = f(w)$ and $C_n = f(D)$, $\mathscr{A}$ acts just like D would when presented with input w. Put more precisely, let $w = a_{1_w} \ldots a_{t_w}$, where length(w) = t. Let $q_{0_w}$ be the initial state of D, and $q_{i_w}$ the state D goes into after scanning $a_{i_w}$. The action of D on w is described completely by the following (2t+1)-tuple, $(q_{0_w},a_{1_w},q_{1_w}\ldots a_{t_w},q_{t_w})$. Let $r_{0_n}$ be the element of R in the stimulus on the first subtrial of trial n, $\sigma_{i_n}$ the element of S on subtrial i and $r_{i_n}$ the response given on subtrial i. The action of $\mathscr{A}$ on $\sigma_n^*$ is similarly described by the following (2m+1)-tuple, where $m = m_n$ to avoid cumbersome notation: $(r_{0_n},\sigma_{1_n},r_{1_n}\cdots \sigma_{m_n},r_{m_n})$. If D has no unconditioned states, $f(D) = C_n$, and $\sigma_n^* = f(w)$, then length (w) = m and the fact that D and $\mathscr{A}$ act the same is shown by the fact that for all i and j, $f(q_{i_w}) = r_{i_n}$ and $f(a_{j_w}) = \sigma_{j_n}$. If D has unconditioned states, these will correspond to unconditioned states in $C_n$. The function f doesn't say anything about the probabilities of the different responses, but the exact probabilities are inessential as long as they are all positive, and this is true of D since $D \in \mathscr{D}(v)$ and it is true for $C_n$ because of condition vii.

Intuitively, Z is the subset of S* whose elements should get a yes answer. Thus, the requirement in condition viii that $\sigma_n^* \in Z \,\&\, r_{n,m_n} \in R_y$ or $\sigma_n^* \notin Z \,\&\, r_{n,m_n} \in R_n$ is equivalent to saying that the answer given on trial n is correct. Condition viii is therefore equivalent to saying that a positive reinforcement occurs on trial n if and only if the answer given on trial n is correct.

$\Delta_C$ is the set of all elements of S* to which $\mathscr{S}$ gives an incorrect response if $C_n = C$. If $\Delta_C \neq 0$, then the conditioning function is incorrect. Condition ix requires that stimuli be presented that will cause incorrect conditioning functions to be deconditioned. It is sufficient to require only that $P(\sigma_n^* \in \Delta_C^t) > \epsilon$ for C's that answer deterministically, and this is why $\Delta_C$ was defined only for these C's.

Theorem: $(\forall v)(\forall G)[\mathscr{S}_0(v,G) \neq 0 \Rightarrow (\forall \mathscr{S} \in \underline{S}(v,f(G))P(r_{n,m_n} \in R_y \Leftrightarrow \sigma_n^* \in f(G)) \underset{n\to\infty}{\to} 1]$

Proof. Suppose $\mathscr{S} \in \underline{S}(v,f(G))$. The condition that $r_{n,m_n} \in R_y \Leftrightarrow \sigma_n^* \in f(G)$ is equivalent to $u(C_n) \in \mathscr{S}_0(v,G)$, where u is the mapping $f^{-1}$ and $u(C_n) \in \mathscr{S}_0(v,G)$ means $\exists D \in \mathscr{S}_0(v,G)$ such that all states in $u(C_n)$ are conditioned the same as in D or are indifferent in D. The strategy of the proof is similar to that of SRTFA; I will show that on each trial there is a positive probability of incorrectly conditioned states becoming deconditioned, not indifferent states becoming correctly conditioned. This will be done in two lemmas, but first, two definitions and one preliminary fact are needed.

Definition: $(\forall C \in \mathscr{C}_\mathscr{S})(\forall D \in \mathscr{S}_1(v))\underline{W(D,C)} = \{(r_i,\sigma_j)$: $(r_i,\sigma_j)$ is conditioned in C, but is conditioned to a different response in C than in f(D)$\}$.

Definition: $\forall D \in \mathscr{S}_1(v)$, $F_n^D$ is the event of all responses on trial n being compatible with $C_n = f(D)$.

$F_n^D$ is a rather special event, since $D \in \mathscr{S}_1(v)$ means all states in f(D) are conditioned. Let $\rho = \min \rho_i$. By condition vi, $\rho > 0$. If $m_n \leq t$, and $W(D,C_n) = 0$, then $P(F_n^D) \geq \rho^t$, since for each subtrial on which an unconditioned stimulus which is conditioned in f(D) occurs, the probability is $\rho$ of the response being given to which the stimulus is conditioned in f(D). Since there are at most t subtrials, the result follows.

Lemma 1. $(\forall n)[(\forall D \in \mathcal{D}_0(v,G))W(D,C_n) \neq 0 \Rightarrow (\exists \delta' > 0)$ s.t. $(\forall D \in \mathcal{D}_0(v,G))P$ (at least one element in $W(D,C_n)$ is deconditioned and no remaining stimuli are conditioned on trial n) $> \delta']$

Proof. Let n be any trial number and $D' \in \mathcal{D}_0(v,G)$. Let $D \in \mathcal{D}_1(v)$ be such that f(D) has all states in $W(D',C_n)$ conditioned as in C, and all other states conditioned as in D'. $W(D,C_n) = 0$, so $D \notin \mathcal{D}_0(v,G)$ and $\Delta^t_D \neq 0$. This means that $\Delta^t_{f(D)} \neq 0$, and hence, by condition ix, $P(\sigma^*_n \in \Delta^t_{f(D)}) > \epsilon$. If $\sigma^*_n \in \Delta^t_{f(D)}$, length$(\sigma^*_n) \leq t$. Since $W(D,C_n) = 0$, $P(F^D_n) \geq \rho^t$. If $\sigma^*_n$ and $F^D_n$, then $P(e_{2,n}) = 1$, since $\Delta^t_{f(D)}$ is the subset of S* to which $\mathcal{L}$ responds incorrectly if $C_n = f(D)$. Thus, on at least one subtrial a conditioned stimulus, say $(r_i,\sigma_j)$, must have been in $W(D',C_n)$. If no $(r_i,\sigma_j) \in W(D',C_n)$, then $F^{D'}_n$ is equivalent to $F^D_n$, since all the states not in $W(D',C_n)$ are conditioned the same in D and D'. Since $F^D_n$ occurs, $F^{D'}_n$ occurs. But this is impossible, since if $F^{D'}_n$ occurs, $P(e_{1,n}) = 1$, because $D' \in \mathcal{D}_0(v,G)$. Since $(r_i,\sigma_j)$ was a stimulus on some subtrial and $e_{2,n}$ occurs, by axiom C6 $(r_i,\sigma_j)$ will be deconditioned with probability d. Putting this together, $P((r_i,\sigma_j)$ is deconditioned on trial n) $\geq d P(\sigma^*_n \in \Delta^t_{f(D)}, F^D_n, e_{2,n}) > d\rho^t\epsilon$. Taking $\delta' = d\rho^t\epsilon$, we get the desired result, since by axioms C5 and C7 no stimuli can be conditioned when $e_2$ occurs.

Lemma 2. $(\forall n)(\forall D \in \mathcal{D}_0(v,G))[W(D,C_n) = 0 \Rightarrow (\exists \delta'' > 0)$ s.t. $\forall$ pairs $(r_i,\sigma_j)$ which are unconditioned in $C_n$ and not indifferent in D, $P((r_i,\sigma_j)$ is conditioned and no state is conditioned differently than in D on trial n) $> \delta'']$.

Proof. Let n be any trial number and $D' \in \mathcal{D}_0(v,G)$ be such that $W(D',C_n) = 0$. If no such D' exists, there is nothing to prove. If there are no unconditioned stimuli that are not indifferent in D', there is likewise nothing to prove, so assume there is at least one such, say $(r_i,\sigma_j)$. Let $r_k$ be the response that $(r_i,\sigma_j)$ is conditioned to in D' and let $D \in \mathcal{D}_1(v)$ be conditioned the same as D' except that $(r_i,\sigma_j)$ is conditioned to $r_\ell$. $r_k$ and D exist since $(r_i,\sigma_j)$ is not indifferent in D'. By argument similar to that of lemma 1, $P(\sigma^*_n \in \Delta^t_{f(D)}) > \epsilon$, and $P(F^{D'}_n) > \rho^t$. Also, $P(F^D_n) > \rho^t$ since $W(D,C_n) = 0$. If $F^D_n$ occurs, $P(e_{2,n}) = 1$ while if $F^{D'}_n$ occurs, $P(e_{1,n}) = 1$. Since D and D' differ only in the way $(r_i,\sigma_j)$ is conditioned, and since different answers occurs in the event

of $F_n^D$ and $F_n^{D'}$, $(r_i,\sigma_j)$ must be the stimulus on some subtrial. Putting this together, $P(\sigma_n^* \in \Delta^t_{f(D)}, F_n^{D'}, e_{1,n}) > \rho^t\epsilon$. Hence, by axiom C2, $P((r_i,\sigma_j)$ is conditioned on trial $n) > c\rho^t\epsilon$. Letting $\delta'' = c\rho^t\epsilon$, we get the desired result, since the fact that $F_n^{D'}$ and $e_1$ occurred means no state can be conditioned differently than in D' by axioms C3, C4, and C7.

For convenience, let $\delta = \min(\delta', \delta'')$. $u(C_n) \in \mathcal{D}_0(v,G)$ if and only if $\exists D \in \mathcal{D}_0(v,G)$ s.t. $W(D,C_n) = 0$ and $C_n$ has no unconditioned states that are not indifferent in D. Let k be any trial and $C_k$ any conditioning function. Choose $D \in \mathcal{D}_0(v,G)$ s.t. $W(D,C_k)$ has the minimum number of elements, say m. Lemma 1 guarantees that there is a probability $\delta^m$ that there will be at least one element D' of $\mathcal{D}_0(v,G)$ such that, for some k', $k \leq k' \leq k+m$, $W(D',C_{k'}) = 0$. k' might be less than $k+m$, since more than one state can be deconditioned on a trial. Moreover, lemma 1 doesn't guarantee that $D = D'$, for it can't be applied unless for all elements B of $\mathcal{D}_0(v,G)$, $W(B,C_n) \neq 0$, and it is possible that on some trial k' there is a $D' \neq D$ such that $W(D',C_{k'}) = 0$, so that lemma 1 will be inapplicable. Also, there is a probability that some of the correctly conditioned states will be deconditioned. Both of these cases are all right, since lemma 2 requires only that there be a D', and does not specify that any state in D' must be conditioned. In a sense, lemma 1 applies to the worst possible case, and the only cases where it might not apply is where what we want to happen has already occurred. Let D' be such that $W(D',C_{k'}) = 0$, and let m' be the number of unconditioned states that are not indifferent in D'. By lemma 2, on each trial P(such a state is conditioned and no state is conditioned differently than in D') $> \delta$, so after m' trials, P(no such states) $> \delta^{m'}$.

Once this occurs, the correct answer occurs with probability 1, so by condition viii, $e_1$ occurs with probability 1. By axioms C4 and C7, the conditioning of all conditioned states remains the same. Thus, only the conditioning of unconditioned states, which must be indifferent, can be changed, and if this occurs, the u of the resulting conditioning function is still in $\mathcal{D}_0(v,G)$.

m and m' are always $\leq gh$, so no matter what $C_k$ is, $\exists k' \leq k + 2gh$ such that $P(u(C_{k'}) \in \mathcal{D}_0(v,G)) > \delta^{2gh}$. By what was said in the preceding

paragraph, $P(\ddot{u}(C_{k+2gh}) \in \mathcal{D}_0(v,G)) > \delta^{2gh}$. Let $I(n,gh)$ be the greatest integer in $\frac{n}{gh}$. Then, regardless of the initial state of conditioning of $\mathcal{S}$, $P(\ddot{u}(C_n) \notin \mathcal{D}_0(v,G)) \leq (1-\delta^{2gh})^{I(n,gh)}$. This approaches 0 as n approaches infinity, so $P(\ddot{u}(C_n) \in \mathcal{D}_0(v,G))$ approaches 1 as n approaches infinity. Q.E.D.

A few remarks concerning this theorem are in order. The theorem gives a lower bound on the rate of learning, but the actual rate of learning will be much faster than this lower bound. In the usual case, the original conditioning function will have all states unconditioned, while the theorem allows for the possibility that all states are conditioned incorrectly. The lower bound also does not use the fact that more than one state can be conditioned (deconditioned) on a given trial. Moreover, $\delta$ was calculated using the minimum of the $\rho_i$, so the fact that there is a higher probability of some responses being given, and hence being conditioned or deconditioned, is ignored. Also, the minimum of c and d is chosen. Very importantly, it takes a sequence of gh trials to get the guaranteed result of the theorem, while in most cases a much shorter sequence is all that is necessary. Also, it is certainly possible for some states to be conditioned correctly even if $W(D,C_n) \neq 0$, which is not taken account of by the theorem. Finally, even if $\sigma_n^*$ is not a member of the appropriate $\Delta_C^t$ or if $F_n^D$ does not occur, there is a probability that some states will be correctly conditioned or that some incorrectly conditioned states will be deconditioned. Although it is obvious that the actual learning rate is much faster than the lower bound given by the theorem, calculating an actual expectation would be brutal. Thus, I have no precise results on how fast learning would actually occur. However, it is probably true that the process as it stands would be adequate for only fairly simple tasks, since it would be too slow for more complex tasks. There are five reasons that this may not be as severe a limitation as might at first seem. First, it may turn out best to think of learning a complex task as combining previously learned simple tasks, and that it is only the simple tasks that have to be learned by the above method. Secondly, it would not be surprising if such basic learning took place slowly, although perhaps not as slowly as

the above set up requires. Thirdly, it might be possible to keep the above framework essentially unchanged and make some adjustments to get a faster rate of learning. Fourthly, in the above work, as in most psychological experiments, the rate of learning is given in terms of the number of trials needed to learn the task, while in ordinary talk, the rate of learning is given in terms of the amount of time needed. What the relationship between number of trials and amount of time is not very clear. It may be that a large number of trials corresponds to a short period of time, in which case the fact that the above learning requires many trials may not be a serious fault. Lastly, the above learning takes place with the minimal amount of information given on each trial, since all the reinforcement does is tell whether or not the final response is correct. No indication is given of where mistakes occurred or what the right procedure would have been. Most learning situations contain this other information, and when it is excluded in an artificial situation, the learning task is indeed made much more difficult.

# CHAPTER 4

## SUMMARY OF TECHNICAL WORK

Our work was an attempt to build a mathematical model of a device that could learn geometry. The best way to visualize it is to think of it as an attempt to connect some type of mathematical learning model to geometry. This is the general plan, but to get a specific problem, it is necessary to choose a particular type of learning model; and to make the problem manageable, one has to limit oneself to a fragment of geometry.

The learning model we chose was an S-R model. This choice is not unproblematical, for cognitive psychologists and linguists like Chomsky have denied the adequacy of S-R models for the type of learning we wanted. Their remarks have been mostly about language learning, but they are also applicable to our work with geometry. I have mentioned similarities between our work and linguistics in the previous chapters, and I will indicate shortly that the situation as far as language learning is concerned is similar to our present situation.

There are four reasons for choosing the S-R model in spite of the criticism it has received. First and foremost, it is the only learning model with any degree of mathematical sophistication. In choosing a learning model, the S-R model wins almost by default, for its critics have not produced a serious competitor. Chomsky, for example, makes a few remarks that indicate he thinks some sort of enumeration procedure is what is needed. He speaks of a device for learning language operating by selecting one member of the class of potential grammars on the basis

of primary linguistic data.[1] These remarks are not developed into a precise formal theory, however. Secondly, in the terminology of Chapter 3, Chomsky favors the enumeration method over the brute force method. While abstract criticism of the brute force method seems plausible, when it comes down to making a concrete choice between brute force and enumeration, it seems to me that our decision to concentrate on brute force is correct. Thirdly, the criticisms of S-R models have consisted of claims, not proofs, that they are inadequate. Whether or not they are in fact inadequate is an open question until such a proof is given. This leads me to my final point, which is that S-R theory is very much an alive area today, and modifications and improvements of S-R models are still being given. A really convincing proof of the inadequacy of S-R models would have to show not only that all present models are inadequate, but that it would be impossible to develop an adequate model within the S-R tradition. This would require the formulation of certain properties that S-R models must have. This formulation is currently lacking (I don't see how it could be given at the present time), and hence any proof of inadequacy is out of the question. Behind these last remarks is the view that a proof of inadequacy of any model of a certain type requires much more precision and rigor than proving the adequacy of a certain model, a point which Professor Suppes is fond of making. What has actually happened, I think, is that critics of S-R models have leveled their criticisms at early, fairly undeveloped versions of the model, and tended to ignore more recent developments and the possibilities of developing more adequate models within S-R tradition.

The fragment of geometry we considered is that part of geometry that can be encoded in the codings given in Chapter 2. Codings apply only to two-dimensional straight-line drawings, and the only geometrical predicates that apply to such drawings which can be recognized from a coding are 'connected' and those involving the recognition of polygons. However, this fragment could be augmented by adding further information and then coupled with the artificial intelligence work to get a device

---

[1] Chomsky, Aspects, p. 24f.

that could deal with real-life situations. It is not clear exactly how to do this, but the fragment we chose to concentrate on seems to be the natural starting place for such a project.

We did not try to connect an S-R model with this coded fragment of geometry directly. In between the S-R model and the fragment of geometry are two types of processing devices, fsa's and Turing machines. One doesn't want to deal with general Turing machines, since these devices have virtually unlimited calculating power. To get a realistic model of human behavior, it will be necessary to place some restriction on the type of Turing machine calculations that are acceptable. These restricted calculations I will call 'Turing-type procedures,' and it is these that we are interested in, though just what restrictions should be made is not clear. Certainly, there should be some sort of limit on the size of the machine and length of calculations involved, and perhaps other restrictions would also be desirable.

Thus, there are four originally unrelated elements that we dealt with: S-R models, fsa's, Turing-type procedures, and the fragment of geometry. The purpose of the work with the coding in Chapter 2 is to provide a connection between the fragment of geometry and Turing-type procedures. The purpose of the work on learning theory in Chapter 3 was to strengthen the connection between S-R models and fsa's that was established in SRTFA. Schematically, the situation as it is now is given by the following diagram:

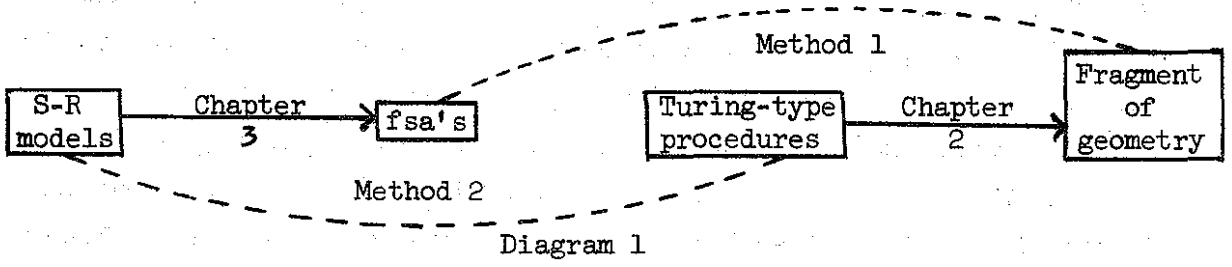



Diagram 1

Present Status of the Technical Problem

Our ultimate goal is to connect S-R models to the fragment of geometry, and, as the diagram indicates, this has not been accomplished. Considerable progress has been made, however. What we did was to start at the two ends of the problem and try to meet in the middle, and we weren't quite successful.

There are two possible ways of completing the connection. It is not possible to connect fsa's and Turing-type procedures directly, since the latter are provably more powerful. One method would be to try to connect the fragment of geometry to fsa's, which is indicated in the diagram as Method 1. I don't believe this method can be completely satisfactory, since the problems mentioned in dealing with the codings in terms of fsa's that were mentioned in Chapter 2 seem to be fundamental. This method might be partially satisfactory, however. The work in Chapter 2 should prove useful in making this connection, if it can be made. Method 2 seems to me to be much more promising. An fsa is essentially a set of triples, which a Turing machine is a set of quadruples, and there is no reason to believe that an S-R model cannot become a Turing machine at asymptote. The work in Chapter 3 would be a useful first step in making this connection. Whichever method is chosen, either the work on the coding or learning will be a necessary link in the final chain connecting S-R models and the fragment of geometry, and the work that is replaced (Chapter 2 if Method 1 is chosen, Chapter 3 if Method 2 is) should be useful in making the final connection. Thus, considerable progress in solving the problem has been made.

An analogy with linguistics can be drawn by replacing 'fragment of geometry' with 'natural languages.' The work of Chomsky, among others, has consisted primarily of an attempt to establish the connection between natural languages and transformational grammars, which are an example of what I called a Turing-type procedure. This work corresponds to the work in Chapter 2, though it has, of course, been more extensively developed than our work. Not much work on learning has been done by linguists, so it is hard to say what the left half of the diagram should look like. If they did start with an S-R model, the result would be exactly like Diagram 1, including the corresponding gap and methods for filling it. Thus, the remaining part of our problem is similar to problems in other areas, and solving it would have far-reaching implications.

This completes the picture of the overall problem. The status of our work in learning theory and what future developments might be have already been indicated, so in conclusion, I want to make some remarks on the present status of the coding problem and what future developments might occur. Working only with the coding places severe restrictions on the geometry that can be done. The size of angles and the length of line segments are not included in the coding, and it is impossible to distinguish concave from convex figures, what is inside a polygon from what is outside and the simple regions in a figure. Whether a figure is connected and the various polygons that it contains is all that one can hope to distinguish. Thus, theorems 1, 2, and 3 of Chapter 2 take care of the positive results that might be expected since they show that 'connected' and 'polygon' can be recognized. Moreover, they give reasonable procedures for accomplishing this. One problem with all three theorems, as Professor Hintikka pointed out in regard to theorem 3, is that each requires arbitrary choices and gives no strategy for making these choices. Finding optimal strategies, or discovering whether particular strategies make much difference, is a natural problem that has not been solved.

The fact that the coding contains such a limited amount of information means that something will have to be done to get more information. One way of doing this would be to allow what we called construction operations, i.e., allow auxiliary lines to be added to a figure and hence to a coding for it. This requires going back to the original figure, and this is undesirable. For example, it would be possible to recognize inside/outside by complicated procedures for adding lines, but this is very counterintuitive. A better method, it seems to me, would be to augment the coding by adding information concerning, say, inside/outside. Unfortunately, I have no suggestions on what would be the best way to do this. It seems as though it would be easy to add information concerning the length of line segments and sizes of angles, but such things as inside/outside are more difficult.

Finally, the most important unsolved problems concern the relationship between codings (or sets of lines), and figures. The outstanding problem is formulating necessary and sufficient conditions for a good set of lines to be a coding. This turns out to be a difficult problem, but there doesn't seem to be any reason a general solution can't be given. The problem would be easier if more information is added to the coding, but it should be solvable without this extra information. Related to this problem are the twin problems of under what groups of transformations codings remain invariant and what the classes of figures that have the same, or equivalent, codings look like.

BIBLIOGRAPHY

Berkeley, George. Essay, Principles, Dialogues, ed. Mary Calkins (New York: Charles Scribner's Sons, 1929).

Chomsky, Noam. Aspects of the Theory of Syntax (Cambridge: M.I.T. Press, 1969).

Chomsky, Noam. Cartesian Linguistics (New York: Harper & Row, 1966).

Chomsky, Noam, and Miller, George A. "Introduction to the Formal Analysis of Natural Languages," Handbook of Mathematical Psychology, ed. R. Duncan Luce, Robert R. Bush, and Eugene Galanter. Vol. II, Ch. 11 (New York: John Wiley and Sons, 1967).

Descartes, Rene. "Meditations," Descartes' Philosophical Writings, trans. Norman Kemp Smith (London: Macmillan & Co., 1952).

Goodman, Nelson. The Structure of Appearance (Indianapolis: The Bobbs-Merrill Company, 1966).

Helmholtz, H. von. Physiological Optics, trans. James P. C. Southhall (Rochester: The Optical Society of America, 1924-25).

Hopcroft, John E., and Ullman, Jeffrey D. Formal Languages and Their Relation to Automata (Reading, Mass.: Addison-Wesley, 1969).

Hume, David. A Treatise of Human Nature (Oxford, Clarendon Press, 1960).

Kant, Immanuel. Critique of Pure Reason, trans. Norman Kemp Smith (New York: St. Martin's Press, 1965).

Kant, Immanuel. Prolegomena to Any Future Metaphysic (Indianapolis: The Liberal Arts Press, 1950).

Locke, John. An Essay Concerning Human Understanding (London: George Routledge and Sons [n.d.]).

Minsky, Marvin and Papert, Seymour. Perceptrons (Cambridge: M.I.T. Press, 1969).

Moler, Nancy, and Suppes, Patrick. "Quantifier-Free Axioms for Constructive Plane Geometry," Compositio Mathematica, Vol. 20 (1968, pp. 143-152).

Moore, G. E. Some Main Problems of Philosophy (New York: Collier Books, 1966).

Nelson, R. J. Introduction to Automata (New York: John Wiley and Sons, 1968).

Papert, Seymour. Lecture on Visual Perception Project at M.I.T. (Stanford University, 1969).

Price, H. H. Perception (London: Methuen & Co., 1954).

Quine, Willard Van Orman. From a Logical Point of View (New York: Harper & Row, 1963).

Quine, Willard Van Orman. Word and Object (Cambridge: The M.I.T. Press, 1965).

Roberts, Fred S., and Suppes, Patrick. "Some Problems in the Geometry of Visual Perception," Synthese, 17 (1967), pp.173-201.

Suppes, Patrick. Set Theoretical Structures in Science (unpub.) (Loose-leaf, Stanford: Institute for Mathematical Studies in the Social Sciences, 1967).

Suppes, Patrick. "Stimulus Response Theory of Finite Automata," Journal of Mathematical Psychology, Vol. 16, No. 3 (October, 1969).

Swartz, Robert J. (Ed.). Perceiving, Sensing, and Knowing (New York: Doubleday, 1965).

Wittgenstein, Ludwig. Philosophical Investigations, trans. G. E. M. Anscombe (New York: The Macmillan Company, 1966).