BULLETIN OF
MATHEMATICAL BIOPHYSICS
VOLUME 31, 1969

LEARNED SPECIFICATION OF CONCEPT NEURONS

WAYNE A. WICKELGREN
Department of Psychology,
Massachusetts Institute of Technology,
Cambridge, Massachusetts 02139

The referential aspect of a concept can be defined by a disjunction of conjunctions of attributes. A single neuron can represent a disjunction of conjunctions of attributes if the assumption that neurons are single-threshold devices is discarded. Instead, one must assume that such concept neurons are composed of hundreds or thousands of (high-threshold) receptive areas, each containing tens or hundreds of synaptic sites. When essentially all of the sites of a receptive area are activated in close temporal contiguity, the receptive area generates a local (spike) response which is assumed to be sufficient to fire the cell body and axon of the neuron. If we assume that all concepts possessed by a single human being can be encoded by single neurons in this manner, there are enough neurons in the human cortex only if we assume that most of these concept neurons are specified by learning. Genetic specification is ruled out by the enormous (infinite?) number of possible concepts humans appear to be able to learn. Therefore, a speculative neural mechanism is presented regarding how "free" neurons could become specified by learning.

Introduction. I guess I must have been about seven when it first occurred to me how completely mysterious it was that reading a story was almost the same as having it happen in one's own life. I realized then that most of the time, while reading, I was in the world of the story almost completely, and that was what made reading stories so much fun. At the same time, a special concern gripped me, "What if I should lose that wonderful ability to have words conjure up (images of) the real world?" This concern came back to me almost every time I read for many years after that. Whenever the concern came, it distracted me from the story until I put it out of my mind. Gradually, the feeling

grew in me that the magic of words would stay with me forever, no matter how incapable I might be to explain that magic, and I no longer had to exert special effort to keep from worrying about this when I read. Now, many years later, I have a theory which, among other things, explains this magic of words. That theory is the subject of this paper.

The serious statement of the problem to be considered in this paper is as follows. Human beings perceive and think in terms of real-world concepts that are not simply described by patterns of peripheral sensory or motor neuron activity. Even the most "concrete" level of human thinking is in terms of physical objects and operations on physical objects—not in terms of patterns of sensory and motor activity. For example, an adult visually recognizes tens, perhaps hundreds of thousands, of familiar particular objects at thousands of discriminably different locations in the frontal plane, tens of discriminably different depths, hundreds of discriminably different sizes, and who-knows-how-many discriminably different rotational perspectives. Furthermore, some of these particular objects (for example, a face) have a set of different real configurations. In addition, most of these particular objects fall into one or more classes of objects, and it is primarily these general classes that we have words for and think in terms of. Many particular and general object concepts have a variety of patterns of auditory, tactile, etc. stimulation that evoke the concept just as effectively as any of the visual patterns. For a human being with language, verbal stimuli evoke the concept just as effectively as any of the nonverbal patterns of stimulation corresponding to the referent itself. This is the human ability for concept recognition. One aspect of the problem is generally called pattern recognition. The present paper considers how concept recognition and learning might be achieved by the human nervous system.

Since the present paper is only concerned with the referential aspect of concepts, not with their combinatorial aspect, attention will be focused, at first, on "concrete" concepts, such as physical objects that have well-defined physical referents. Later in the paper, the referential aspect of more abstract concepts will be discussed very briefly. The combinatorial aspect of concepts (e.g., how the meaning of a phrase, clause, sentence or paragraph, etc. is derived from the meanings of its constituents) will not be discussed.

Pattern Recognition. An attentive English-speaking human being has a high probability of thinking of the concept "dog" when confronted with a real dog, a picture of a dog, the printed word "dog", the spoken word "dog", etc. This can easily be demonstrated behaviorally by instructing the subject to verbalize the first one or more thoughts that come into his mind. No one, I believe, would call this an example of the human ability for pattern recognition. We

must give this ability another name because there is no common pattern in all of the various sensory events that can evoke the concept "dog". However, the ability seems closely analogous to pattern recognition, and so an appropriate name is "concept recognition".

Having generalized the problem of pattern recognition to the problem of concept recognition, the question arises as to whether the problem of pattern recognition deserves any separate mention at all from the problem of concept recognition in human beings.

The argument in favor of a separate problem of pattern recognition is that certain aspects of concept recognition do involve the recognition of a common pattern in different examples of a concept. For instance, it is alleged that there is often a common pattern in different examples of a particular capital letter from the same or similar type fonts, regardless of position on the retina, the presence of a "little" random visual noise, and sometimes also size and rotation.

The argument against considering pattern recognition as a problem separate from concept recognition is that there may not be any functional distinction between those cases where the examples of a concept share a common pattern and those cases in which they do not. Furthermore, it is not easy to define those cases in which the examples of a concept share a common pattern and those cases in which they do not, since this is as much dependent upon the nature of the perceiver as upon the nature of the perceived. In any event, the present paper will consider only the more general problem of concept recognition giving no special status to the problem of pattern recognition.

Genetic vs. Learned Specification of Concepts. The recent studies in sensory neurophysiology (e.g., Hubel and Wiesel 1959, 1962, 1963) provide solid support for the notion that there are several levels of analysis of the sensory input that are genetically specified. Thus, a visual stimulus is coded in many redundant ways, by many overlapping grids of "dot" detectors of different grain size, by many overlapping "grids" of "edge", "slit", and "corner", detectors, etc. At least in the cat, these visual analyzing systems are present at birth (Hubel and Wiesel, 1963), and the anatomical regularity of the analyzers also points to genetic specification. The complete inventory of visual analyzers and the exact specification of these analyzers is not yet known for any sensory modality of any organism, but existing data make it clear that we must consider the genetically specified levels of analysis to extend several synapses back from the receptors. If one considers a slit of light at a particular place on the retina to be a concept, then I suppose one must consider this sort of concept to be specified almost entirely by genetic mechanisms.

On the other hand, it is perfectly clear that most of the objects we have

names for have at least an arbitrary association with their name, which cannot be specified genetically and must be learned. Furthermore, human beings have the ability to lump chosen events into the same concept arbitrarily.

Certainly we cannot have a single genetically-specified neuron standing for each concept that it is possible for human beings to have, because there are too (infinitely?) many possible concepts. This suggests two alternative solutions to the problem of the representation of complex concepts.

First, complex concepts may be represented, not by single neurons, but only by sets of neurons, each of which represents a component attribute of an example of the concept. In this case, the attribute neurons may be genetically specified. However, any associations between attribute neurons of the same concept or between the attribute neurons of one concept and those of another concept or of a specific response representative must be learned, since these associations can be completely arbitrary. Call this the theory of "attribute specification of concepts".

Second, concepts may be represented by single neurons, because one individual does not appear to have more concepts than he has neurons. However, if single neurons encode concepts, then they must be specified by learning, not by genetic mechanisms. Call this the theory of "learned specification of concept neurons".

Attribute Specification of Concepts

Associative Interference Problem. The attribute specification of concepts has two major disadvantages. First, since the same attributes are used over and over again in different combinations for the different events signaling the same and different concepts, there is a considerable interference problem for associative learning. For example, the spoken "words" B (phonemically /biy/) and K (phonemically /key/) are letter names. If we assume (a) that the representative of a spoken word is the set of representatives of its component phonemes and (b) that there is a strong association between the representative of the spoken words /biy/ and /key/ on the one hand, and the representative of the concept "letter name" on the other hand, then the associations from /bey/ and /kiy/ to "letter name" should be as strong as those from /biy/ and /key/ to "letter name". How are we able to remember that /biy/ and /key/ are letter names, but /bey/ and /kiy/ are not?*

One possible alternative is to question the usual assumption of associative memory. Perhaps there is some kind of contingent associative memory, such that /b/ potentiates the association from /iy/ to "letter name" and /iy/ poten-

* This problem was posed to me by Harris Savin.

tiates the association from /b/ to "letter name". It may be that contiguity of activation of three or more neurons not only increases the strength of association between each pair, but also strengthens an association from each neuron to the *association* (connection) between each pair of neurons in the set contiguously activated. Conditionable presynaptic inhibitory or facilitatory synapses would be one obvious neural analogue of contingent associative learning. Unfortunately for this theory, there is no anatomical or physiological evidence for the existence of synapses on presynaptic axonal terminals in the cerebral cortex (Eccles, 1966). Furthermore, there is some behavioral evidence suggesting that truly contingent associative learning may be impossible (Wickelgren, 1969).

The other alternative is to question the phonemic coding assumption, and there are at least two ways to do this. First, one can substitute context-sensitive allophonic coding for phonemic coding. (For a full discussion of the advantages of context-sensitive allophonic coding over phonemic coding, see Wickelgren, 1969.) For example, /stop/ is coded as $/_{\#}s_{t}, {}_{s}t_{o}, {}_{t}o_{p}, {}_{o}p_{\#}/$. Now instead of about 50 phonemes, one has on the order of 50^3 context-sensitive allophones. With the increase in the number of attributes of a word, the discriminability of encoding of words increases enormously and associative interference is correspondingly reduced. However, associative interference is by no means eliminated, and there would seem to be considerable advantage in escaping from a system in which this associative interference is based on meaningless phonetic attributes into a system in which the associative interference is based on potentially useful semantic factors. The evidence from studies of human memory strongly supports the hypothesis that long-term verbal memory is in a semantically encoded system (McGeoch and McDonald, 1931; Underwood and Goad, 1951; Baddeley and Dale, 1966), while short-term verbal memory is predominantly encoded phonetically (Conrad 1962, 1964; Wickelgren 1965a, b, c, 1966).

Second, since it hardly seems conceivable that semantic attributes can be filtered out of the events signaling concepts, semantic encoding virtually requires transformation from phonetic attribute encoding (whether phonemic or, more likely, context-sensitive allophonic) to an encoding of each word as a unit that is completely distinct from other words. Later the semantic relationships between words must be established by means of learned associations. Thus, we are left with the conclusion that, at least at the level of verbal long-term memory, a word is not represented only by a set of attributes of any sensory event signaling it. The simplest alternative seems to be that each word (concept) is encoded by a single neuron. Semantic analysis of that word (concept) in relation to other words (concepts) is achieved only after a single

neuron has come to stand for the set of sensory attributes signaling at least one example of the word (concept).

Associative Transfer Problem. There is still a second disadvantage of attribute specification of concepts. That disadvantage concerns the fact that many sensory events signal the same concept. As has already been noted, these sensory events need not have any attributes in common, but even when they do have attributes in common, these common attributes are often not sufficient to distinguish events signaling one concept from events signaling other concepts. Not only does this multiplicity of sets of attributes signaling the same concept magnify the associative interference problem just discussed, but it also creates a problem in transfer of associations to a concept gained when one event is signaling the concept to situations in which a completely distinct sensory event is signaling the same concept.

Human beings do not seem to have such a problem. Associations formed to a word in one type font, color, size, location, rotational orientation (within limits), etc. will transfer to the same word in a different type font, color, size, location, rotational orientation, etc., or to the script form of the word, or to a picture of the object the word stands for, or to a whole host of different sensory events that signal the object itself. One can try to evade this difficulty by making an assumption (similar to one suggested by Hebb, 1949) that the attribute representatives of one event signaling a concept have come to be associated to the attribute representatives of every other event signaling the same concept, as well as to the attribute representatives of other concepts to which this concept is associated. Either in retrieval alone or in both storage and retrieval, the stimulus event activates either simultaneously or successively all the attribute sets associated with sensory events signaling the same concept. If this were done successively, it would undoubtedly take longer than human beings take to respond in rote learning tasks. If it were done simultaneously, it would solve the associative transfer problem, but only at the expense of making the associative interference problem orders of magnitude worse.

Attribute specification of concepts using filterable sensory attributes is incompatible with the facts. Attribute specification of concepts using semantic attributes seems to require prior representation of each concept by a single concept neuron (which is the theory favored in this paper), but perhaps someone can suggest another way that works equally well. I cannot.

Learned Specification of Concept Neurons

Concepts as Disjunctions of Conjunctions. We have said that a concept is often signaled by many events which have no common pattern sufficient to

distinguish events signaling one concept from events signaling other concepts. From a referential point of view, it is sufficient to say that the internal representative of a concept is a disjunction of conjunctions of the attributes of the sensory events signaling that concept. The purpose of the present section is to show that a single neuron could represent a disjunction of conjunctions of sensory attributes. To do this I will state rather carefully exactly what assumptions are being made concerning the class of neurons assumed to stand for concepts and just how these assumptions relate to the assumptions usually made in the neural network tradition inspired by Rashevsky (1938) and McCulloch and Pitts (1943).

Neurons as Logical Elements. The pioneering paper of McCulloch and Pitts (1943) shaped subsequent work on theoretical neural networks in many respects. Two of these respects are of primary importance to the present paper: (a) the assumption that neurons are logical operators (followed in this paper) and (b) the assumption that neurons are single-threshold operators (not followed in this paper).

McCulloch and Pitts explicitly stated that single neurons or small systems of neurons computed logical functions of their inputs. This interpolated a more explicit subgoal between the biophysics and biochemistry of synaptic and intraneuronal transmission, on the one hand, and the behavior of the organism, on the other hand. The subgoal was to define the elementary logical operations computed by neurons or small systems of neurons. Both the general assumption that the theoretical language of functional neurophysiology is some kind of mathematical logic and any specific theory about the nature of neural logic are to be validated in two separate steps. First, it must be shown that the physical and chemical properties of synaptic and intracellular transmission result in neurons or small systems of neurons that compute the logical functions specified by the theory. Second, it must be shown that the behavior of the organism is controlled by these elementary logical functions. The present paper will follow the McCulloch and Pitts tradition of interpolating a logical level between nerve membranes and organismic behavior.

Single-threshold Neurons. However, McCulloch and Pitts (1943) and everyone else in the nerve-net tradition assume that neurons are single-threshold devices. A weighted sum of the input lines (which are either in state 0, state 1 or state -1) is compared to a threshold. If the weighted sum exceeds a threshold, the output neuron goes into state 1. If not, it remains in (or goes into) state 0. Sometimes weights are equal; sometimes they are not. Usually time is quantized $(0, t, 2t, 3t, \ldots)$; occasionally time is continuous. Sometimes presynaptic inhibition is allowed in addition to postsynaptic inhibition;

other times it is not. Rather often, someone voices the concern that the combination of inputs (for comparison to the threshold) may not be a simple weighted arithmetic sum, but almost no one wants to get involved in that kind of complication (for an exception see Rall 1964 and 1967). What no one seems to question is the basic assumption that a neuron is a single-threshold device. As I shall attempt to show, several problems are solved by assuming that many neurons are multiple-threshold devices. In particular, I will assume that many (cortical) neurons are powerful logical elements which compute disjunctions of conjunctions.

Multiple-threshold Neurons. Let us imagine (and as far as I know, this is largely imagination) that the dendritic tree of a large class of neurons is subdivided into several hundred or thousand connected, convex, disjoint (non-overlapping) receptive areas, each of which contains tens or hundreds of synaptic sites. This total number of synapses per neuron is of approximately the right order of magnitude (see Cragg, 1967). Each receptive area is a single-threshold element, probably with all of its synaptic sites chemically similar to each other and chemically distinct from the sites in all neighboring receptive areas. Application of the famous 4-color conjecture in topology indicates that only 4 chemically different types of synaptic sites would be necessary to satisfy this condition.

If enough of the synaptic sites in a particular receptive area receive input at about the same time, then the threshold of the receptive area is exceeded and a "local (spike) response" is initiated. For simplicity, let us assume that a local response initiated in any receptive area is sufficient to fire the cell body, initiating a spike in the axon. Also, let us assume that each receptive area has a very high threshold, so high that to have the probability of a local response be above 0.99, virtually all of the synaptic sites in the receptive area must receive input at about the same time. Such a neuron computes a disjunction of conjunctions. If such neurons exist, simplicity demands that they be discussed as two-stage, multiple-threshold devices, not as single-stage, single-threshold devices with non-linear combination of inputs. Of course, we are not assuming that all neurons are disjunctive-conjunctive neurons; only that a large subset of them are.

There is anatomical and physiological evidence for such multi-threshold neurons in the cerebral cortex. The anatomical evidence is merely that the richly branching dendritic trees of many cortical neurons are ideally suited to the computation of logical functions such as a disjunction of conjunctions or even more complex logical functions. If all synaptic sites contributed by passive electronic conduction to a single weighted sum which was compared to

a threshold, it is estimated that many dendritic synapses would contribute nothing at all (Eccles, 1966). Since it is unlikely that these dendritic synapses serve no function in the activation of a cell, the anatomical evidence is against the single threshold theory. There is also physiological evidence for local (spike) responses in dendrites of the pyramidal neurons in the archicortex (Spencer and Kandel, 1961; Anderson, 1966) and in the neocortex (Purpura and Shofer, 1964). There is, however, some argument concerning the frequency of such local responses in the neocortical pyramidal cells (Eccles, 1966). In addition, Diamond (1968) has demonstrated in Mauthner neurons that a single dendrite can be inhibited with little or no effect on other dendrites. Thus, the neural evidence does not support the general validity of the single threshold theory and indicates that some kinds of neurons could compute disjunctions of conjunctions.

What is the Critical Neural Event? It should be pointed out that there is an unresolved question about neural coding that may or may not be successfully finessed by an abstract version of the above formulation. The abstract version is to substitute "goes into state 1" for "initiates a spike" either in a receptive area or a cell body and axon. Whether state 1 consists of a single spike (rather unlikely) or some number of spikes initiated in some short time period (much more likely) will not be specified by the theory.

Conceivably state 1 could be a particular pattern of spikes, as opposed to sheer frequency of spikes, but there is little evidence to support this possibility. In particular, motor units depend on frequency, not specific patterns, and so it would seem to be simpler to use frequency as the code throughout the nervous system to facilitate mapping from one subsystem onto another and ultimately onto motor units. Nevertheless, the abstract formulation of disjunctive-conjunctive neurons can be noncommittal on the issue of pattern *vs.* frequency of spikes, though detailed physiological testing of the theory requires specifying the theory in this respect.

Note, however, that the abstract formulation specifies only two states of functional significance for a receptive area or a cell body, above threshold and below threshold. A more detailed theory would, undoubtedly, specify many states both above and below threshold. Hopefully, the two state version is not a bad first approximation. In any event, throughout the paper, when a neuron is said to stand for (represent) a pattern or concept, it means that the neuron is in state 1, if and only if that pattern or concept is being perceived or thought of.

Concept Neurons. Assuming that the two-state representation is not a bad approximation, we have seen how a neuron could encode a disjunction of con-

junctions of attributes. Each conjunction of attribute-neurons must synapse on approximately all of the synaptic sites in one or more receptive areas, dividing up the sites in each area approximately equally among the attributes. The disjunctive aspect of the concept is handled by the different receptive areas, each of which is sufficient to fire the cell body and axon. The conjunctive aspect of the concept is handled by the different synaptic sites in each receptive area, almost all of which must receive input at about the same time in order to initiate a spike in the receptive area and therefore in the cell body.

Thus, we are postulating the existence of one, or more likely several (redundant), concept neurons standing for every concept an individual has, with each of the events that signal a concept causing one or more spikes in one or more receptive areas of the neuron(s) standing for that concept. So the bark of a dog, the sight of a dog from many different perspectives at many different distances, the sound of the word "dog", the sight of the word "dog", and many other events more indirectly associated to the concept dog are all assumed to activate the neuron(s) standing for dog by virtue of their activating one or more receptive areas of the "dog-neuron(s)".

All this is certainly not done in one step from the peripheral sensory neurons to the learned concept neurons. The recent advances in our knowledge of the first several levels of the cat visual system (e.g., Kuffler, 1953; Hubel and Wiesel 1959, 1961, 1962, 1965) make it clear that the initial attribute analysis of a stimulus is transformed into several other attribute analyses, which presumably allow characterization of a particular sensory event as a conjunction of a smaller number of attribute-representatives. For example, a particular triangle can be represented by a conjunction of a large number of retinal ganglion cells (dot-representatives) or by only 3 simple cortical cells (line-representatives). There appear to be several genetically-specified attribute analyses in the visual system, and undoubtedly in every other modality as well.

Sitting on top of the genetically-specified hierarchy of analyses in each modality, there are, according to the present theory, several levels of learned concept neurons, some of which may be specific to a single sensory modality, but others of which receive input from many or all modalities.

For example, there may be auditory concept neurons standing for distinctive features of allophones or phonemes in speech recognition. These may be separate from the feature-representatives used in articulation, or they may be the same feature-representatives. Conjunctions of feature-representatives might project onto phoneme or allophone representatives. Alternatively, there may be no special neurons on either the auditory or the articulatory side that stand for distinctive features. Instead, it is conceivable that concept neurons standing for allophones or phonemes could handle (within each con-

junction of cues signaling an allophone or phoneme) the much more complex encoding (attribute analysis) of speech that is given by a tonotopically-organized level of the auditory system.

In any event, conjunctions of phoneme neurons or, more likely, context-sensitive allophone neurons (see Wickelgren, 1969) project onto word neurons which project onto concept neurons, or perhaps context-sensitive allophone neurons project directly onto concept neurons. Concept neurons, of course, are activated by many events other than the sound of the word that stands for them. Thus, reading a story is very much like having it happen to you.

How much of this hierarchy in the speech system is genetically specified and how much is specified by learning cannot be said at the present time. However, it is clear that there are too many possible words or concepts that human beings *can* possess for there to be genetic specification of neurons standing for each possible word or concept. However, the number of words and concepts *actually* possessed by any given individual does not appear to exceed 10^{10}, which is a rough estimate of the number of neurons in the human cortex (see Cragg, 1967, for the most up-to-date estimate). Thus, if some reasonable theory could be devised for learned specification of genetically unspecified neurons, then the notion of single neurons standing for concepts would have much greater plausibility. This will be attempted in the next section.

Learned Specification. Granted that a single neuron could encode a disjunction of conjunctions of attributes in the manner just described, how could such precise specification come about by learning? The specification must be by learning, since there are too many possible concepts for genetic specification. This section considers the problem of learned specification of concept neurons, coming to the conclusion that a solution to this problem is only slightly more remarkable than a solution to the problem of genetic specification, which we know must have a solution.

Let us imagine that genetically-specified attribute neurons send their multiply-branched axons into the region(s) of the (association?) cortex where the genetically-unspecified (free) concept neurons are located. Whenever a set of attribute neurons is activated in contiguity, their axons set up an "electrochemical gradient" which attracts the axons toward each other. When they get very close to one another, they do not synapse with each other (axon to cell body or dendrite), as in the more familiar neural contiguity conditioning assumptions, e.g., Kappers neurobiotaxis (Kappers, Huber and Crosby, 1936) or Hebb (1949). Rather, they synapse extremely near each other on a single receptive area of the nearest free neuron, specifying that neuron to stand for the conjunction of attributes. All the sites in the receptive area are divided up

approximately equally by the attribute neurons. Thus, in a conjunction of few attributes, each attribute neuron makes more synaptic contacts in the receptive area than in a conjunction of many attributes. This handles the conjunctive aspect of concepts, and is no more implausible than what must happen in development to achieve genetic specification of neurons.

The disjunctive aspect of concept neuron specification may or may not require additional properties of the nervous system. Somehow, each of the conjunctions of attributes that signals a given concept must synapse in different receptive areas of the same neuron. In order to suggest a solution to this problem, we must first specify the circumstances under which human beings learn to consider two events to be signaling the same concept. The most obvious such circumstance is when the two events are frequently experienced in close contiguity. Undoubtedly, there are other circumstances under which our complex intellects deduce that two events signal a common concept, and probably, two events have to satisfy other requirements in addition to contiguity in order to be considered as signaling a common concept. However, let us limit consideration to the effects of event contiguity, assuming that the events meet whatever other requirements might be necessary for being lumped into a common concept.

So we are considering how two event-representatives that are not now connected to the same concept neuron come to be connected to different receptive areas of the same concept neuron through contiguity of activation. If we can handle pairs of events, we can obviously handle any number of events, up to the capacity of the concept neuron. There are two distinguishable cases. First, the concept neuron may already be specified by the conjunction of attributes corresponding to one event, and the question is simply, how does the other conjunction of attributes make contact with another receptive area of the same concept neuron? Second, the concept neuron may be completely unspecified, and both conjunctions of attributes are presented, in contiguity, for the first time. Then the question is, how do the two sets of attribute neurons migrate to the same concept neuron so as to connect to separate receptive areas of that concept neuron?

Solution of the first case is easy. It requires no assumptions beyond those necessary for solving the conjunctive aspect of learned specification of concept neurons. The axons of the second set of attribute neurons are attracted to the terminals of the first set, since both sets are activated in contiguity. But when they arrive near the terminals, all the synaptic sites of one receptive area are taken, so they make contact with the synaptic sites of a different receptive area on the same concept neuron.

Solution of the second case is more difficult. We can have the axons of both

sets of attribute neurons growing to the same concept neuron by the mechanism already assumed, but how do the two sets remain differentiated when they make synaptic contact with the concept neuron? Would not this look like a bigger conjunction of attributes to the concept neuron instead of being two conjunctions? The answer, I suggest, is in whether the contiguity of activation of the two sets of attribute neurons is simultaneous or successive. If the contiguity is simultaneous, then there is no basis for separating the set into two subsets. However, if there really are two sets, activated in close, but successive, contiguity, then there is a basis for separation, and it seems necessary to propose some manner in which the nervous system could achieve separation in contacts on the concept neuron.

Let us consider some examples of rapid successive activation of sets of attribute neurons. First, there are the different perspectives of the same object, which are never simultaneously activating their different sets of attribute neurons, but often do so in rapid succession. Second, there are the cues from different sensory modalities for the same object, e.g., the bark and the sight of a dog. These could be activating their attribute neurons simultaneously, but we shall assume that usually or always there is a selective attention mechanism that gates activation of the attribute neurons in blocks corresponding to modalities and often in blocks within sensory modalities (verbal-non-verbal, frequency bands, location in auditory or visual space, etc.). With the selective attention mechanism operating within modalities, one could have a basis for separating parts of the same object, each part alone being sufficient for identification of the entire object on a later occasion. Thus, a third example might be the facial profile vs. the body of a dog, even when both are simultaneously present.

Given the rapid successive activation of two sets of attribute neurons, how do we get them to grow to the same concept neuron, but end on different receptive areas? I suppose there are many possibilities, but one approach is to distinguish between two stages of the process; (a) the growing to the vicinity of a single concept neuron, which is assumed to work as before, with simultaneous or successive contiguity being approximately equivalent and (b) the making of synaptic contacts in a single receptive area, for which we assume strict simultaneity of activation to be critical.

Conceivably all concept learning by contiguity takes place by pairing one new event signaling the concept with one previously learned event signaling the concept (case 1). If so, the requirements for learned specification of concept neurons seem to be no more remarkable than the requirements for genetic specification, namely the growth of neurons in directions determined by electrochemical gradients. Case 2 concept learning would be a little more difficult to

achieve than case 1 concept learning, probably requiring a selective attention mechanism.

Consolidation. Up to now we have completely ignored the question of how rapidly these growth processes take place. Pretty obviously they require much longer than the period of time of event pairing necessary to produce learning. Thus, we must assume some sort of consolidation time, during which the information about contiguity of activation is preserved and used to direct the growth process. Furthermore, if events signaling the same concept are not always learned one at a time, we must postulate that the consolidation mechanism preserves the information regarding the difference between simultaneous and successive contiguity.

What sort of consolidation mechanism might achieve these objectives? One possibility is as follows. Imagine a consolidation system consisting of thousands of subsystems. Each subsystem is connected to all attribute neurons and maybe to all concept neurons too. When a set of attribute neurons (and/or concept neurons) is activated simultaneously, the connections from one subsystem of the consolidation system to the set of activated attribute neurons are facilitated.

Subsystems of the consolidation system are composed of one or more pacemaker neurons connected in an excitatory manner to other neurons in the same subsystem, but in an inhibitory manner to neurons in other subsystems. Thus, only one subsystem can be active at a time. Subsystems fatigue and then transfer control to another subsystem. The subsystems are conditionable to each other, so the one to which control is transferred is the subsystem that was activated after the just-active subsystem on the previous occasion.

When a new set of attributes is to be consolidated, a subsystem that is not currently in use is "hooked-up" to the new set by facilitation of the synapses from the subsystem to that set of attribute neurons.

When no new learning is taking place, the consolidation system consolidates old learning by activating subsystems in the order they were used in the recent past. The facilitated synapses between a subsystem and a set of attribute-neurons remain facilitated for only a limited period of time (hours or days) and thereafter are available for use in new learning.

The effect of such a consolidation system is to use the time when the nervous system is not actively engaged in new learning to consolidate old learning. The mechanism of the consolidation involves a purely central reinstigation of the same pattern of simultaneous and successive activation of sets of neurons that was produced by the original learning experience. Thus, one could have periodic reinstigation of the electrochemical gradients necessary for both of the

postulated growth processes, axonal attraction and synaptic contact in a single receptive area.

The above theory of consolidation specifies a close dependence of the consolidation of the long-term memory trace upon certain passive and active short-term memory traces. While this theory is tenable, it is equally plausible that the long-term memory trace is formed by a process completely independent of any active or passive short-term memory traces (Albert, 1966a). One possible theory of this type of consolidation is as follows. Imagine that the electrochemical gradient set up by simultaneous activation of a set of neurons establishes some sort of intermediate-term (chemical) memory in the attribute neurons simultaneously activated, or conceivably, the surrounding glia. This memory causes the neurons simultaneously activated to grow toward some common point, like their center of gravity or a concept neuron activated simultaneously with the set of attribute neurons.

One learning experience must often be insufficient for specification of a concept neuron by the conjunction of attributes characteristic of the event. In such cases, it is assumed that the axonal attraction process results in branches of the axonal trees of each attribute neuron in the conjunction being brought closer together, but not necessarily close enough together to make contact with a receptive area of a single concept neuron. Further experience with the same set of attributes, or a subset, the *relevant* attributes, of the original set, would eventually result in the relevant subset impinging upon a receptive area of a free neuron, specifying it to respond to that subset of relevant attributes. Thus, the initial learning experiences of young organisms may have no effect on performance, but may make later experience more likely to result in learning that will affect performance.

Although each of the two particular theories of consolidation expressed in this section is speculative in the extreme, their plausibility in no way affects the plausibility of the theory of learned specification of concept neurons. On the contrary, since there is evidence that some aspects of the consolidation process (but not all) last several hours or days (Albert 1966a and b; Deutsch, Hamburg and Dahl, 1966; Hamburg, 1967; Russell, 1959; Weiskrantz, 1966), an axonal-synaptic growth process is made slightly more plausible by its presumed need for consolidation times of that order of magnitude.

Forgetting. With time or interference, I would assume that unused synaptic connections degenerate, followed by a gradual degeneration (withdrawal) of the ends of axonal branches that had converged on a single receptive area of a concept neuron. In this way, earlier connections that are later of no use would not remain to complicate retrieval and make new learning needlessly more

difficult. Although forgetting is often implicitly considered to be undesirable, a stronger case can probably be made for its advantages in reducing interference in both acquisition and retrieval than for its disadvantages. The fact that we do forget is another reason for thinking that it is functionally advantageous to forget at about the rate that we do.

The present theory also explains why forgetting often appears to be complete by a performance measure, such as recall or recognition, but relearning is much faster than original learning. The explanation is that, in these cases, forgetting has proceeded to the point where most of the synaptic connections have degenerated, but the axonal branches are still relatively near each other, facilitating relearning.

Associative Structure. In addition to conjunctions of attribute neurons becoming connected to a concept neuron, other concept neurons or conjunctions of concept neurons can become connected to any given concept neuron. Furthermore, conjunctions of "concrete" concept neurons (those specified by disjunctions of conjunctions of attribute neurons) can come to specify more "abstract" concept neurons that may or may not have direct input from genetically specified attribute neurons. Thus, one can have hierarchical conceptual organization and various kinds of associative structures containing loops (closed chains of associations), convergence and divergence of associations, etc.

Fallacious Arguments Against Connectionism

Several arguments seem, almost invariably, to be raised by people upon hearing of any specific-neuron, connectionist theory. First, it is said that we lose several thousand neurons a day (Brody, 1955); why do we not frequently find ourselves unable to recognize or think of certain very familiar concepts? (Of course, we often fail to recognize or think of "less familiar" concepts.) Second, people who suffer brain injuries that certainly have destroyed large numbers of neurons occasionally appear to have little or no cognitive deficit. Third, Lashley (1929) and Lashley and Wiley (1933) performed a number of lesions in rats that appeared to show no specific deficits due to loss of a particular region of the brain, but rather a general decrease in maze learning or retention with increasing amount of cortical tissue removed.

These arguments against specific neuron encoding of concepts are all fallacious for various reasons. In the first place, common sense tells one that loss of neurons (whether gradually each day in scattered locations or suddenly in a single location) must have some deleterious effect on mental functioning. When such effects are not found, the most likely possibility is that the tests are too

crude to show the defects. Another possibility, which fits in perfectly with the present theory of learned specification of concept neurons, is that the losses may sometimes be primarily in future intellectual potential or the duration of competent intellectual functioning, with any immediate losses in important knowledge being quickly made up by specification of previously unspecified neurons. In fact, the instances of recovery of function and the resiliency of the human brain to gradual loss of neurons until age 60–80 constitute one source of support for the learned specification aspect of the present theory.

In the second place, as many people have pointed out, complex tasks, like maze learning, that can be learned using many different kinds of cues are hardly suitable for determining whether the encoding of certain kinds of concepts is restricted to a certain part of the cortex. What sort of task is simple is, of course, a difficult question to answer.

In the third place, if a substantial portion of the cerebral cortex consists of neurons that are specified by learning rather than by genetic processes, then there is reason to suspect greater individual variability in the types of concepts represented and in the location of similar concepts. The absence of precise localization of concepts in parts of the cortex, if it could be established, would not be evidence against the present theory. On the contrary, depending on just what the evidence was, it could be strongly supportive of learned specification of concept neurons.

In the fourth place, postulating the existence of a single neuron that represents a concept does not mean postulating that there is *only* one such neuron representing each concept. If replication is desirable for combating the effects of gradual neural death, accidental destruction of one region of the brain, or for facilitating the formation of connections to all other concept neurons in the cortex, then many or all concepts may be represented by several neurons in distant locations (including being in different hemispheres). Furthermore, human concepts are redundant, e.g., dogs are pets.

In addition, although it is outside the scope of the present paper, if one makes the more accurate assumption that neurons have many states (rather than just two), then some replication of representation is possible through generalization gradients. That is, if a concept neuron is activated submaximally by an input that maximally activates a neuron encoding a "similar" concept, then this alone probably provides some replication of concept representation. How useful this would be is difficult to say without a quantitative theory.

Are Neurons Unreliable? One of the concerns of research on neural networks has been to obtain reliable behavior from nets of unreliable component neurons (e.g., McCulloch, 1959). Absolutely no provision for this has been made in the

present theory, because I do not believe there is substantial evidence that neurons are "very" unreliable, that is, orders of magnitude more unreliable than behavior, which is also unreliable. Obviously, by assuming a large safety factor in the ability of a local response in a receptive area to initiate a spike in the cell body, we can avoid unreliability in the disjunctive aspect of concept neurons. The conjunctive aspect could have essentially the same unreliability problem posed for single-threshold neurons, which is the assertion that the threshold can be greatly affected by various drugs (e.g., alcohol) and temperature changes that are alleged not to affect intellectual functioning very much. How much is "not very much?" Certainly there are changes in intellectual functioning, but our knowledge of what they are is limited. Are these behavioral changes less than we would expect from the neural changes? Well, we do not know much about the neural changes either, including possible compensatory changes, such as changes in transmitter release compensating for changes in postsynaptic threshold.

Temporal and Spatial Summation. However, there is a more general problem closely related to the problem of threshold change that must be solved to make the present theory plausible. That is the problem of trade-off between temporal and spatial summation of inputs within a receptive area. In order to compute a conjunction, it has been assumed that a receptive area must have a very high threshold, requiring nearly simultaneous input at virtually all sites in order to be exceeded. This ignores temporal summation almost completely. Would not the threshold also be exceeded by more rapid input on a fraction of the sites in a receptive area? If so, the receptive area would not always be computing a conjunction of its inputs. Since we wish the receptive area to compute a conjunction, this is a serious problem.

Of course, we could completely discard the notion that even a receptive area of a neuron computes a single sum of its inputs to be compared to a threshold. Instead, we could just assert that all sites must have changed state in order for a local response to be initiated. Since this flies in the face of current notions of spike generation, it is preferable to specify the temporal summation properties of a single threshold theory of receptive areas that would make the receptive area compute a conjunction of its inputs.

The obvious answer is to assume that the input from a single presynaptic spike is sufficient to produce the maximum contribution (that a single synaptic site is capable of producing) to the excitatory postsynaptic potential (epsp) of a receptive area. Further input at the same site can maintain the epsp contribution of that site at its maximum for a longer period of time, but it cannot increase the epsp contribution beyond the level produced by a single input.

Now if the threshold of the receptive area is equivalent to maximum epsp contributions from nearly all the sites or nearly maximum contributions from all the sites, then the receptive area will compute a conjunction.

The only remaining question concerns how long a single input will maintain the epsp contribution of a site near its maximum. This could be set at whatever time period seems most desirable, and might be different for different neurons. Conceivably, it could be a dynamically adjustable parameter, though I have no idea of the mechanism.

This work was supported primarily by grant, MH 08890-04, from the National Institute of Mental Health, U.S. Public Health Service.

LITERATURE

Albert, D. J. 1966a. "The Effects of Polarizing Currents on the Consolidation of Learning." *Neuropsychologia*, **4**, 65–77.

———. 1966b. "Memory in Mammals: Evidence for a System Involving Nuclear Ribonucleic Acid." *Ibid.*, **4**, 79–92.

Andersen, P. O. 1966. "Correlation of Structural Design with Function in the Archicortex." In *Brain and Conscious Experience*, pp. 59–79. J. C. Eccles, Ed. New York: Springer-Verlag.

Baddeley, A. D. and H. C. A. Dale. 1966. "The Effect of Semantic Similarity on Retroactive Interference in Long- and Short-Term Memory." *J. Verb. Learn. Verb. Behav.*, **5**, 417–420.

Brody, H. 1955. "Organization of the Cerebral Cortex. III. A Study of Aging in the Human Cerebral Cortex." *J. Comp. Neurol.*, **102**, 511–556.

Conrad, R. 1962. "An Association Between Memory Errors and Errors Due to Acoustic Masking of Speech." *Nature*, **193**, 1314–1315.

———. 1964. "Acoustic Confusions in Immediate Memory." *Brit. J. Psychol.*, **55**, 75–84.

Cragg, B. G. 1967. "The Density of Synapses and Neurones in the Motor and Visual Areas of the Cerebral Cortex." *J. Anat.*, **101**, 639–654.

Deutsch, J. A., M. D. Hamburg and H. Dahl. 1966. "Anticholinesterase-Induced Amnesia and Its Temporal Aspects." *Science*, **151**, 221–223.

Diamond, J. 1968. "The Activation and Distribution of GABA and L-Glutamate Receptors on Goldfish Mauthner Neurons: An Analysis of Dendritic Remote Inhibition." *J. Physiol.*, **194**, 669–723.

Eccles, J. C. 1966. "Cerebral Synaptic Mechanisms." In *Brain and Conscious Experience*. J. C. Eccles, Ed. New York: Springer-Verlag.

Hamburg, M. D. 1967. "Retrograde Amnesia Produced by Intraperitoneal Injection of Physostigmine." *Science*, **156**, 973–974.

Hebb, D. O. 1949. *The Organization of Behavior*. New York: Wiley.

Hubel, D. H. and Wiesel, T. N. 1959. "Receptive Fields of Single Neurons in the Cat's Striate Cortex." *J. Physiol.*, **148**, 574–591.

——— ———. 1961. "Integrative Action in the Cat's Lateral Geniculate Body." *Ibid.*, **155**, 385–398.

Hubel, D. H. and Wiesel, T. N. 1962. "Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex." *Ibid.*, **160**, 106–154.

——— ———. 1963. "Receptive Fields of Cells in Striate Cortex of Very Young, Visually Inexperienced Kittens." *J. Neurophysiol.*, **26**, 994–1002.

——— ———. 1965. "Receptive Fields and Functional Architecture in Two Nonstriate Visual Areas (18 and 19) of the Cat." *Ibid.*, **28**, 229–289.

Kappers, C. U. A., G. C. Huber and E. C. Crosby. 1936. *The Comparative Anatomy of the Nervous System of Vertebrates, Including Man.* Vol. I. New York: Macmillan.

Kuffler, S. W. 1953. "Discharge Patterns and Functional Organization of Mammalian Retina." *J. Neurophysiol.*, **16**, 37–68.

Lashley, K. S. 1929. *Brain Mechanisms and Intelligence.* Chicago: University of Chicago Press.

——— and L. E. Wiley. 1933. "Studies of Cerebral Function in Learning. IX. Mass Action in Relation to the Number of Elements in the Problem to be Learned." *J. Comp. Neurol.*, **57**, 3–55.

McCulloch, W. S. 1959. "Agathe Tyche of Nervous Nets—the Lucky Reckoners." *Mechanisation of Thought Process*, **2**, 612–625. London: Her Majesty's Stationery Office.

——— and W. Pitts. 1943. "A Logical Calculus of the Ideas Immanent in Nervous Activity." *Bull. Math. Biophysics*, **5**, 115–133.

McGeoch, J. A. and W. T. McDonald. 1931. "Meaningful Relation and Retroactive Inhibition." *Amer. J. Psychol.*, **43**, 579–588.

Purpura, D. P. and R. J. Shofer. 1964. "Cortical Intracellular Potentials During Augmenting and Recruiting Responses. I. Effects of Injected Hyperpolarizing Currents on Evoked Membrane Potential Changes." *J. Neurophysiol.*, **27**, 117–132.

Rall, W. 1964. "Theoretical Significance of Dendritic Trees for Neuronal Input-Output Relations." In *Neural Theory and Modeling*, pp. 73–97. R. F. Reiss, Ed. Stanford, California: Stanford University Press.

———. 1967. "Distinguishing Theoretical Synaptic Potentials Computed for Different Soma-Dendritic Distributions of Synaptic Input." *J. Neurophysiol.*, **30**, 1138–1168.

Rashevsky, N. 1938. *Mathematical Biophysics.* First edition, 1938. Third edition, New York: Dover, 1960.

Russell, W. R. 1959. *Brain, Memory and Learning.* New York: Oxford University Press.

Spencer, W. A. and E. R. Kandel. 1961. "Electrophysiology of Hippocampal Neurons. IV. Fast Prepotentials." *J. Neurophysiol.*, **24**, 273–285.

Underwood, B. J. and D. Goad. 1951. "Studies of Distributed Practice: I. The Influence of Intra-List Similarity in Serial Learning." *J. Exp. Psychol.*, **42**, 125–134.

Weiskrantz, L. 1966. "Experimental Studies of Amnesia." In *Amnesia.* C. W. M. Whitty and O. L. Zangwell, Eds. London: Butterworths.

Wickelgren, W. A. 1965a. "Acoustic Similarity and Retroactive Interference in Short-Term Memory." *J. Verb. Learn. Verb. Behav.*, **4**, 53–61.

———. 1965b. "Acoustic Similarity and Intrusion Errors in Short-Term Memory." *J. Exp. Psychol.*, **70**, 102–108.

———. 1965c. "Distinctive Features and Errors in Short-Term Memory for English Vowels." *J. Acoust. Soc. Amer.*, **38**, 583–588.

———. 1966. "Distinctive Features and Errors in Short-Term Memory for English Consonants." *Ibid.*, **39**, 388–398.

———. 1969. "Context-Sensitive Coding, Associative [Memory and Serial Order in (Speech) Behavior." *Psychol. Rev.* (in press).

RECEIVED 7-24-68