A Theory for Cerebral Neocortex
Author(s): D. Marr
Source: *Proceedings of the Royal Society of London. Series B, Biological Sciences*, Vol. 176, No. 1043 (Nov. 3, 1970), pp. 161-234
Published by: The Royal Society
Stable URL: http://www.jstor.org/stable/76043
Accessed: 06/06/2014 14:41

Proc. Roy. Soc. Lond. B. **176**, 161–234 (1970)
Printed in Great Britain

A theory for cerebral neocortex

By D. Marr
Trinity College, Cambridge

(*Communicated by G. S. Brindley, F.R.S.—Received 2 March* 1970)

Contents

It is proposed that the learning of many tasks by the cerebrum is based on using a very few fundamental techniques for organizing information. It is argued that this is made possible by the prevalence in the world of a particular kind of redundancy, which is characterized by a 'Fundamental Hypothesis'.

This hypothesis is used to found a theory of the basic operations which, it is proposed, are carried out by the cerebral neocortex. They involve the use of past experience to form so-called 'classificatory units' with which to interpret subsequent experience. Such classificatory units are imagined to be created whenever either something occurs frequently in the brain's experience, or enough redundancy appears in the form of clusters of slightly differing inputs.

A (non-Bayesian) information theoretic account is given of the diagnosis of an input as an instance of an existing classificatory unit, and of the interpretation as such of an incompletely specified input. Neural models are devised to implement the two operations of diagnosis and interpretation, and it is found that the performance of the second is an automatic consequence of the model's ability to perform the first.

The discovery and formation of new classificatory units is discussed within the context of these neural models. It is shown how a climbing fibre input (of the kind described by Cajal) to the correct cell can cause that cell to perform a mountain-climbing operation in an underlying probability space, that will lead it to respond to a class of events for which it is appropriate to code. This is called the 'spatial recognizer effect'.

The structure of the cerebral neocortex is reviewed in the light of the model which the theory establishes. It is found that many elements in the cortex have a natural identification with elements in the model. This enables many predictions, with specified degrees of firmness, to be made concerning the connexions and synapses of the following cortical cells and fibres: Martinotti cells; cerebral granule cells; pyramidal cells of layers III, V and II; short axon cells of all layers, especially I, IV and VI; cerebral climbing fibres and those cells of the cortex which give rise to them; cerebral basket cells; fusiform cells of layers VI and VII.

It is shown that if rather little information about the classificatory units to be formed has been coded genetically, it may be necessary to use a technique called codon formation to organize structure in a suitable way to represent a new unit. It is shown that under certain conditions, it is necessary to carry out a part of this organization during sleep. A prediction is made about the effect of sleep on learning of a certain kind.

§ 0. Introduction

0.1. *The form of a neurophysiological theory*

The mammalian cerebral neocortex can learn to perform a wide variety of tasks, yet its structure is strikingly uniform (Cajal 1911). It is natural to wonder whether this uniformity reflects the use of rather few underlying methods of organizing information. The present paper rests on the belief that this is so, and describes a kind of analysis which is capable of serving many aspects of the brain's function. The theory is necessarily general, but it in principle allows the exact form of the analysis for any particular cerebral task to be computed.

There is an analogy between the shape of the general theory set out here, and that of a recent theory of cerebellar cortex (Marr 1969). The essence of the latter theory was a principle, that motor sequences are driven by learned contexts, which was clearly applicable to the kind of function with which the cerebellum was thought to be associated. The key ideas concerned the way information was stored, and the way stored information could be used; but the theory did not explicitly demonstrate how any particular motor action was learned. For this, it would be necessary to have a much fuller understanding of the nature of the elemental movements for which the Purkinje cells actually code, and of the information present in the relevant mossy fibres. The theory was however useful, because it postulated the existence of a 'fundamental operation' of the cerebellar cortex, and offered a candidate for it. The present theory is once removed from the description of any task the cerebrum might perform, in the same way as was the cerebellar theory from the description of any particular motor action.

Something of this kind is probably an inevitable feature of the theory of any interesting learning machine, but in the particular case of the cerebral cortex, it is likely there exists a second, more concrete analogy between its working, and that of the cerebellar cortex. The evidence for this is the analogy between the structures of the two types of cortex. The cerebral cortex is of course irregular and very complicated, but there do exist similarities between it and the cerebellar cortex: the fundamental cerebellar components—the granule cells, Purkinje cells, parallel fibres, climbing fibres, basket cells and so on—have recognizable counterparts in the cerebral cortex. In view of the great power the codon representation possesses for the economical storage of information (Marr 1969), it cannot be that this analogy is accidental. There must be a deeper correspondence.

0.2. *The nature of the present general theory*

It was the suspicion that there may exist deep reasons for these similarities that formed the starting point of the present enquiry. The motivation for the development of the theory was provided by two intuitions. The first was that in the generalization of the basic cerebellar circuit, the analogue of the Purkinje cell (called an *output cell*) need not have a fixed 'meaning'. In the cerebellum, each Purkinje cell probably has predetermined 'meanings', in that the responses its outputs can

evoke are likely to be determined by embryological and early post-natal development. In a more general application of this kind of model, it is clear that what the output cell 'means' might be free to be determined by some aspect of the structure of the information for which the system is being used.

The second intuition was that the codon representation, in the kind of model applicable to the cerebellar cortex, may in fact be capable of doing more than the simple memorizing task to which it can obviously be applied (Blomfield & Marr 1970). This feeling was tied to the idea that the recognition of a learned input ought properly to be viewed as a process of diagnosing whether the current input belonged to the class of learned inputs. This immediately suggests that the behaviour of an output cell should not be an all-or-none affair, but should convey a measure of how *certain* is the outcome of the diagnostic process. This has the attraction that it could ultimately correspond to how 'like' a tree is the object at which one is currently looking.

These two ideas were bound by the constraint that more or less whatever theory was set up, it had to be grounded in information theory; or if not firm reasons why this is undesirable must be given. It was evident from the start that no very orthodox information theoretic approach would be of any use; but the general ideas behind the formulation of an information measure are so powerful that it would have been surprising had they turned out to be totally irrelevant.

The result of these ideas was a general theory which divides neatly into two parts. The first, with which this paper is concerned, describes the formation and operation of a language of so-called *classificatory units* by means of which the sensory input can eventually be usefully interpreted (§ 1). The formation of a classificatory unit is imagined to occur roughly whenever enough related inputs happen to make it worth forming a special description for them. The main results are the information theoretic theorems of § 2 on the diagnosis and interpretation of an input within a class, and the theory of § 5 for class formation. The power of these results is that they lead to specific neural models, and to operations in those models, through which a preliminary interpretation of the histology of cerebral cortex can usefully be made.

The first part of the theory may therefore be described as a model for concept formation and recognition, where concepts are 'classificatory units'. It argues that there exists a basic information-handling scheme which is applied by the cerebral cortex to a wide range of different kinds of information—that there exists a 'way' in which the cerebral cortex 'works'. This scheme has a wide application, subject to reservations about the need in certain circumstances for special coding devices to cope with particular forms of redundancy. But in principle, it can be applied to anything from the recognition of a tree to the recognition of the necessity to take a particular course of action.

The theorems of § 2 provide a complete analysis of the problem of interpreting an input within a particular class, but the ideas of § 5 provide only a partial analysis of the formation of the classes themselves. This problem cannot be dealt with using only

the hardware developed in this paper; and its solution requires the results of the second part of the general theory.

The second part of the theory embodies a second pair of ideas. One of these also arises from the cerebellar theory, where it was seen that a codon representation is extremely successful at straight memorizing tasks (Brindley 1969; Marr 1969). The other is the everyday concept of an associative memory. The cerebellar theory is a kind of associative memory theory, and it is not difficult to extend the idea of the codon representation to the case of a general associative memory. This is developed in the theory of Simple Memory (Marr 1971). Once this has been done, it is possible to see how current descriptions of the environment can be stored, and recalled by addressing them with small parts of such inputs. This is the facility needed to complete the theory of the formation of classificatory units. It is, however, only a small part of the use to which such a device can be put: almost the entire theory of the analysis of temporally extended events, and of the execution *ab initio* of a sequence of movements, rests upon such a mechanism. Though simple, it is important (and long) enough to warrant a separate development, and is therefore expounded elsewhere, together with the theory of archicortex to which it gives rise.

0.3. *Outlines of the present theory*

This paper starts with a discussion of the kind of analysis of sensory information which the brain must perform. The discussion has two main strands: the structure of the relationships which appear in the afferent information; and the usefulness to the organism of discovering them. These two ideas are combined by the 'Fundamental Hypothesis' of §1.6 which asserts the existence and prevalance in the world of a particular kind of relationship. This forms an explicit basis for the subsequent theoretical development of classificatory units as a way of exploiting these relationships. The fundamental hypothesis is a statement about the world, and asserts roughly speaking, that the world tends to be redundant in a particular way. The subsequent theory is based, roughly, on the assumption that the brain runs on this redundancy.

The second section contains the fundamental theorems about the diagnosis and interpretation of events within a class. It assumes that the classes have been set up, and studies the way in which they allow subsequent incoming information to be interpreted. These theorems receive their neural implementation in the model of figure 8.

The rest of the paper is closely tied to the examination of specific neural models. After the technical statistics of §3, the main section §4 on the fundamental neural models appears. This discusses the structures necessary for the implementation of the basic theorems, and derives explicitly those models which for various reasons seem preferable to any others. The first main result of the paper consists in the demonstration that the two theorems of §2 correspond to closely related operations in the basic neural model.

The second main result concerns the operations involved in the discovery of new

classificatory units. It shows how a climbing fibre enables a cortical pyramidal cell to discover a cluster in the space of events which that cell receives. This result, together with the previous ones which show how classificatory units work when represented, completes the main argument of the paper.

Finally, in §6, the available knowledge of the structure of the cerebral cortex is briefly reviewed, and parts of it interpreted within the models of §4. This section is incomplete, both because of a lack of information, and because Simple Memory theory allows the interpretation of other components; but it was thought better at this stage to include a brief review than to say nothing. Far too little is known about the structure of the cerebral cortex.

0.4. *Definitions and notation*

0.4.1. *Time*, t, is discrete, and runs through the non-negative integers ($t = 0, 1, 2, \ldots$). t scarcely appears itself in the paper, but most of the objects with which the theory deals are essentially functions of t.

0.4.2. An *input fibre*, or *fibre*, $a_i(t)$, is a function of time t which has the value 0 or 1, for each i, $1 \leqslant i \leqslant N$. $a_i(t) = 1$ will have the informal meaning that the fibre a_i carries a signal, or 'fires' at time t. A signal is usually thought to correspond to a burst of impulses in a real axon. The set of all input fibres is denoted by A, and the set of all subsets of A by $\mathfrak{A}$.

0.4.3. An *input event*, or *event*, on A assigns to each fibre in A the value 0 or 1. Events are usually denoted by letters like E, F, and the value which the event E assigns to the fibre a_i is written $E(a_i)$, and equals 0 or 1 ($1 \leqslant i \leqslant N$). It is convenient to allow the following slight abuse of notation: E can also stand for the set of a_i which have $E(a_i) = 1$. The phrase 'a_i in E' therefore means that $E(a_i) = 1$, i.e. that the fibre a_i fires during the event E.

0.4.4. A *subevent* on A, usually denoted by letters like X, Y, assigns the value 0 or 1 to a subset of the fibres $a_1, \ldots, a_N$. For example, if

$$X(a_i) = 1 \quad (1 \leqslant i \leqslant r),$$
$$X(a_i) = 0 \quad (r < i \leqslant s),$$

$X(a_i)$ is undefined for $i > s$, then X is a subevent on A. As in the case of full events, X can also mean the set of fibres a_i for which $X(a_i) = 1$: in the example therefore, X can stand for the set $\{a_1, \ldots, a_r\}$.

0.4.5. If X is a subevent, the set of fibres to which X assigns a value is called the *support* of X, and is written $S(X)$. Thus in the above example, $S(X) = \{a_1, \ldots, a_s\}$.

0.4.6. A set of events is called an *event space*, and is denoted by letters like $\mathfrak{E}$, $\mathfrak{F}$. A set of subevents is called a *subevent space*, and is denoted by letters like $\mathfrak{X}$, $\mathfrak{Y}$.

0.4.7. Greek letters are usually reserved for probability distributions. The letter λ, for example, often denotes the probability distribution induced over $\mathfrak{A}$ (the set of all possible events on $a_1, \ldots, a_N$) by the input events. Thus $\lambda(E)$ is the number of occurrences of the event E divided by the total elapsed time. If, instead of considering the whole of $A = \{a_1, \ldots, a_N\}$, attention is restricted to $A' = \{a_1, \ldots, a_r\}$, then

the space $\mathfrak{A}'$ of events on A' corresponds to a set of subevents on the original fibre set A. Every event in $\mathfrak{A}$ defines a unique event in $\mathfrak{A}'$, obtained by ignoring the fibres $a_{r+1}, \ldots, a_N$. Thus the full distribution λ over $\mathfrak{A}$ induces a distribution λ' over $\mathfrak{A}'$ obtained by looking only at the fibres $a_1, \ldots, a_r$. λ is called the *projection* onto $\mathfrak{A}'$ of λ. If $\mathfrak{X}$ is a subevent space, then the phrase 'λ' is the distribution induced over $\mathfrak{X}$ by the input' refers to the λ' induced from the full input probability distribution λ by projecting it onto $\mathfrak{X}$. If $\mathfrak{B}$ is any subset of $\mathfrak{A}$, then the *restriction* $\lambda|\mathfrak{B}$ of λ to $\mathfrak{B}$ is defined as follows:

$$(\lambda|\mathfrak{B})(E) = \lambda(E) \quad \text{when } E \text{ is in } \mathfrak{B},$$
$$(\lambda|\mathfrak{B})(E) = 0 \quad \text{elsewhere.}$$

0.4.8. Finally, it is often convenient to use various pieces of shorthand. The following is a list of the abbreviations used.

$\{\ |\ \}$ is a method of defining a set. For example, $\{a_i|1 \leqslant i \leqslant N\}$ means 'the set of fibres a_i which satisfy the condition that $1 \leqslant i \leqslant N$'.
s.t. means 'such that',
$\in$ means 'is a member of the set': e.g. $a_i \in E$,
$\notin$ means 'not $\in$',
$P(X|Y)$ is the conventional conditional probability of X given Y,
$\Rightarrow$ means 'implies',
$\Leftarrow$ means 'is implied by',
$\Leftrightarrow$ means 'implies and is implied by',
iff means 'if and only if',
$\exists$ means 'there exists',
$|\ |$ means 'the number of elements in': e.g. $|E|$ means 'the number of fibres that are active in the event E',

The following set-theoretic symbols are also used:

$E \cup F$ = the union of E and F,
$E \cap F$ = the intersection of E and F,
$E \setminus F$ = the set of elements which are in E but not in F,
$E \triangle F$ = the set of elements which are in exactly one of E and F,
$E \subseteq F$ means E is contained in or equal to F,
$E \subset F$ means E is contained in F and does not equal F.

The reader who is not familiar with this notation should not be put off by it. All the important arguments of the paper have been written out in full. An adequate understanding of its content may be achieved without reading the paragraphs in small type, which is where these symbols usually appear.

0.5. *Information measures*

The only universal measures of suitability, fit, and so forth, are information measures. Three are of principal importance in this paper, and are defined below. Others are derived as they are needed. All the spaces with which the paper is

concerned are finite, and therefore only discrete probability distributions need be considered. Definitions are given here only for the finite case, although every expression has its more general form.

0.5.1. *Entropy* (Shannon 1949).

The *entropy* of the discrete probability distribution $p_1, \ldots, p_s$ will be denoted by the letter h. Thus

$$h(p_1, \ldots, p_s) = \sum_{i=1}^{s} -p_i \log_2 p_i.$$

All logarithms are to base 2.

0.5.2. *Information gain* (Shannon 1949, and see Rēnyi 1961).

Let μ, ν be two discrete probability distributions over the same set of events:

$$\mu = (p_1, \ldots, p_s), \quad \sum_i p_i = 1,$$

$$\nu = (q_1, \ldots, q_s), \quad \sum_i q_i = 1.$$

Then the *information gain* due to μ given ν is

$$I(\mu|\nu) = \sum_i p_i \log_2 p_i/q_i.$$

0.5.3. *Information radius* (Sibson 1969).

Let $\mu_1, \ldots, \mu_n$ be discrete probability distributions over the same s events. $\mu_i = (p_{i1}, \ldots, p_{is})$, $\sum_j p_{ij} = 1$. Let $\mu = (p_1, \ldots, p_s)$, and write $\mu \gg \mu_i$ if $p_k = 0$ implies that $p_{ik} = 0$. Let $w_1, \ldots, w_n$ be positive numbers. Then the *information radius* of the μ_i with weights w_i, is

$$K\begin{pmatrix} \mu_1 \cdots \mu_n \\ w_1 \ldots w_n \end{pmatrix} = \operatorname*{infimum}_{\mu \gg \mu_1, \ldots, \mu_n} \frac{\sum_{i=1}^{n} w_i I(\mu_i|\mu)}{\sum_{i=1}^{n} w_i}.$$

This infimum is achieved uniquely when

$$\mu = \frac{\sum_{i=1}^{n} w_i \mu_i}{\sum_{i=1}^{n} w_i}.$$

K, the information radius, is an information measure of dissimilarity.

$$K\begin{pmatrix} \mu_1 & \mu_2 \\ 1 & 1 \end{pmatrix}$$

will be abbreviated to $K(\mu_1\mu_2)$. The nature of K is explained more fully where it is used.

§1. Foundations

1.0. *Introduction*

This section is concerned with the problem of what the brain does. The background and arguments it contains are directed towards the justification of the Fundamental Hypothesis (1.6). It is shown that despite the complications which arise in the early

processing of sensory information, this hypothesis is often valid for information with which the brain has to deal. The discussion proceeds by first exploring notions connected with the idea of eliminating information theoretic redundancy—an idea which has had a somewhat chequered career in neuropnysiology (see Barlow 1961 for discussion and references). Secondly, ideas connected with biological utility are developed; and finally these are combined with the ideas of the first part to produce the philosophy from which the theory is derived.

1.1. *Information theoretic redundancy*

1.1.1. *Redundancy and early processing of visual information*

The notion that the processing of sensory information is an operation designed to reduce the redundancy in its expression is attractive, and one that is helpful for understanding certain aspects of early coding. For example, the coding in the optic nerve of relative rather than absolute brightness prevents the repeated transmission of the average brightness of the visual field. The use of on-centre off-surround coding there is peculiarly suitable for another reason, namely that the visual world has a tendency to be locally homogeneous. Given that a particular point in the visual field has a certain luminance and colour, the chance that neighbouring points also do is high. This kind of redundancy would not be present if, for example, the world was like scattered, multi-coloured pepper.

The visual world has this tendency towards continuity because matter is cohesive: the existence of edges and boundaries is a consequence of this. It may be possible to view the next stages of visual processing—by the 'simple' and 'complex' Hubel & Wiesel (1962) cells of area 17—as a further recoding designed round the redundancy associated with the existence of edges, bars, and corners. The test of this is whether using these cells, it is easier to represent scenes from the real visual world than an arbitrary, peppery optic nerve input; and it probably is.

There are many other ways in which redundancies arise in visual information. The next most obvious are those introduced by the operations of translation, magnification, and by rotation. For these operations at least, the question of what to do with the redundancy to which they give rise poses no great difficulties of principle. The brain is, for example, much less interested in where an image is on the retina than on the relative positions of its various parts. In this case, the clear object of a portion of the processing must be to recode the input, perhaps gradually, in such a way that relative positions are preserved. This should probably be done so that if two objects are seen momentarily, each in a different position, orientation, and having a different size, then the accuracy with which they may be compared should depend upon the magnitude of these differences.

Various similar points can be made about early processing in the other sensory modalities; but enough has probably been said to make the two main points. They are first, that notions of pure redundancy reduction probably are involved in the early analysis of sensory information. Secondly, redundancy can occur in many forms. The variety is especially obvious nearer the periphery. Each form requires a

special mechanism to cope with it, and so, especially lower down in the brain, it is natural to expect a diversity of specialized coding tricks. Some of these have been found, and some have not.

1.1.2. *Redundancy and later visual processing*

A great deal of the redundancy in visual information arises out of the permanence of the world. This, which includes the tendency of matter to cohere, makes it natural to code for changes, and to look for common subevents, like lines, corners, and so forth, which concern only a small fraction of the total population of input fibres. Common subevents are often called *features*, and the ideas associated with the analysis of features are probably the most promising available concerning later processing. Their potential advantage is most clearly seen in the analysis of objects: the great hope they hold is in the possibility that objects may be recognized by checking for the presence of particular features. These features are imagined to be drawn from a central pool which is shared by all other objects, and which is not too large.

This kind of scheme for later visual processing introduces five main categories of problem:

(i) The discovery of the relevant feature vocabulary.
(ii) Coding features in a suitably invariant way.
(iii) Coding the relative positions of the features.
(iv) Partitioning the features so that information from one object is separated from information about other objects.
(v) The decision process itself.

'Object', in the case of visual information, has a fairly well-defined meaning, because of the coherence of matter; but these general ideas have a wider application. For example, an 'impression' of an auditory input may be obtained from its power spectrum: in such cases, the 'objects' are less tangible. But for now, it is enough to consider just the special, visual case.

Problems (i) and (v) are very general, and are dealt with later (§1.4, §2, §5). Problem (ii) is special, and only two points about it will be made here. First, lines and edges are preserved by magnification, so parts of problem (ii) are automatically solved. Secondly, it is only necessary to localize the components of any particular image to an extent that will prevent their confusion with other images. The exact positions of the edges and corners of an object need not be retained, because the general restraint of continuity of form will mean that exact relative positions can always be recovered from a knowledge of approximate relative positions, the number of terminations, and approximate lengths of segments. Hence the problems associated with translation of an image across the retina can begin to be solved quite early by recoding into elements which signal the existence of their corresponding features within a region of a particular size. The exact size will depend upon how unusual is the feature.

This in itself is of no use unless some way can be found of representing these

approximate relative positions: this is problem (iii). Fortunately, it is very easy to see how distance relations may be held by a codon representation (Marr 1969). The key is an idea of 'nearness'. Suppose $\{f_1, \ldots, f_n\}$ is a collection of features, endowed with approximate distance relations $d(f_i, f_j)$ between each pair. Suppose subsets of the set $\{f_1, \ldots, f_n\}$ are formed in such a way that those features which are near one another are more likely to be included in the same subset than those which are not. Then the subsets would contain information about the relative positions of the f_i (see Petrie 1899 for an intriguing natural occurrence of this effect). Techniques like multidimensional scaling can be used to recover metric information explicitly in this kind of situation (Kruskal 1964; Kendall 1969), but for the present purpose, it is enough to note that two different spatial configurations would produce two different subset collections.

There is thus no difficulty of principle in the idea of analysis of shape by roughly localized features: but it is clear that all these techniques rely a great deal on the ability to pick out the components of a *single* shape in the first place. That is, a successful solution to problem (iv) is a prerequisite for this kind of solution to problems (ii) and (iii). This involves searching for hard criteria which will enable the nervous system to split up its visual input into components from different objects.

The most obvious suitable criteria arise from the tendency of matter to cohere: they are continuity of form, of colour, of visual texture, and of movement. For example, most parts of a fleeing mouse are distinguished from the background by their movement. A solution in this case would be to have a mechanism which causes signals about movement in adjacent regions of the visual field to enhance one another, and to suppress information from nearby stationary objects. It is not difficult to devise mechanisms for this, and analogous ones for the other criteria.

These ideas about joining visual data up using certain fixed criteria, are collectively called techniques for *visual bonding*. It would be surprising if the visual system did not contain mechanisms for implementing at least some kinds of visual bonding, since the methods are powerful, and can be innate.

It can be seen from this discussion that although ideas about redundancy elimination probably do not determine the shape of later visual processing, they are capable of contributing to its study. Those problems of principle ((i) and (v)) which arise quite quickly can and will be dealt with: the crucial point is that technical problems ((ii)–(iv)) will usually involve the elimination of redundancy associated with special kinds of transformation—perhaps specific to one sensory mode. These problems can either be solved by brute memory (e.g. perhaps rotation for visual information) or by suitable tricks, like visual bonding. The point is that these problems usually can be overcome somehow; and this is the optimism one needs to propel one to study in a serious way the later difficulties, which are genuinely matters of principle.

1.1.3. *Redundancy and information storage*

There is a quite different possible application of information theoretic ideas, and it is associated with the notion of coding information to be stored. It is a matter of everyday experience that some things are more easily remembered than others. Patterns are easier to recall than randomly distributed lines or dots. It cannot be argued that the random picture contains more information in any *absolute* sense, since the calculation of its information content depends entirely upon the norm with which it is compared. If the norm is itself, the random picture contains no information. There can be no doubt that a normal person would have to store more information to remember the random picture than the patterned one; but this, in the first instance anyway, is a remark about the person, not about the pictures.

This illustrates the fundamental point of this section—that the amount of information a memory has to store to record a given signal depends upon the structure of the signal, and the structure of the memory. Let $\mathfrak{X}$ be an event space, and let σ be the probability distribution corresponding to the afferent signal: thus $\sigma(E)$, for E in $\mathfrak{X}$, in the probability that E will occur next. (The present crude point can be made without bringing in temporal correlations.) Let μ be the probability distribution which describes what the memory expects. Then the amount of information the memory requires to store σ is

$$h(\sigma:\mu) = \int_{\mathfrak{X}} -\log_2 \mu(E)\, \mathrm{d}\sigma(E).$$

This expression exists if and only if

$$\mu(E) = 0 \Rightarrow \sigma(E) = 0.$$

$h(\sigma:\mu)$ and $h(\sigma)$, the entropy of σ, are related by the following result. Assuming the memory can store σ, then:

Lemma. $I(\sigma|\mu)$ exists, and $h(\sigma:\mu) = h(\sigma) + I(\sigma|\mu)$.

Proof. If the memory can store σ, $\mu(E) = 0 \Rightarrow \sigma(E) = 0$, and hence $I(\sigma|\mu)$ exists. Now

$$\begin{aligned} h(\sigma:\mu) &= \int -\log_2 \mu(E)\, \mathrm{d}\sigma(E) \\ &= \int \left\{ \log_2 \frac{\sigma(E)}{\mu(E)} - \log_2 \sigma(E) \right\} \mathrm{d}\sigma(E) \\ &= I(\sigma|\mu) + h(\sigma). \end{aligned}$$

The term $h(\sigma)$ is inevitable, but the term $I(\sigma|\mu)$ reflects the fundamental choice a memory has when instructed to store a signal σ. It can either store it straight, at cost $h(\sigma:\mu)$, or it can change its internal structure to a new distribution, μ' say, and store the signal relative to that. The amount of information required to change the structure from μ to μ' is at least $K(\mu\,\mu')$, where K is the information radius (§ 0.5.3); but, though an expensive outlay, it can lead to great savings in the long run if μ' is a good fit to the incoming information.

These arguments are too general to warrant further precise development, but

they do illustrate the two possibilities for a memory which has to store information: either it can store it raw, or it can develop a new language which better fits the information, and store it in terms of that. To this point, the next section § 1.2 will return.

Finally, this result shows how important it is to examine the structure of a memory before trying to compute the amount of information needed to store any given signal; it would therefore be disappointing to leave it without some remarks on the type of internal distributions μ we may expect to find in the actual brain. The obvious kind of answer is the distributions induced by a codon representation—as in the cerebellum. The reliability of a memory is measured by the number of wrong answers it gives when asked whether the current event has been learned. This in turn depends upon the number of possible input events: in cases where this is huge, the memory need only arrange that the proportion of wrong to right answers remains low. In smaller event spaces, a memory may have to represent the learned distribution a good deal more accurately. The first case may well correspond to the situation in the cerebellum and allows codons of a relatively small size: the second may require them to be much larger. The result relevant to this appears in § 3, but the situation even in the cerebellum may in fact be rather more complicated (Blomfield & Marr 1970).

1.2. *Concept formation and redundancy*

1.2.1. *The relevance of concepts*

It was shown in § 1.1.3 that one policy available to a memory faced with having to store a signal is to construct for it a special language. In the present context, this is bound to suggest the notion of concept formation.

It is difficult to doubt that one of the most important ways in which the nervous system eventually deals with sensory information is to form concepts with which to decompose and classify it. For example, the concepts *chair*, *sun*, *lover*, *music* all have their use in the description of the world; and so, at a lower level, do the notions of *line*, *edge*, *tone* and so forth.

Concepts, in general, are things which ease the nervous system's task; and although they do this in various ways, many of these ways produce their advantage by characterizing (and hence circumventing) a particular source of redundancy. One especially important example of how a concept does this is by expressing a part or the whole of that which many 'things' or 'objects' have in common. This 'common' element may take many forms: the objects' representations by sensory receptors may be related; some aspect of their functions may be the same; they may have common associations; or they may simply have occurred frequently in the experience of the observing organism.

This notion has the corollary that concept formation should be a natural consequence of the discovery of a large enough source of redundancy in the input generating a brain's experience. For example, if it is noticed that a certain collection of features commonly occurs, this collection should be recoded as a new and separate

entity: for this new entity, special recognition apparatus should be set up, and this then joins the vocabulary of concepts through which the brain interprets and records its experience.

Finally, concepts have been discussed as a means of formulating relationships between collections of other 'things', 'objects', or 'features'. This appears to rest upon the imprecise notions of 'thing', 'object' or 'feature': but there is in fact no undefinable notion present, for these can simply be regarded as concepts (or roughly, occurrences of concepts) that have previously been formed. This inductive step allows the argument to be taken back to the primitive input elements on which the whole structure is built; and in neurophysiology, there is no fundamental problem to finding a meaning for these: they are either the signals in axons that constitute the great afferent sensory tracts, or the features automatically coded for in the nervous system.

1.2.2. *Obstacles*

Something of a case can therefore be made for a connexion between concept formation and the coding out of redundancy, but it would be wrong to suggest this is all that is involved. Concept formation is a selective process, not always a simple recoding: quite as important as coding out redundancy is the operation of throwing away information which is irrelevant. For the moment however (until §1.4) it is convenient to ignore the possibility that a recoding process might positively be designed to lose information, and to concentrate on the difficulties involved in recoding a redundant signal into a more suitable form.

The general prospects for this operation are not good: this is for the same reason that the proofs of Shannon's (1949) main coding theorems are non-constructive. There exists no general finite apparatus which will 'remove redundancy' from a signal in a channel. Different kinds of signal are redundant in esoteric ways, and any particular signal demands an analysis which is specially tailored to its individual quirks. Hence the only hope for a general theory is that a particular *sort* of redundancy be especially common: a system to deal with that would then have a general application. Fortunately, it is likely there does exist such a form; and with its detailed discussion the next section is concerned.

1.3. *Problems in spatial redundancy*

1.3.0. *Introduction*

The term *spatial redundancy* means that redundancy which is preserved by any reordering of the input events (of which only a finite number have occurred); it thus fails to take account of causal or correlative relations which hold between events at different times. It is the only kind of redundancy with whose detection this paper deals. The complications introduced by considering temporal correlations as well are severe, and anyway cannot be discussed without some way of storing temporally extended events. This requires Simple Memory theory, and must therefore be postponed.

The particular kind of spatial redundancy with which this section is concerned is the sort which arises from the fact that some objects look alike. This will be interpreted as meaning that some objects share more 'features' than others, where 'features' are previously constructed classes, as outlined in § 1.1.2. It is conjectured that this kind of information forms the basis for the classification of objects by the brain: but before examining in detail the mechanism by which it is done, some arguments must be presented for the general notion that something of this sort is possible.

1.3.1. *Numerical taxonomy*

Evidence to support this hypothesis may be derived from recent studies in automatic classification techniques. The most important work in this field concerns the use of cluster methods to compute classes from information about the pairwise dissimilarities of the objects concerned (Jardine & Sibson 1968). There are two steps to the process. The first computes the pairwise dissimilarities of the objects from data about the features each object possesses. For this, the information radius (Sibson 1969; Jardine & Sibson 1970) is used, and in the case where the features are of an all-or-none type (i.e. an object either does or does not possess any given feature), this takes a simple form. Suppose object $\boldsymbol{O}_1$ possesses features $f_1, \ldots, f_n$, and object $\boldsymbol{O}_2$ possesses features $f_{r+1}, \ldots, f_m$, $1 < r < n < m$. Then $K(\boldsymbol{O}_1\ \boldsymbol{O}_2)$, the information radius associated with $\boldsymbol{O}_1$ and $\boldsymbol{O}_2$, (regarded as point distributions), is simply $r+(m-n)$, the number of features which exactly one object of the pair possesses.

The second step of the classification process uses a cluster method to compute classes from the information radius measurements. Various arguments can be put forward to show that some cluster methods are greatly to be preferred to others (Sibson 1970). Unlike the measurement of dissimilarity, these have not been given an information theoretic background; but to do so would require a firm idea of the purpose of the classification. The kind of assumption one would need would be to require that the classification provide the best way of storing the information relative to some measure—for example, a product distribution generated by assigning particular probabilities to the individual features. There is considerable choice, however, and it is unlikely that any particular measure could be shown to be natural in any sense.

It is not argued that any cluster process actually occurs in the brain: the importance of this work to the present enquiry is more indirect, and consists of two basic points. The first arises out of the *type* of redundancy these methods detect. It is that the objects concerned do not have randomly distributed collections of features: what happens is that classes of objects exist which produce collections of features that overlap much more than they should on the hypothesis of randomness. This fact, together with some kind of convexity condition which asserts that an object must be included if enough like it are, is fundamental to the classifying process.

The second point is that cluster analysis works. A large amount of information has

been analysed by such programs, especially information about the attributes of various plants. It has been found that these methods do give the classifications which people naturally make. This is important, for it shows that people probably use some process associated with the detection of this kind of redundancy for the classification of a wide range of objects. The motivation for studying methods for detecting this kind of redundancy now becomes strong.

1.3.2. *Mountain climbing in a probabilistic landscape*

In the brain, one may expect feature detectors to exist, if the recognition of objects is based on this sort of analysis. If spatial redundancy (§1.3.0) is present in the input, there will exist collections of features which tend to occur together. This phenomenon can be given the following more picturesque description. Let the input fibres $a_1, \ldots, a_N$ represent feature detectors, and let $\mathfrak{A}$ be the set of events on $\{a_1, \ldots, a_N\}$ (§0.4). Endow $\mathfrak{A}$ with the distance function d, where $d(E, F) =$ the number of fibres at which the events E and F disagree. $(\mathfrak{A}, d)$ is a metric space, and in fact $d(E, F) = K(E, F)$, where K is the information radius.

Imagine the space $(\mathfrak{A}, d)$ laid out, with the probability $p(E)$ of each event $E \in \mathfrak{A}$ represented by an extension in a new dimension. $p(E)$ is called the 'height' of E. It will be clear that if E occurs more frequently than F, $p(E) > p(F)$ and E is higher than F. In this way, the environment may be regarded as landscaping the space $\mathfrak{A}$, in which the mountains correspond to areas of events which are frequent, and the valley to events which are rare.

The important point about the choice of d for the metric on $\mathfrak{A}$ is that nearby inputs (under d) possess nearly the same features. Hence if a number of inputs commonly occur with very similar collections of features, they will turn out as a mountain in $(\mathfrak{A}, d)$ under p. The detection of such collections is thus equivalent to the discovery in the space $(\mathfrak{A}, d)$ of the mountains induced by p. The problem of discovering such mountains is solved in §5. Two other problems concern the choice of the feature detectors $\{a_1, \ldots, a_N\}$ with which to form the space $\mathfrak{A}$; and the question of what exactly one does with a mountain when it has been discovered. These are dealt with next. The point that this section illustrates is that the mountain idea over the space $(\mathfrak{A}, d)$ characterizes the kind of redundancy in which we are interested.

1.3.3. *The partition problem*

The prospects for discovering mountains in the space $\mathfrak{A}$, given that they are there, are good; but whether they are there or not depends largely on the choice of the feature detectors $\{a_1, \ldots, a_N\}$. There can be no guarantee that an arbitrarily chosen collection of features will generate a probabilistic landscape of any interest.

The discovery of an appropriate $\mathfrak{A}$ needs methods whereby features which are likely to be related are brought together. This is called the *partition problem*, and is in general extremely difficult to solve. The problem for which visual bonding was introduced in §1.1.2 was an example of how special tricks can in certain circumstances be used to solve it.

If no bonding tricks are known, however, the discovery of suitable spaces must rest upon measuring correlations of various kinds over likely looking populations of events. This is an operation whose rate of success depends upon the size of the available memory. It needs the theory of Simple Memory, and will be discussed more fully there. Suffice it here to say that the problem is not totally intractable despite the huge sizes of all the relevant event spaces. The reason is that only a very small proportion of the possible events can ever actually occur, simply because of the length of time for which a brain lives. This means, first, that the memory can be quite coarse; and secondly, that if anything much happens twice, it is almost certain to be significant.

1.4. *The recoding dilemma*

1.4.0. *Introduction*

The attraction of mountains is that when applied to the correct space, they provide a neat characterization of the type of redundancy which, there is reason to believe, is important for the classification of objects, and probably much else besides. The question that has now to be discussed is what to do with a mountain when it has been discovered. The obvious thing to do is to lump the events of a mountain together and call it a class. The problems arise because there is virtually no hope of ever saying *why* this is the right thing to do, using purely information theoretic ideas; and until this is specified, it will be impossible to say exactly how the lumping should be done.

The basic difficulty is that the lumping process involves losing information—about the difference between the events lumped together. The simplest reason why this process might be justifiable, or even desirable, is reliability. It would be implausible to suppose that the interpretation of an input might fail because of the failure of a single fibre. Hence a recognition apparatus for the particular event X must admit the possibility that an input Y with $d(X, Y) = 1$ or 2 (say) should be treated like X. But it is only by introducing such an assumption that this kind of step could be made, at least within the framework of the arguments set up so far.

1.4.1. *Information theoretic assumptions of a suitable nature*

The problem about trying to develop information theoretic hypotheses to act as justification for ignoring the difference between two events is that from an absolute point of view, one might just as well confuse two events with $d(X, Y)$ large as with $d(X, Y)$ small: there is no deep reason for preferring pairs of the second sort. It is natural to hope that in some sense, less information is lost by confusing nearby events, but in order for this to be true, something has to be assumed about the way two events can be compared. This effectively means comparing them to one—or a family of—reference distributions, whose choice must be arbitrary, and equivalent to some statement that nearby events are related. The theory thus becomes self-defeating, and the realization that this must be so allows exactly one observation to be made—namely that information theoretic arguments alone can never suffice to form a basis for a neurophysiological theory.

1.4.2. *Landslide*

The mountain structure of 1.3.2 depends on two things: the environmental probability distribution p, and the metric d. But it has been shown in 1.4.1 that the particular choice of d for the metric cannot be justified in any absolute way. The view that these mountains are important can therefore receive no support from any theory, based solely on ideas about storage, which does not assume that the first information to be thrown out is that which distinguishes the different parts of one mountain. In order to see how this might in fact be so, it is therefore necessary to return to the real world, to discover how some information may be important, while some may be expendable.

1.5. *Biological utility*

1.5.0. *The general argument*

The question with which this section is concerned is why should it ever be an advantage to classify together the events of a mountain. To answer this requires a clear idea of what the brain classifies *for*: only when this is known can it be deduced what kind of information is irrelevant, and hence which events may be classified together. The answer which will be proposed is that the classifications the brain eventually derives are ones which allow the deduction of the presence or absence of a *property* or *properties*, not necessarily directly observable, from such information as is at the time available. The word 'property' means here a slightly generalized idea of a feature: that is, it includes specifications of things an object can do, or can have done to it, as well as, for example, the sound it makes or the colour it has.

1.5.1. *Examples*

It is helpful at this point to give some concrete instances of the general statement made above. In its purest form, it implies a simple learning device, to which instances of the property concerned are transmitted through one channel, while information from which this property is to be diagnosed is conveyed through another. This corresponds exactly to the situation proposed for the cerebellar cortex in a recent theory of that structure (Marr 1969): the first channel is the climbing fibre input, and the second, the mossy fibres. There clearly exist stern limitations to this idea in any more general application, since in the cerebellar model, a property can only be diagnosed in conditions which are virtually a replica of a previous state in which the property was known to hold. It is, nevertheless, a primitive example of the central idea.

The property concerned need not be the immediate implementation of a particular elemental movement: it might be whether or not a particular branch can support the weight of a particular monkey. The animal concerned clearly needs to be able to make this discrimination, and to be able to do so by methods other than direct experiment. The information available is the appearance of the branch, from which it is possible to produce a reliable estimate of its strength. It is supposed that the

animal used data obtained by direct experiment (in play during his youth), to set up the appropriate classification apparatus.

These two cases illustrate the idea of a classificatory scheme designed for the diagnosis of properties not directly or immediately observable. It is helpful to make the following

Definition. An *intrinsic* property is one the presence or absence of which is known, and which is being used to decide whether another property is present. The word 'intrinsic' is used for this because if a property-detecting fibre a_i is in the support of a space over which there is a mountain, then that property is in a real sense an intrinsic part of the structure of the mountain. The second part of the definition follows naturally: an *extrinsic* property is one whose presence or absence is currently being diagnosed. These two words have only a local meaning: they are simply a useful way of describing which side of a decision process a particular property lies.

Classification for biological utility may therefore be regarded as the diagnosis of important but not immediately observable properties from information which is easy to obtain; and although this to some extent begs the question of what is an important property, it, nevertheless, represents some advance. Its strength is that it shows what information may be lost—namely the difference between events which lead to a correct diagnosis of a given property. The weakness of this approach is that it contains no scope for generalization from situations in which a property is known to hold, to new situations; and therefore seems to reduce operations in the brain to a simple form of memory.

1.5.2. *The dichotomy*

It may fairly be said that the remarks of this and the last sections force a dichotomy. On the one hand, there are the attractive and elegant ideas associated with coding for features, and their connexion with mountains and pure classification theory. These have been shown to be an insufficient basis for a theory, but they have a strong intuitive appeal. On the other hand, there are the nakedly practical ideas associated with strict biological utility. These have the advantage of giving a criterion for what information can be ignored, but in this crude shape, they suggest a memorizing system which performs more or less by brute force. There is no hope for either of these approaches unless they can be reconciled; and for this task, the next section is reserved.

1.6. *The fundamental hypothesis*

1.6.0. *The nature of a reconciliation*

Before trying to discover how these two views may be united, one must have a clear idea of the nature of any statement which could bring them together. The first view was of a kind of classification scheme which might be used by the brain. It consisted of selecting regions of commonly occurring subevents in event spaces over a collection of feature-detecting fibres, such that the subevents selected differed

rather little from one another. The second view suggested that the main function of the analysis of sensory information was to deduce properties of importance to the needs of the animal from such information as is available. These can only be reconciled if classification by mountain selection *does* prove a good guide to the presence of important properties: to decide whether this is so, properties of the real world must be considered.

1.6.1. *Validity for properties which are usually intrinsic*

Let $\mathfrak{A}$ be the event space on the feature-detecting fibres $\{a_1, \ldots, a_N\}$, and let λ be the probability distribution induced over $\mathfrak{A}$ by the environment. d is the natural metric defined in §1.3.2. In a general input subevent, the value of each fibre will be 0, or 1, or will be undefined. The last case can arise, for example, in the case of visual information, when part of an object is hidden behind something else. In this way, a property which is usually observable may sometimes not be. It will now be shown that classes obtained by lumping together events of a mountain over $(\mathfrak{A}, d)$ can usually act as diagnostic classes for such properties.

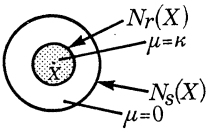

FIGURE 1. An illustration of the form of redundancy being discussed: the probability distribution μ induced by the environment over $N_s(X)$ has non-zero values only in $N_r(X)$.

Let $X \in \mathfrak{A}$ be an event of $\mathfrak{A}$, and let $N_r(X) = \{Y \mid Y \in \mathfrak{A} \text{ and } d(X, Y) \leqslant r\}$. A 'mountain' in $\mathfrak{A}$ might correspond to some distribution like μ where

$$\mu(Y) = \kappa, \quad Y \in N_r(X),$$
$$\mu(Y) = 0, \quad Y \in N_s(X) \backslash N_r(X),$$

where $s > r$, r is small, and κ is some positive constant. As soon as enough values of the a_i are known to determine an event as lying within $N_s(X)$, it follows that the event lies within $N_r(X)$ (see figure 1). Write p_i = probability that ($a_i = 1$ given $E \in N_r(X)$). Then if an event is diagnosed as falling within $N_s(X)$ without knowing the value of a_i, it can be asserted that $a_i = 1$ with probability about p_i. This is useful if p_i is near 0 or 1.

This kind of effect is a natural consequence of any mountain-like structure of λ over $\mathfrak{A}$, and allows that, in certain circumstances, these classes can be used to diagnose properties which are usually intrinsic. The values of a_i are not necessarily as expected—the piece of the object that is hidden may in fact be broken off; but the spikier the mountain (i.e. the smaller the local variance of λ), the nearer the p_i will be to 0 or 1, and the more certain the outcome.

1.6.2. *Extrinsic properties*

The argument for this kind of classification is that whenever there is a tendency for intrinsic properties to occur together in this way, it is extremely likely that there will also exist other properties, perhaps not directly observable ones, which also generalize over such groups of events. Hence, although the reason may not at the time be apparent, it will be good strategy for the animal to tend to make these classifications. Thus later, when a property is discovered to hold for one event in a given class of events, the animal will be inclined to associate it with members of the whole class. The generalization may or may not be found to be valid, but as long as it is successful sufficiently often, the animal will survive.

One other way of looking at this kind of generalization is to alter slightly the way one expresses the relevant kind of redundancy. It is equivalent to the assertion that once a context is sufficiently determined, one property may be a reliable indicator of another. The example cited earlier was of a monkey judging the strength of a branch. In practice, the thickness of a branch of a tree is a fairly reliable indicator of its strength, so that unless the branch is rotten, it will support the monkey if it is thick enough. Rottenness, too, can be visually diagnosed, so that a completely reliable assessment can be made on the basis of visual information alone. The context within which thickness and strength are related is roughly that the object in question is a branch of a tree, and is not rotten.

This kind of relationship is common in everyday experience; so common indeed that further examples are unnecessary. But although the general notion of this kind of redundancy has a clear importance, it is not obvious how the details might work in any particular case, nor that they may work the same way in any two. This problem must be tackled before any methods can be given for prescribing limits to the classes.

1.6.3. *Refining a classificatory unit*

The rough heuristic for picking out likely looking classes has been discussed at length. It was hinted that there may exist no *a priori* 'correct' way of assigning limits: where, for example, is the boundary between red and orange? The view that the present author takes is that although there are likely to exist fairly good general heuristics for class delimitation—like some kind of convexity property analogous to that which the cluster analysts use—there are probably no universal rules. It will be extremely difficult to give even these heuristics a satisfactory physical derivation: the kind of argument required is very indirect. But to say there exist no precise, generally applicable rules is merely to say that different properties have different relations to their indicators, and so is not very surprising. If, for example, an important extrinsic property is attached to a group of subevents, then its cessation marks the boundary of the class. If the property ceases to hold in a gradual way, the class will have problematical boundaries. This does not necessarily mean the class is not a useful one: the dubious cases may be rare, or may fall less dubiously into other classes. In any case, those falling well inside will be usefully dealt with.

It is therefore proposed that the exact specification of the boundaries to the classes should proceed by experiment. A new class is tentatively formed, upon the discovery of a promising mountain. If it turns out to have no attached extrinsic properties, it probably remains a slightly vague curiosity. If an extrinsic property more or less fits the provisional class, its boundary can be modified in a suitable way: this operation requires simple memory. If an extrinsic property is attached to it in no very sensible way—that is, instances of the property are scattered randomly or inconsistently over the class—then the class is no use as a reliable indicator, even with the available scope for shifting the boundaries. This does not necessarily render the class useless, for the property might be one which puts the animal in danger, and the class may contain all inputs associated with this kind of danger. For example, only a few kinds of snake are dangerous, but the class of snakes includes the class of dangerous snakes. It may be impossible to produce a reliable classification of snakes into dangerous and not dangerous without classifying some of them by species. This requires the consideration of more information than is necessary for diagnosis as a snake, and may be impossible without a potentially lethal investigation.

The investigation of the viability of a prospective class should probably be a very flexible process, drawing on the play of an animal when it is young, and upon the experience of life later on. Those classes which turn out, with slight alteration, to be useful will survive, while those which do not will not. Provided the initial class selection technique is neither wrong too often, nor fails too frequently to provide a guess where it should, the animal will be well served; and an instinct to explore his surroundings should enable him to remove any important errors.

1.6.4. *The Fundamental Hypothesis*

The conditions for the success of the general scheme of classification by mountain selection with later adjustments can now be explicitly characterized. It will work *whenever an extrinsic property is stable over small changes in its diagnostic intrinsic properties*. A given extrinsic property may possess more than one cluster of intrinsic properties which diagnose it, but as long as this condition is satisfied within each, the scheme will work. If a small change in intrinsic properties destroys an extrinsic property, either the boundary of the class passes near that point, or this extrinsic property cannot be diagnosed this way. In the former case, slight boundary changes can probably accommodate the situation: in the latter, there are two possible remedies. Either instances of the extrinsic property can be learned by rote—this can only be successful if the relationship of the extrinsic to the intrinsic properties is fixed—it is in any case arduous; or the intrinsic context has to be recoded. To the general recoding problem, there exists no general solution (by the remarks of §1.2.2).

The present theory is thus based on the existence of a particular kind of redundancy, not because it is redundancy as such, but because it is a special, useful sort. This is expressed by the following *Fundamental Hypothesis*:

Where instances of a particular collection of intrinsic properties (*i.e. properties*

already diagnosed from sensory information) tend to be grouped such that if some are present, most are, then other useful properties are likely to exist which generalize over such instances. Further, properties often are grouped in this way.

§2. The fundamental theorems

2.0. *Introduction*

The discussion has hitherto been concerned with the type of analysis which may be expected in the brains of sophisticated living animals. It was suggested that an important aspect of the computations they perform is the induction of extrinsic from intrinsic properties. This conclusion introduces three problems: first, collections of frequent, closely similar subevents have to be picked out. The Fundamental Hypothesis asserts that it is sensible to deal with such objects. This problem, the *discovery problem*, is dealt with in §5. Secondly, once a subevent mountain has been discovered, its set of subevents must be made into a new classificatory unit: this is the *representation problem*, and is dealt with in §4. Finally, on the basis of previous information about the way various extrinsic properties generalize over these collections of subevents, it must be decided whether any new subevent falls into a particular class. This is the *diagnosis problem*, and is dealt with now.

2.1. *Diagnosis: generalities*

A common method for selecting the hypothesis from a set $(\Omega_1, \ldots, \Omega_n)$ which best fits the occurrence of an event E, is to choose that Ω_i which maximizes $P(E|\Omega_i)$. Such a solution is called the maximum likelihood solution, and is the idea upon which the theory of Bayesian inference rests (see e.g. Kingman & Taylor 1966, p. 274, for a statement of Bayes's theorem). This method is certainly the best for the model in which it is usually developed, where the Ω_i may be regarded as random variables, and the conditional probabilities $P(E|\Omega_i)$, for $1 \leqslant i \leqslant n$, are known. The maximum likelihood solution will, for example, show how, and at what odds, one would have to place a bet on the nature of E in order to expect an overall profit. It is of course important to know all the conditional probabilities; and if the Ω_i are not independent, various complications can arise.

The situation with which the present theory must deal is different in several ways, of which two are of decisive importance. First, the prime task of the diagnostic process is to deal with events E_j which have never been seen before, and hence for which conditional probabilities $P(E_j|\Omega_i)$ cannot be known. It will further often be the case that E_j occurs only once in a brain's lifetime, yet that brain may correctly be quite certain about the nature of E_j.

Secondly, the prior knowledge available for inferring that E_j is (say) an Ω_i comes from the Fundamental Hypothesis. That is, the knowledge lies in the expectation that if E_j is 'like' a number of other E_k, all of which are an Ω_i, then E_j is probably also an Ω_i. This does not mean that $P(E_j|\Omega_i)$ is likely to be about the same as

$P(E_k|\Omega_i)$: frequency and similarity are quite distinct ideas. Hence if the Fundamental Hypothesis is to be used to aid in the diagnosis of classes—the assumption on which the present theory largely rests—then that diagnosis is bound to depend upon measurements of similarity rather than upon measurements of frequencies.

The analysis of frequencies of the events E_j is therefore relatively unimportant in the solution of the diagnosis problem; but it is of course extremely important for the discovery problem. The prediction that a particular classificatory unit will be useful rests upon the discovery that subevents often occur which are similar to some fixed subevent: the role of frequency here is transparently important. But when the new classificatory unit has been formed, diagnosis itself rests upon similarity alone.

An example will help to clarify these ideas. The concept of a poodle is clearly a useful one, since animals possessing most of the relevant features are fairly common. Further, a prize poodle is in some sense a poodle *par excellence*, and is as 'like' a poodle as one can get; but it is also extremely rare. The essential point seems to be that in a prize poodle are collected together more, and perhaps all, of the features upon which diagnosis as a poodle depends (or ought, in the eyes of poodle breeders, to depend).

These arguments imply that for the diagnosis of classificatory units by the brain, Bayesian methods are probably not used. Conditional probabilities of the form $P(E|\Omega)$ are thus largely irrelevant. The important question, when trying to decide whether E is an Ω, is how many of the events like E are definitely known to be an Ω. The computation of this raises entirely different issues.

2.2. *The notion of evidence*

The diagnosis of an input requires that an informed guess be made about it on the basis of the results for other inputs. If, for example, the present input E (say) has already occurred in the history of the brain, and has been found to deserve classification in a particular class, then its subsequent recognition as a member of that class is strictly a problem of memory, not of diagnosis. On the other hand, E may never have occurred before, though it might be that all E's neighbours have occurred, and have been classified in a particular way. The Fundamental Hypothesis asserts that this is good ground for classifying E in the same way.

The existence of an event similar to E, and known to be classified as, say, an Ω, therefore constitutes *evidence* that E should also be classified as an Ω. It will be clear that the more such events there are, the stronger the case for classifying E as an Ω.

It is appropriate to make two general remarks about evidence. The first concerns the absolute weight of evidence provided by Ω-classified events at different distances from E. Any theory must allow that for some categories of information, nearby events consitute strong evidence, whereas for others, they do not. Diagnoses within different categories will not necessarily employ the same weighting functions in the analyses of their evidence.

The second point about evidence concerns its adequacy. It may, for example, never be possible to diagnose correctly the class or property on the basis of evidence

from events on the fibres $\{a_1, \ldots, a_N\}$: they simply may not contain enough information. On the other hand they may contain irrelevant information, whose effect is to make the classifying task appear to be more difficult than it really is. This observation emphasizes the importance of picking the support of the mountain correctly.

The requirements of the diagnostic system can now be stated. It must:

(i) Operate only over a suitably chosen space of subevents (suggested by the Simple Memory). This space is called the *diagnostic space* for the property in question, Ω.

(ii) Record, as far as condition (iii) requires, which events of the diagnostic space have hitherto been found to be Ω's or not to be Ω's.

(iii) Be able, given a new event E, to examine events near E, discover whether they are Ω's or not, apply the weighting function appropriate to the category of Ω, and compute a measure of the certainty with which E itself may be diagnosed as an Ω.

The three crucial points now become:

P 1. How is the evidence stored?
P 2. How is the stored evidence consulted?
P 3. What is the weighting function (of (iii))?

The solutions to these which are proposed in this paper are not unique, but it is conjectured that they are the solutions which the nervous system actually uses. The key idea is that of an *evidence function*, which will in practice turn out to be a subset detector analogous to a cerebellar granule cell. The three points are resolved in the following way:

P 1. Evidence is stored in the form of conditional probabilities at modifiable synapses between 'evidence function' cells and a so-called 'output cell' for Ω, (eventually identified with a cortical pyramidal cell).

P 2. Evidence is consulted by applying an input event E, which causes evidence cells relevant to E to fire. The output cell then has active afferent synapses only from the relevant evidence cells. The exact way in which it deals with the evidence is analysed in § 2.3.

P 3. The weighting function comes about because nearby events will use overlapping evidence cells, just as very similar mossy fibre inputs are translated into firing in overlapping collections of cerebellar granule cells. The exact size of subset detector cells used for collecting evidence depends upon the category of Ω: recognition of speech may, for example, require a generally higher subset size than the 4 or 5 used in the cerebellar cortex.

Let $\mathfrak{X}$ be the diagnostic space for Ω, and let c be a function on $\mathfrak{X}$ which takes the value 0 or 1. c may, for example, be a detector of the subset A' of input fibres, in which case, for E in $\mathfrak{X}$, $c(E) = 1$ if and only if the event E assigns the value 1 to all the fibres in the collection A'; but c can in general be any binary function on $\mathfrak{X}$. Let $P(\Omega|c)$ denote the conditional probability (measured in the brain's experience so far) that the input is an Ω given that $c = 1$.

Definition. The pair $\langle c, P(\Omega|c)\rangle$ is called the *evidence* for Ω provided by the *evidence function* c.

The most important evidence functions are essentially subset detectors, (justified in §4.2.1), and it is convenient to give these functions a special name.

Definitions. (i) For all E in $\mathfrak{X}$, let $c(E) = 1$, if and only if $E(a_i) = 1$, $1 \leqslant i \leqslant r < N$. In this case, c is called an *r-codon*, or *r-codon function*, and is essentially a detector of the subset $\{a_1, \ldots, a_r\}$ of the input fibres.

(ii) For all E in $\mathfrak{X}$, let $c(E) = 1$ if and only if at least θ of

$$E(a_i) = 1,\ 1 \leqslant i \leqslant R < N.$$

In this case, c detects activity in at least θ of the R fibres $\{a_1, \ldots, a_R\}$, and is called an (R, θ)-codon.

The larger subset size, the fewer events E exist which have $c(E) = 1$, and so the more specifically c is tied to certain events in the space $\mathfrak{X}$. Let $|\mathfrak{X}|$ denote the number of events in $\mathfrak{X}$, and let κ be the number of events E in $\mathfrak{X}$ with $c(E) = 1$: then the fraction $\kappa/|\mathfrak{X}|$ is called the *quality* of the evidence produced by c in $\mathfrak{X}$. The qualities of various kinds of codon function are derived in §3.2.

2.3. *The diagnosis theorem*

The form of evidence has now been defined, and the rules for its collection have been set out. The information gained from the classification of one event, E, has been transferred to its neighbours in so far as they share subsets with E, and the subsets can be chosen to be of a size suitable for information of the category containing Ω. Thus problems P 1 and P 3 of §2.2 have been solved in outline: the details are cleared up in §§3 and 4. It remains only to discover the exact nature of the diagnostic operation: that is, to see exactly what function of the evidence consulted about E should serve as a measure of the likelihood that E is an Ω.

The problem may be stated precisely as follows. Let $\mathfrak{C} = \{\langle c_i, P(\Omega|c_i)\rangle\}_{i=1}^{M}$ be the collection of evidence available for the diagnosis of Ω over the space of events $\mathfrak{X}$. Let E be an event in $\mathfrak{X}$, and suppose

$$c_i(E) = 1 \quad (1 \leqslant i \leqslant k),$$
$$c_i(E) = 0 \quad (k < i \leqslant M).$$

That is, the evidence relevant to the diagnosis of E comes only from the functions $c_1, \ldots, c_k$, and is in the form of numbers $P(\Omega|c_1), \ldots, P(\Omega|c_k)$. The question is, what function of these numbers should be used to measure how certain it is that E is an Ω? The answer most consistent with the heuristic approach implied by the Fundamental Hypothesis is that function which gives the best results; this may be different for different categories. But a general theory must be clear about basic general functions if it can, and an abstract approach to this problem produces a definite and simple answer.

Suppose that, in order to obtain some idea of what this function is in the most general case, one assumes nothing except that E has occurred, and that the relevant

evidence is available. Then E effectively causes k different estimates of the probability of Ω to be made, since k of the c_i have the value one, and $P(\Omega|c_i = 1)$ is the information that is available. That is, E may be regarded as causing k different measurements of the probability that Ω has occurred. The system wishes to know what is the probability that Ω has actually occurred; and the best estimate of this is to take the arithmetic mean of the measurements. This suggests that the function which should be computed is the arithmetic mean of the probabilities constituting the available, relevant evidence; in other words, that the decision function, written $\mathcal{P}(\Omega|E)$ has the form

$$\mathcal{P}(\Omega|E) = \sum_{i=1}^{M} c_i(E) P(\Omega|c_i) \Big/ \sum_{i=1}^{M} c_i(E).$$

The conclusion one may draw from these arguments is that if one takes the most general view, assuming nothing about the diagnosis situation other than the evidence which E brings into play, then the arithmetic mean is the function which measures how likely it is that E is an Ω. The diagnosis theorem itself simply gives a formal proof of this. The meaning of the result is discussed in 2.4.

Lemma (Sibson 1969). Let T_i be a random variable which takes the value 0 with probability q_i, and 1 with probability $p_i = (1-q_i)$, for $1 \leqslant i \leqslant l$. Let T be another such variable, with corresponding probabilities q and p. Let p, q be chosen to minimize $\sum_{i=1}^{l} I(T_i|T)$, and let $p_0 = (1/l) \sum_{i=1}^{l} p_i$. Then $p = p_0$, and is unique.

Proof. Let $p_0 \neq 1, \neq 0$, and let T_0 be its corresponding binary valued random variable.

$$\begin{aligned}
&\sum_i I(T_i|T) - \sum_i I(T_i|T_0) \\
&\quad = \sum_i p_i \log_2 p_i/p + \sum_i q_i \log_2 q_i/q - \sum_i p_i \log_2 p_i/p_0 - \sum_i q_i \log_2 q_i/q_0 \\
&\quad = \sum_i p_i \log_2 p_0/p + \sum_i q_i \log_2 q_0/q = lI(T_0|T).
\end{aligned}$$

Hence
$$\sum_i I(T_i|T) = \sum_i I(T_i|T_0) + lI(T_0|T)$$

and I is always $\geqslant 0$. Thus
$$\sum_i I(T_i|T) \geqslant \sum_i I(T_i|T_0),$$

equality occurring only when $I(T_0|T) = 0$, i.e. when $T = T_0$. Hence the minimum value of $\sum_i I(T_i|T)$ is achieved uniquely when $p = p_0$.

Diagnosis theorem. Let Ω be a binary-valued random variable, and let $p_1, \ldots, p_k$ be independent estimates of the probability p that $\Omega = 1$. Then the maximum likelihood estimate for p is $p_0 = (1/k) \sum_i p_i$.

Proof. The estimate p_i of p may be regarded as being made through noise whose effect is to change the original binary signal Ω, which has distribution $(p, 1-p)$, into the observed binary random variable T_i (say), with distribution $(p_i, 1-p_i)$. The information gain due to the noise is $I(T_i|\Omega)$. Hence that value of p which attributes

least overall disruption to noise, and is therefore the maximum likelihood solution, is the one which minimizes $\sum_i I(T_i|\Omega)$. By the lemma, p is unique and equals p_0, the arithmetic mean of the p_i.

This result applies when the p_i are independent, or are so to speak symmetrically correlated. For example, if $T_1, \ldots, T_{k-1}$ are independent, but $T_k = T_{k-1}$, the result is clearly inappropriately weighted towards T_{k-1}. On the other hand, if k is even, and $T_1 = T_2, T_3 = T_4, \ldots, T_{k-1} = T_k$, this is not harmful. The general condition is complicated; but if $c_1, c_2, \ldots, c_M$ form a complete set of r-codons over the fibres $\{a_1, \ldots, a_N\}$, or a large random sample of such r-codons, then they are symmetrically correlated in the above sense.

$p = p_0$ gives the best single description of $p_1, \ldots, p_k$ in the sense that it minimizes $\sum_i I(T_i|T)$. The diagnosis theorem deals with a situation in fact rather far removed from the real one, and the next section is concerned with reservations about its application. It is not clear that any single general result can be established in a rigorous way for this diagnostic situation.

2.4. *Notes on the diagnosis theorem*

The key idea behind the present theory is that the brain decomposes its afferent information into what are essentially its natural cluster classes. The classes thus formed may be left alone, but are likely to be too coarse. They will often have to be decomposed still further, until the clusters fall inside the classes which in real life have to be discriminated; and they will often later have to be recombined, using, for example, an 'or' gate, into more useful ones, like specific numeral or letter detectors. These various operations are of obvious importance, but the basic emphasis of this approach is that the natural generalization classes in the naïve animal are the primary clusters. Diagnosis of a new input is achieved by measuring its similarity to other events in a cluster, and the similarity measure $\mathcal{P}$ of § 2.3 is proposed as suitable for this purpose. Its advantages are that it can be derived rigorously in an analogous situation in which the c_i are proper random variables; and that the result does not absolutely require that the c_i be independent. Moreover, the conditions under which dependence between the c_i is permissible (the 'symmetric' correlation of § 2.3) include those (when the c_i are a large sample of r-subset detectors) which resemble their proposed conditions of use (§ 4).

Nevertheless, the inference that if $\mathcal{P}(\Omega|E)$ is sufficiently high, then E is probably an Ω, rests upon the Fundamental Hypothesis. This observation raises a number of points, about the structure of the evidence functions, and about ways in which exceptions to the general rule can be dealt with. The various points are discussed in the following paragraphs.

2.4.1. *Codons for evidence*

The validity of the statement that a high $\mathcal{P}(\Omega|E)$ implies that E is an Ω rests upon the structure of the evidence functions used to obtain $\mathcal{P}$. The neural models of § 4

employ codons (i.e. subset detectors), but their physiological simplicity is not their only justification. In §4.2 it is shown, as far as the imprecision in its statement allows, that the Fundamental Hypothesis requires the use of rather small subset detectors for collecting evidence. It is not clear that advantage can at present be gained by sharpening the arguments set out there.

2.4.2. *Use of evidence of approximately uniform quality*

The reason for using functions c_i over $\mathfrak{X}$ at all, rather than simply collecting evidence with fibres a_j, is that the untransformed a_j would often not produce evidence of suitable quality. It may be possible simply to use fibres, especially for storing associational evidence (see §2.4.5); but it is probably also often necessary to create very specific codon functions giving high quality evidence for very selective classificatory units. This process must involve learning whenever the classes concerned are too specialized for much information about them to be carried genetically.

The quality of a piece of evidence is a measure of how specific it is to certain events in the diagnostic space $\mathfrak{X}$. In general, a given diagnostic task will require discriminations to be made above a minimum value p (say) of $\mathcal{P}$, and the quality of the evidence used will have to be sufficient to achieve such values of $\mathcal{P}$. The higher the quality of the evidence, the more there has to be to provide an adequate representation of $\mathfrak{X}$; and hence economy dictates that evidence for a particular discrimination should have as poor a quality as possible, subject to the condition on $\mathcal{P}$. Evidence of less than this minimal quality will serve only to degrade the overall quality, and so must be excluded. Hence, evidence should tend to have uniform quality. Mixing evidence of greatly different qualities is in general wasteful.

This condition is satisfied by the models of §4, where evidence is provided by (R, θ)-codons, and most of the evidence for a single classificatory unit has the same values of R and θ.

2.4.3. *Classifying to achieve a particular discrimination*

The quality of evidence function for a particular classificatory unit depends upon the minimum value p of $\mathcal{P}$ which is acceptable for a positive diagnosis, and this in turn will depend on how fine are the local discriminations which have to be made. The size of the clusters diagnosing the numeral '2' (say) in the relevant feature space depends upon the necessity for discriminating '2' from instances of other numerals and letters. The usual condition is probably that the part of the diagnostic space (over the relevant features) occupied by instances of a '2' must be covered by clusters contained wholly in that part. This condition fixes the minimum permissible value of p for diagnosis of a '2', which in turn fixes the subset sizes over any given diagnostic space. There may however be important qualifications necessary about this approach: the observations of §§2.4.4 and 2.4.5 can seriously affect the value of p.

2.4.4. *Evidence against* Ω

$\mathcal{P}$ will be most successful as a measure for diagnosis when the properties being diagnosed are stable over small changes in the input event. As E moves away from the centre of an Ω-cluster in the diagnostic space $\mathfrak{X}$, the values of $P(\Omega|c)$ where $c(E) = 1$ gradually decrease, and $\mathcal{P}$ decreases correspondingly. Provided these things happen reasonably slowly, all the remarks about symmetrical correlations of the evidence functions will hold in an adequate fashion.

The possibility must, however, be raised that within a general area of $\mathfrak{X}$ which tends to give a diagnosis of Ω, there exist special regions in which for some reason, Ω does not hold. Provided the region in which Ω does not hold is itself a cluster within the larger Ω-cluster, this state of affairs is not inconsistent with the Fundamental Hypothesis. This contingency can be dealt with in the same way as the diagnosis of Ω, by collecting evidence for 'not Ω'—evidence against Ω—within either $\mathfrak{X}$, or a space related to $\mathfrak{X}$. The form of the analysis is exactly the same as for Ω, except that the classificatory unit for 'not Ω' must be capable of overriding that for Ω. It is of course important for the successful diagnosis of Ω that diagnostic spaces for Ω and for 'not Ω' should both be appropriate, and both have evidence functions of suitable quality: but the mechanism which discovers the diagnostic space $\mathfrak{X}$ for Ω can clearly be used to discover the appropriate space for 'not Ω'.

It is interesting that this situation corresponds exactly to one proposed for the primary motor cortex. It has been suggested by Blomfield & Marr (1970) that the superficial cortical pyramidal cells there detect inappropriate firing of deep pyramidal cells. They presumably detect clusters in information describing the difference between an actual and an intended movement. These clusters in effect correspond to the need for deletion of activity in certain deep pyramids (an instance of the Fundamental Hypothesis), and the superficial pyramids cause the deletions to be learned in the cerebellar cortex. This distinction between the classes represented by deep and superficial cortical pyramidal cells may well not be restricted to area 4.

2.4.5. *Competing diagnoses and contextual clues*

It is often the case that a single retinal image could originate from two possible objects, yet contextual clues leave no doubt about which is the true source, and that source is the only one which is experienced. Such circumstances demonstrate the great importance of indirect information to the correct diagnosis of a sensory input. The present theory contains three ways by which such information may affect a diagnosis.

First, contextual information—for example, concerning the place one is in—may be included in the specification of the diagnostic space for Ω. There presumably exist classificatory units in one's brain for the places in which one commonly finds oneself, and other units which describe less common locations more pedantically: and these probably either fire all the time one is in the appropriate location, or (roughly) fire whenever other parts of the brain 'ask' where one is. Such information may be treated like more conventional sensory input.

Secondly, diagnostic criteria within categories can be relaxed by changing p. It is analogous to the ideas proposed in explanation of the collaterals of the cerebellar Purkinje cells (Marr 1969; Blomfield & Marr 1970). *A priori* information is sometimes available which makes units in one category more likely to be present following the diagnosis of units in another. In such cases, a general relaxation of the minimum acceptable value p of $\mathcal{P}$ over the relevant category will be appropriate.

Thirdly, and perhaps most important, is the matter of 'associational' contextual information. No additional theory is required, since such information can be treated as evidence in the usual way. It is probably for this kind of information that evidence functions are least often needed: direct association of classificatory unit detectors (cortical pyramidal cells) will often be adequate. The matter is touched on in §4.1.8, and dealt with at more length in Marr (1971, §2.4).

2.4.6. *General remarks about* $\mathcal{P}$

The direct technical importance of the Fundamental Hypothesis to the application of the results of the diagnosis theorem raises the wider issue of the extent to which one can feel justified in applying information-theoretic arguments to the kind of situation with which the diagnosis theorem deals. The Fundamental Hypothesis simply summarizes the view that clusters are useful. This is a heuristic approach, and it is not obvious that the diagnosis problem deserves any better than a heuristic approach itself. It probably matters rather little exactly what measure of similarity or fit is used: the redundancies on which the success of the system depends are so gross that there is probably more than one working alternative to $\mathcal{P}$.

If this is so, the diagnosis theorem loses much of its importance as a derivation of the 'correct' measure, since there may be no genuine sense in which *any* measure is correct, as long as it has a certain general form. The measure $\mathcal{P}$ does however seem intuitively plausible, and the reader may be happy to accept it without much justification. Theorem 2.3 is the best argument this author has discovered in its support; but it is not binding.

The measure $\mathcal{P}$ can be given a direct meaning in terms of the events of $\mathfrak{X}$. Let $\mathfrak{X}_i$ be the set of events E of $\mathfrak{X}$ with $c_i(E) = 1$. Then $P(\Omega|c_i)$ is the probability that if an an event of $\mathfrak{X}_i$ occurred, it was an Ω. Suppose that $\mathfrak{X}$ is the set of all events of size L on the fibres $\{a_1, \ldots, a_N\}$, and that the evidence functions $c_1, \ldots, c_M$ are the set of all r-codons. Let F be the new input event of $\mathfrak{X}$, which must be diagnosed; and let E be an arbitrary event of $\mathfrak{X}$. Write $d(E, F) = x$, d being the usual distance function of §1.2.

The number of r-subsets which E and F share is $\binom{L-x}{r}$, taking $\binom{y}{z}$ to be zero when $y < z$. Hence the weighting function which describes the 'influence' of E on the diagnosis of F is

$$\binom{L-x}{r} \Big/ \binom{L}{r}.$$

Thus the arithmetic mean obtained by the theorem of § 2.3 is

$$\mathcal{P}(\Omega|F) = \frac{\sum_{\substack{E \text{ in } \mathfrak{X} \\ E \text{ an } \Omega}} \lambda(E) \binom{L-x}{r} \Big/ \binom{L}{r}}{\sum_{E \text{ in } \mathfrak{X}} \lambda(E) \binom{L-x}{r} \Big/ \binom{L}{r}},$$

where λ is the probability distribution induced over $\mathfrak{X}$ hitherto by the environment.

2.5. *The interpretation theorem*

The diagnosis theorem 2.3 was concerned with the diagnosis of the property Ω over the diagnostic space $\mathfrak{X}$ on fibres $\{a_1, \ldots, a_N\}$. The events E in this situation specify the values of all the fibres $\{a_1, \ldots, a_N\}$; but it will frequently occur in practice that some values of the a_j will be undefined, and a decision has to be made on the basis of incomplete information. The problem is that this will mean that many of the evidence functions c_i are also undefined, thus leaving little if any evidence actually accessible to the input in question. For example, suppose a recognition system has been set up for a particular face: then a pencil sketch of that face can be recognized as such, even though much information—the colour of the eyes, skin, hair and so forth—is missing. Such a sketch can itself be analysed and set up as a new classificatory unit if that seems useful, and the mechanics of this process are the same as for the original. But this is a notion quite separate from the idea that the sketch is in some way related to the original face, and it is this idea with which the present section is concerned. The crux of the relationship is that the original face is the one which in some way best relates the sparse information contained in the features presented by the sketch. The result which follows characterizes this relationship precisely.

$\mathfrak{X}$, as usual, is the event space on $\{a_1, \ldots, a_N\}$. Let X be a subevent of $\mathfrak{X}$ which specifies the values of (say) $a_1, \ldots, a_r$ for some $r < N$. Then the event E in $\mathfrak{X}$ is a *completion* of X, written $E \vdash X$, if

(i) E specifies the values of all a_i, $1 \leqslant i \leqslant N$,

(ii) $E(a_i) = X(a_i)$ where $X(a_i)$ is defined.

Let $C = \{c_i | 1 \leqslant i \leqslant M\}$ be the set of functions on $\mathfrak{X}$ which provide evidence for the diagnosis of Ω. Since X is not a full event of $\mathfrak{X}$, $c_i(X)$ is undefined $(1 \leqslant i \leqslant M)$. Now there clearly exists a sense in which $c_i(X)$ might be defined: for example,

either $\quad c_i(E) = 1 \quad$ for all $\quad E$ in $\mathfrak{X}$ such that $E \vdash X$,

or $\quad c_i(E) = 0 \quad$ for all $\quad E$ in $\mathfrak{X}$ such that $E \vdash X$;

but such a circumstance is exceptional, and cannot be relied upon to provide adequate diagnostic criteria.

Let $\{E_1, \ldots, E_K\}$ be the set of all completions of X in $\mathfrak{X}$. Then clearly if $\mathcal{P}(\Omega|E_i)$ has the same value, q, for all $1 \leqslant i \leqslant K$, there are strong grounds for asserting that on the basis of the evidence from C, the estimate for $\mathcal{P}(\Omega|X)$ is also q. This result is a

special case of the following theorem. If $\mathcal{P}(\Omega|X)$ denotes the maximum likelihood value of the probability of Ω given X, taken from the evidence, $\mathcal{P}(\Omega|E)$ denotes the estimate arrived at in the diagnosis theorem, and $P(E_i|X)$ is a conventional conditional probability, then we have the

Interpretation theorem. Let X be a subevent of $\mathfrak{X}$ with completions $E_1, \ldots, E_K$. Then

$$\mathcal{P}(\Omega|X) = \sum_{i=1}^{K} \mathcal{P}(\Omega|E_i)\, P(E_i|X),$$

and is unique.

Proof. The argument is similar to that of the diagnosis theorem. Let $T_i(X)$ be a binary-valued random variable such that $T_i(X) = 1$ with probability $\mathcal{P}(\Omega|E_i) = p_i$ (say), for each i, $1 \leqslant i \leqslant K$. Let $\mathcal{P}(\Omega|X)$ correspond to a binary-valued random variable T where $T(X) = 1$ with probability p. Then each completion E_i of X corresponds to an estimate p_i of p, and $P(E_i|X)$ specifies the weight to be attached to this estimate. Hence by the same argument as that of the theorem 2.3, the maximum likelihood solution for T is that which minimizes

$$\sum_{i=1}^{K} P(E_i|X)\, I(T_i|T).$$

By an extension of the argument of the lemma 2.3., the value of p which achieves this is unique, and is

$$p = \sum_{i=1}^{K} P(E_i|X)\, p_i.$$

Hence

$$\mathcal{P}(\Omega|X) = \sum_{i=1}^{K} \mathcal{P}(\Omega|E_i)\, P(E_i|X),$$

and is unique.

Remarks. In general, no information about $P(E_i|X)$ will be available, so that $\mathcal{P}(\Omega|X)$ will usually be the arithmetic mean of $\mathcal{P}(\Omega|E_i)$ over those $E_i \vdash X$.

This theorem shows that incomplete information should be treated in a way which looks like an extension of the methods used for complete information, and the reservations of §2.4 apply equally here. The result does, however, have the satisfying consequence that the models of §4 designed to implement the diagnosis theorem automatically estimate the quantity derived in the interpretation theorem when presented with an incompletely specified input event.

§ 3. The codon representation

This section contains the technical preliminaries to the business of designing the concrete neural models which form the subject of the next. The results are mainly of an abstract or statistical nature, and despite the length of the formulae, are essentially simple.

3.1. *Simple synaptic distributions*

Let $\mathfrak{P}_1$, $\mathfrak{P}_2$ be two populations of cells, numbering N_1 and N_2 elements respectively. Suppose axons from the cells of $\mathfrak{P}_1$ are distributed randomly among the cells of $\mathfrak{P}_2$ in

such a way that a given cell $\boldsymbol{c}_1 \in \mathfrak{P}_1$ sends a synapse to a given cell $\boldsymbol{c}_2 \in \mathfrak{P}_2$ with probability z_{12}. z_{12} is called the *contact probability for* $\mathfrak{P}_1 \to \mathfrak{P}_2$.

If L of the cells in $\mathfrak{P}_1$ are firing, the probability that a given cell $\boldsymbol{c}_2 \in \mathfrak{P}_2$ receives synapses from exactly r active cells in $\mathfrak{P}_1$ is

$$\binom{L}{r} z_{12}^r (1-z_{12})^{L-r}. \tag{3.1.1}$$

Hence the probability that $\boldsymbol{c}_2$ receives at least R active synapses is X where

$$\begin{aligned} X(R, L, z_{12}) &= \sum_{r \geqslant R} \binom{L}{r} z_{12}^r (1-z_{12})^{L-r} \\ &= 1 - \sum_{r=0}^{R-1} \binom{L}{r} z_{12}^r (1-z_{12})^{L-r}. \end{aligned} \tag{3.1.2}$$

$X(R, L, z_{12})$ is called the *formation probability for* $\mathfrak{P}_1 \to \mathfrak{P}_2$.

Suppose the cells of $\mathfrak{P}_2$ receive synapses from no cells other than those of $\mathfrak{P}_1$ and that they have threshold R. The probability that exactly s cells in $\mathfrak{P}_2$ are caused to fire is

$$\binom{N_2}{s} X^s (1-X)^{N_2-s}, \quad \text{where} \quad X = X(R, L, z_{12}). \tag{3.1.3}$$

Hence the probability that at least S fire is

$$\sum_{s=S}^{N_2} \binom{N_2}{s} X^s (1-X)^{N_2-s}. \tag{3.1.4}$$

It is of some interest to know how well represented the L active cells of $\mathfrak{P}_1$ are by the cells of $\mathfrak{P}_2$ which they cause to fire. For most purposes, and all with which this paper is concerned, it is sufficient that any change in the cells which are firing in $\mathfrak{P}_1$ should cause a change in the cells of $\mathfrak{P}_2$. This is in general a complicated question, but a simple and useful guide is the following. Suppose the L cells of $\mathfrak{P}_1$ cause exactly R synapses to be active on each of S cells of $\mathfrak{P}_2$. Then the probability that at least one of the L active cells in $\mathfrak{P}_1$ sends a synapse to none of the active cells in $\mathfrak{P}_2$ is $(1-R/L)^S$. If R/L is small, this is approximately

$$\mathrm{e}^{-RS/L}. \tag{3.1.5}$$

3.2. *Quality of evidence from codon functions*

Codon functions, introduced in § 2.2, are associated with particular subsets of the input fibres in the sense that knowledge of the values of the fibres in a particular subset is enough to determine the value of the codon function. The larger the subset, the smaller the number of events at which the function takes the value 1, so the more specific that function is to any single event. Hence the general rule that r-codon functions provide better evidence the larger the value of r. This point is illustrated by the discrimination theorem which follows, and by various estimators of the quality of evidence to be expected from a codon function of a given size.

It is convenient to use the event space $\mathfrak{X}$ on fibres $\{a_1, \ldots, a_N\}$ such that in each event of $\mathfrak{X}$, exactly L of the fibres a_i have value 1. The set of such events is called the *code of size* L on $\{a_1, \ldots, a_N\}$. This involves no absolute restriction, but enables one to deal only with codon functions which assign the value 1 to all the fibres in their particular subsets, rather than allowing any arbitrary (but fixed) selection of 0's and 1's.

Let $\mathfrak{X}$ be the code of size L on $\{a_1, \ldots, a_N\}$, and let $\mathfrak{F}$ be a set of events of $\mathfrak{X}$—for example, $\mathfrak{F}$ may be the set of events with the property Ω. Let $\mathfrak{W}_r$ be the collection of all subsets of $\{a_1, \ldots, a_N\}$ of size r.

Definition. $\mathfrak{W}_r$ *discriminates* $\mathfrak{F}$ from the rest of $\mathfrak{X}$ if given $X \in \mathfrak{X}$, $X \notin \mathfrak{F}$, there exists a subset $C \in \mathfrak{W}_r$ such that $C \subseteq X$ but $C \nsubseteq Y$, for any $Y \in \mathfrak{F}$.

Theorem. Let $\mathfrak{F} \subset \mathfrak{X}$; then there exists a unique integer $R = R(\mathfrak{F})$ such that $\mathfrak{W}_r$ discriminates $\mathfrak{F}$ from $\mathfrak{X}$, all $r \geqslant R$.

Proof. If $\mathfrak{W}_r$ discriminates $\mathfrak{F}$ from $\mathfrak{X}$, any $\mathfrak{W}_s$ s.t. $\mathfrak{W}_s \supset \mathfrak{W}_r$ also discriminates $\mathfrak{F}$ from $\mathfrak{X}$. If $\mathfrak{F}$ can be discriminated by $\mathfrak{W}_r$, then $\mathfrak{F}$ can be discriminated by $\mathfrak{W}_{r+1}$, some set $\mathfrak{W}_{r+1}$ of $(r+1)$-subsets, since there will exist a set $\mathfrak{W}_{r+1}$ of $(r+1)$-subsets the set of whose r-subsets contains $\mathfrak{W}_r$. Finally, $\mathfrak{F}$ is always discriminated by $\mathfrak{W}_L = \{E | E \in \mathfrak{X}\}$. Hence there exists a unique lower bound R s.t. $\mathfrak{F}$ is discriminated from $\mathfrak{X}$ by all $\mathfrak{W}_r$ for $r \geqslant R$.

This shows that for a given discrimination task, $\mathfrak{F}$ from $\mathfrak{X}$, for which codon functions are to be used, the codons must be bigger than some lower bound R which depends on $\mathfrak{F}$.

Definition. R is called the *critical* codon size for $\mathfrak{F}$, and is written R_{crit}.

An *a priori* estimate of the likely value of the evidence obtained from a codon can be made by examining the number of events of various kinds over which the codon takes the value 1. Let $\mathfrak{X}$ be the code of size L on $\{a_1, \ldots, a_N\}$: $\mathfrak{X}$ contains $\binom{N}{L}$ events. Let λ denote the uniform probability distribution over $\mathfrak{X}$: i.e. $\lambda(E) = 1 \Big/ \binom{N}{L}$, all $E \in \mathfrak{X}$; and for $\mathfrak{F} \subseteq \mathfrak{X}$ write $\lambda(\mathfrak{F}) = \sum_{E \in \mathfrak{F}} \lambda(E)$. Then $\lambda(\mathfrak{F})$ simply measures the number of events in $\mathfrak{F}$.

The following results are useful.

3.2.1. Each input fibre is involved in L/N of the events in $\mathfrak{X}$ (under the distribution λ).

3.2.2. Let $\mathfrak{F} = \{E | (L - |E \cap F|) < \rho\}$ where F is some fixed event of $\mathfrak{X}$, and ρ is a positive integer. That is, $\mathfrak{F}$ is the ρ-neighbourhood of F. Then the number of events in $\mathfrak{F}$ is related to

$$\lambda(\mathfrak{F}) = \binom{N}{L}^{-1} \sum_{x=0}^{\rho} \binom{L}{L-x} \binom{N-L}{x}.$$

3.2.3. Now suppose c is an R-codon corresponding to an R-subset of the event F of §3.2.2. The number of events E such that $E \in \mathfrak{F}$ (of 3.2.2) and $c(E) = 1$ is related to

$$\lambda(\mathfrak{F} \cap \mathfrak{C}) = \binom{N}{L}^{-1} \sum_{x=0}^{\rho} \binom{L-R}{L-R-x} \binom{N-L}{x},$$

where $\mathfrak{C} = \{E |\ c(E) = 1\}$.

3.2.4. $\lambda(\mathfrak{C}) = \binom{N}{L}^{-1} \binom{N-R}{L-R}$, c an R-codon.

3.2.5. Suppose $\mathfrak{F}$, the ρ-neighbourhood of F, is a diagnostic class of $\mathfrak{X}$ for which the R-codon c (corresponding to a subset of F) is used to calculate evidence. Let Ω be the

property of being in $\mathfrak{F}$: then the value of $P(\Omega|c)$ that would be generated by the uniform distribution λ over $\mathfrak{X}$ is given by

$$\frac{\lambda(\mathfrak{F}\cap\mathfrak{C})}{\lambda(\mathfrak{C})} = \binom{N-R}{L-R}^{-1} \sum_{x=0}^{\rho} \binom{L-R}{L-R-x}\binom{N-L}{x} = p_R \text{ say,}$$

where c is an R-codon. Provided ρ is such that

$$\binom{N-L}{\rho} \text{ is large compared to } \binom{N-L}{\rho-1}$$

(that is ρ is smaller than say $\frac{1}{4}(N-L)$), $p_R \leqslant p_{R+1}$ if $\rho \leqslant (N-L)(L-R)/(N-R)$: so that for the simple case where the diagnostic class is a ρ-neighbourhood of some event F, increasing the codon size will, under any likely conditions, increase the expected quality of the evidence.

3.2.6. In the more complicated case where c is an (R,θ)-codon intersecting F in exactly S elements, we have

$$\frac{\lambda(\mathfrak{F}\cap\mathfrak{C})}{\lambda(\mathfrak{C})} = \frac{\sum_{\substack{x_i\geqslant 0 \\ \Sigma x_i = L \\ x_1+x_4\geqslant\theta}} \binom{S}{x_1}\binom{L-S}{x_1}\binom{N-L-R+S}{x_3}\binom{R-S}{x_4}}{\sum_{x=\theta}^{\mathrm{Min}(R,L)} \binom{R}{x}\binom{N-R}{L-x}}.$$

§4. The general neural representation

4.0. *Introduction*

This section is concerned with the design of neural models for implementing the theorems of §2. It is assumed that the exact nature of the classificatory units required has already been decided: only the representation problem is dealt with here. The discovery and refinement of new classificatory units is postponed until §5, where it is discussed within the context of the models developed now.

The central difficulty with producing neural models for a specific function is that there are many ways of doing the same thing: although the crucial averaging operation probably has to be performed at exactly one cell, there are many ways in which the supporting structure may vary. Both the form of the evidence, and the exact conditions under which it is used, are undefined; so the rigorous derivation of the basic neural models cannot proceed very far. This does not, however, commit the discussion to unredeemed vagueness. The injection at strategic points of a little common sense allows enough precision in the models to make their comparison in §6 with the known histology of non-specific cerebral neocortex a useful venture.

4.1. *Implementing the diagnosis theorem*

4.1.1. *Diagnosis by a single cell*

Theorem 2.3 suggests that the best estimate of the likelihood that a given event falls within a particular class is achieved by taking the average of the conditional probabilities offered by the relevant evidence. Suppose first that this operation is carried out by a single cell called the *output cell*: the arguments for this appear in

§4.1.7. Let $\boldsymbol{\Omega}$ be the cell in question, and Ω its associated property. $\boldsymbol{\Omega}$ receives afferent synapses from each of the evidence function cells $\boldsymbol{c}_i$ (cells which emit a signal—usually a burst of impulses—if and only if the input event E satisfies $c_i(E) = 1$). It is assumed that the strength of the synapse from the cell $\boldsymbol{c}_i$ for c_i to $\boldsymbol{\Omega}$ depends linearly on $P(\Omega|c_i)$. If, for Ω, the number of evidence functions c_i with $c_i(E) = 1$ is independent of E, $\boldsymbol{\Omega}$ has simply to add the values of $P(\Omega|c_i)$ for which $c_i(E) = 1$ since

$$\mathcal{P}(\Omega|E) = \sum_{i=1}^{M} k^{-1} c_i(E) P(\Omega|c_i) \propto \sum_{i=1}^{M} c_i(E) P(\Omega|c_i)$$

if k is independent of E. That is, $\boldsymbol{\Omega}$ has simply to add the weights of all the synapses from currently active evidence cells, and signal the result. It is easy to imagine that the firing rate of the cell $\boldsymbol{\Omega}$ should vary monotonically with the value of this sum.

The theory therefore requires that *the strength of the synapse from* $\boldsymbol{c}_i$ *to* $\boldsymbol{\Omega}$ *should depend linearly upon* $n_1 n_2^{-1}$ *where* n_1 *= the number of times* $c_i = 1$ *and a positive diagnosis was achieved, and* n_2 *= the number of times* $c_i = 1$. This condition can clearly be generated by some process in which a combination of pre- and post-synaptic firing causes the synapse to facilitate, while pre- without post-synaptic activity causes its power to decrease.

4.1.2. *Synaptic weights: the range of relevance*

Economical use of the full range of synaptic strength demands that the maximum strength of each synapse should be achieved at roughly the maximum value of $P(\Omega|c_i)$ taken over those c_i concerned with Ω. This value is not necessarily 1—indeed will rarely be 1: suppose it is q. Then the range of strengths available to each evidence synapse must represent the whole of $[0, q]$: it cannot be limited to $[p, q]$ for some $p > 0$, since the accurate calculation of $\mathcal{P}(\Omega|E)$ may often depend in part upon evidence suggesting it is very unlikely that E is an Ω.

Furthermore, all the evidence synapses at $\boldsymbol{\Omega}$ which are likely to be used with one another must have their strengths normalized to the same range $[0, q]$ in order that an unbiased sum may be taken. Any two synapses should be interchangeable, yet give the same output cell firing frequency. The range $[0, q]$ is called the *range of relevance* for evidence associated with Ω.

4.1.3. *The plausibility range*

Let $[0, q]$ be the range of relevance for evidence associated with Ω. The maximum value which $\mathcal{P}(\Omega|E)$ can achieve is at most q, and hence the maximum firing rate of $\boldsymbol{\Omega}$ should be reached at or near this value. Unlike the synaptic strengths, however, there is no need to be able to cover the whole range $[0, q]$, since the lower values may make the presence of Ω extremely unlikely. Let p be that value of $\mathcal{P}(\Omega|E)$ at and below which it is impossible that E ever is an Ω; then $[p, q]$ is called the *plausibility range* associated with Ω, and $0 \leqslant p < q \leqslant 1$. It is evident that some accuracy will be gained by representing only the plausibility range through the $\boldsymbol{\Omega}$-cell firing

frequency. Both p and q will depend upon the nature of the information with which Ω is dealing; there will exist no universally valid values.

The simplest view of the output cell coding of $\mathcal{P}(\Omega|E)$ thus requires that $\boldsymbol{\Omega}$ should not fire at all unless $\mathcal{P}(\Omega|E)$ exceeds some minimum value p, and that its maximum firing rate should be achieved at or near some maximum value q. The only restriction so far placed on the nature of the coding within the plausibility range is that it be monotonic increasing with $\mathcal{P}(\Omega|E)$. If the outputs of two cells have to be compared —to decide for example into which of two classes the current input falls—then unless unreasonable complications are introduced, they have to code $\mathcal{P}(\Omega|E)$ the same way. That is, they must have the same plausibility range $[p, q]$, and they have to code $\mathcal{P}(\Omega|E)$ identically (within the limits of permissible error) inside the plausibility range. Since it is often necessary to decide between classes of the same kind, it may be concluded that all output cells for diagnosing competing classes should be cells of the same construction: they should share a common plausibility range, and a common coding within it.

4.1.4. *Variable k*

The final complication to be added to the simple scheme of §4.1.1 which simply summed the weights of the active afferent synapses is that the number of such synapses may vary. $k = \sum_i c_i(E)$, and in general depends upon E. $\boldsymbol{\Omega}$ must therefore be associated with some mechanism which can compensate for this, and its effect must be to divide the total $\sum_i c_i(E) P(\Omega|c_i)$ by $k(E) = \sum_i c_i(E)$ for the current event E. The output cell firing frequency must therefore be monotonically related to

$$k^{-1}(E) \sum_i c_i(E) P(\Omega|c_i)$$

within the plausibility range for Ω.

4.1.5. *Computing* $\mathcal{P}(\Omega|E) - p$

The four possibilities for the sequence of operations carried out in the computation of $\mathcal{P}(\Omega|E) - p$ are represented by the bracketing in the following formulae.

$$k^{-1}(\sum_i (c_i(E)\,(P(\Omega|c_i) - p))), \tag{1}$$

$$(k^{-1}(\sum_i c_i(E) P(\Omega|c_i))) - p, \tag{2}$$

$$\sum_i k^{-1}\,(c_i(E)\,(P(\Omega|c_i) - p)), \tag{3}$$

$$(\sum_i k^{-1} c_i(E) P(\Omega|c_i)) - p. \tag{4}$$

In (1) and (2), the summation is performed before the division, whereas in (3) and (4) it is performed after. In (1) and (3), the subtraction is performed before the other operations, which are done on the residues: in (2) and (4) the subtraction is done last.

The smaller the numbers can be kept, the more accurate will be the final result; so other things being equal, computations which keep numbers small are to be preferred to ones which do not. Other things are equal in the choice between (1) and (2), and in the choice between (3) and (4). It is therefore natural to prefer (1) to (2), and (3) to (4).

In all these computations, a subtraction, summation and division have to be performed, so it is important to consider whether they can plausibly be executed by a real cortical neuron. Many types of cortical pyramidal cell will be identified in §6 as output cells, especially those types found in layers III and V of Cajal.

The synapses for $P(\Omega|c_i)$ are assumed to be excitatory, and only those with $c_i(E) = 1$ carry a signal. Hence there is no difficulty about arranging that only those $P(\Omega|c_i)$ with $c_i(E) = 1$ are considered. The summation of the active synapses is, as remarked in 4.1.1, an operation which it is quite plausible to assume possible in the dendrites of $\boldsymbol{\Omega}$.

The subtraction must be performed by inhibition. The actual amount of inhibition, in both (1) and (3), depends upon $k(E) = \sum_i c_i(E)$, which will vary with E, so the amount must depend upon the number of active evidence cells $\boldsymbol{c}_i$. This means that one or more inhibitory interneurons must have dendrites which sample the fibres from the $\boldsymbol{c}_i$-cells, and whose axons terminate on the dendrite of $\boldsymbol{\Omega}$ itself, near enough to the active c_i-cell synapses to interact with them in an additive way. The dendritic field of $\boldsymbol{\Omega}$ may be very large, in which case many inhibitory interneurons, each with a rather local dendritic field, will be needed to ensure each dendrite contributes its proper share to the sum.

Both (1) and (3) require that the subtraction be performed before the summation, and the idea of subtraction performed uniformly over the $\boldsymbol{\Omega}$ dendritic tree makes both schemes possible from this point of view. The great problems arise over the division, which has to be done if $k(E)$ varies significantly. (1) and (3) differ in the order in which the summation and the division are taken, so the discussion of division falls into two parts. First, can it be done at all; and secondly, if it can, does it appear that either of (1) and (3) is more likely?

Suppose for the moment that division can be performed. Observe that it has certainly to occur *after* an estimate of the total value of $k(E)$ has been made. This is because a division by (n_1+n_2) becomes complicated if one insists on dividing by n_1 first, and then performing some operation on n_2, since the nature of the operation to be performed using n_2 depends on the value of n_1. If division is to take place, therefore, an explicit estimate of $k(E)$ has to be made by the neural machinery. The actual division process has then to involve this estimate.

A distinction can be made between the mechanics of this process for (1) and for (3). If the division is done before the summation, it has to be done over the whole $\boldsymbol{\Omega}$ dendrite, and must therefore involve some kind of uniform field where intensity depends on $k(E)$. If, on the other hand, the summation is done first, the division might be a quite localized process.

4.1.6. *A model for division*

This is not the place for a detailed discussion of dendrite theory, but it is worth pointing out, by way of general support for the theory's plausibility, that there exists an extremely simple model for the process of division. Suppose G is a spike generator, and I is a spike inhibitor, as in figure 2. The spike generator produces impulses with some frequency ν, and models the result of the summation process. The spike inhibitor I has two inputs, one from G and one of strength which varies with $k(E)$, the

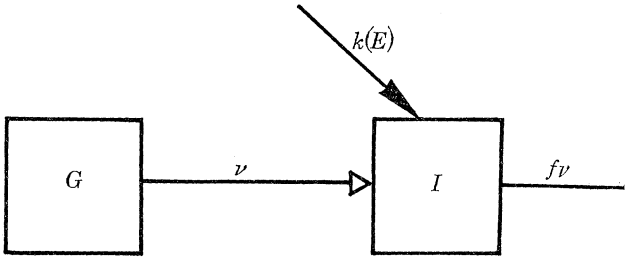

FIGURE 2. A model for division. The spike generator G emits spikes at a rate ν and the inhibitor I allows a fraction f to be transmitted, where $f \propto k^{-1}(E)$. Spikes are therefore emitted at a rate $f\nu \propto \nu k^{-1}(E)$.

number of currently active evidence cells. I is such that each incoming spike is transmitted with probability f, where f varies inversely with $k(E)$. That is, each incoming spike has a chance $f = Kk^{-1}(E)$ of crossing I, where K is some suitable normalizing constant. I may thus be regarded as a conducting medium with only a fraction f of its maximum ability to sustain a spike. The output spike frequency is then monotonically related to $\nu k^{-1}(E)$.

There are of course other models which have the same effect, but one fact seems to commend this above the rest: it is that spikes have been observed in the large dendritic stems of the cerebellar Purkinje cells (Eccles, Ito & Szentágothai 1967, p. 79) and of the hippocampal pyramidal cells (Spencer & Kandel 1961). It is therefore not unreasonable to suppose that the main apical dendrites of cortical pyramidal cells are also able to support spikes; and if so, that this is how the sum of the residues is communicated to the soma. It is, however, well known that many cortical pyramidal cells, especially those of layers III and V, have somas surrounded by basket cell synapses (Cajal 1911). These cells are well placed to make an estimate of $k(E)$, the amount of parallel fibre activity, and are almost certainly inhibitory. Their action might therefore have the effect that a proportion of the spikes from the dendrite fails to be transmitted to the axon, this proportion depending in a suitable way on the value of $k(E)$. The estimate of $k(E)$ itself could be the combined work of many basket cells, their contributions being summed at the soma itself.

If this model is correct, it provides an explanation of how the division process is performed, in the case in which it follows the summation of the residues. It thus favours the order of computation described by formula (1) of §4.1.5.

4.1.7. *Arguments for diagnosis by a single cell*

It is necessary now to justify the choice of using one rather than a collection of cells at which to compute a single decision. The arguments are these: first, the weights of the synapses from the evidence cells must vary with $P(\Omega|c_i)$ which depends, for each cell $\boldsymbol{c}_i$, on the number of positive diagnoses coincident with the firing of $\boldsymbol{c}_i$. Hence in order that every evidence synapse has the correct weight, all the output cells representing Ω at whose synapses the evidence is collected must fire every time a positive diagnosis is achieved. Hence either the output cells must be completely interconnected, or they must drive some super-output cell, which fires them all if it is itself fired.

Secondly, if evidence for Ω is collected and judged by many cells, the weight each ell has in the final decision ought to depend upon the amount of evidence it has considered. This could be arranged by some suitable trick, but the combination of this and the first point, though not compelling, favours the view that each decision process be carried out by one cell. If therefore, as also seems likely, there do exist several representations of any given concept, they are probably independent.

4.1.8. *Dual purpose output cells*

This concludes the discussion of the implementation of the theorem 2.3, but before leaving the topic to discuss the form of evidence functions, something must be said about driving the cell $\boldsymbol{\Omega}$ by information of two distinct types. If a single diagnosis could be achieved by two quite unrelated sets of evidence, with different plausibility ranges, it would be necessary to locate the relevant synapses on different, independent regions of dendrite. For example, use of $\boldsymbol{\Omega}$ with direct sensory information may involve synapses on the apical dendritic tree of a cortical pyramidal cell, whereas associational information may be held in the basilar dendrites. These systems could possess different values for both limits, p and q, of the plausibility range. They would require entirely different systems of inhibitory subtraction cells, and although the basket cells for the division function could in each case send synapses to the soma, their dendrites would have to sample the correct, disjoint populations of evidence fibres. The cell $\boldsymbol{\Omega}$ would then effectively become two cells in one, and it would succeed in this rôle as long as the other cells of its class also had the same specifications, and the same dual plausibility ranges.

If $\boldsymbol{\Omega}$ can be driven by sensory or by associational information, it is possible that conditional probabilities for sensory evidence should not count those instances of Ω which arise by association. This is because in the second rôle, $\boldsymbol{\Omega}$ may be being used symbolically, not directly. $P(\Omega|c_i)$ for sensory information should probably not be influenced by instances of this rôle.

Finally, the advantages of such dual rôle cells may be important. If all the various conditions are satisfied, they can probably combine in a satisfactory way information of two kinds in a single diagnostic process. This would to some extent be against the rules, but as long as the contravention is uniform over cells of the

relevant category, it would probably work. The effect would be to make it easier to see what you expect to see.

The results of this section are summarized in figure 3.

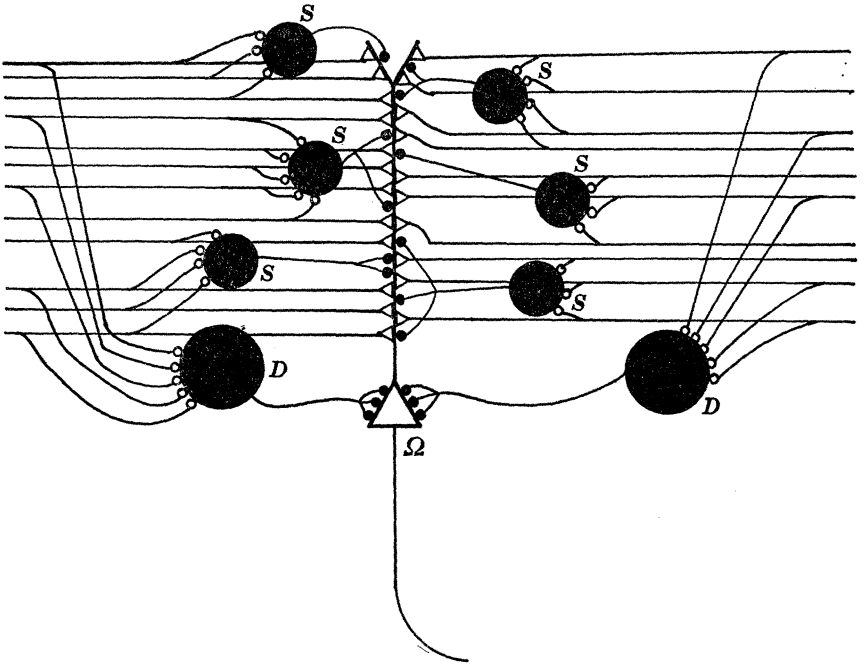

FIGURE 3. The output cell **Ω** has three kinds of afferent synapse: Hebb synapses (open triangles) from evidence cells, and two kinds of inhibitory synapse. Those from the ***S***-cells are spread over the dendritic tree, and perform a subtraction: those from the ***D***-cells, concentrated at the soma, perform a division.

4.2. *Codon functions for evidence*

4.2.0. *Standard evidence functions*

Two constraints have been placed on the evidence functions c_i for a particular output cell **Ω**: that the evidence they provide should be of sufficient quality, and that the amount of correlation between the c_i for Ω should be either negligible or regular in a way which does not cause improper bias. The choice of evidence function ought to depend upon the particular circumstances for which it is required: if especially efficient functions exist and can be constructed for a particular purpose, their use will permit an economy in the amount of structure required for that process. But it will frequently occur either that rather little is known about exactly what information will come to be held in a particular piece of cortex, or that there is nothing particular about that information which makes it a suitable candidate for special methods. For such cases, it is natural to seek a class of functions from which a 'standard' form of evidence may be constructed.

There are various conditions such a class should satisfy. Most important, they

should have a simple neural representation. Secondly, and also essential, there should be different categories of function corresponding to different expected qualities of the evidence to which they give rise. This is an economy condition, since it is wasteful to use better (and hence in general, more) evidence than necessary. Thirdly, according to the Fundamental Hypothesis § 1.6, the expected quality of the evidence produced by the function c will depend upon the distribution of the events E with $c(E) = 1$ over the event space $\mathfrak{X}$. If the property Ω which the cell $\mathbf{\Omega}$ is signalling is stable over relatively small changes in the input event E, the best evidence functions c will be those whose events F with $c(F) = 1$ are grouped together, as seen through the natural metric d of § 1.3.2.

4.2.1. *Arguments for codon functions*

These three conditions do have implications about the kind of evidence one may expect: they strongly suggest one particular family of functions, the generalized (R, θ)-codons. First, observe that figure 4 shows the simplest kind of afferent

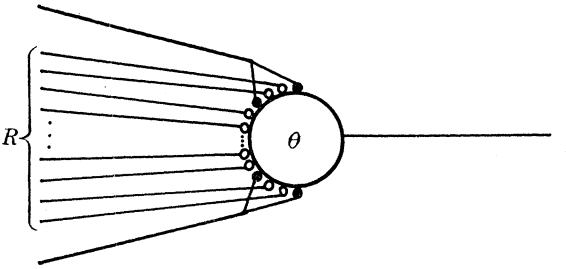

FIGURE 4. An (R, θ)-codon cell. There are R excitatory afferent synapses (open circles), and enough inhibition (filled circles) to give the cell a threshold of θ.

system possible for a cell. There are R afferent fibres, $a_{i_1}, \ldots, a_{i_R}$, each with an excitatory synapse of some fixed weight—1, say. The cell has threshold θ, which may be determined by some suitably arranged inhibition. Then the cell will emit a signal whenever at least θ of the R fibres $a_{i_1}, \ldots, a_{i_R}$ are active: hence the set of firing conditions for the cell constitutes an (R, θ)-codon on any event space over fibres which include $a_{i_1}, \ldots, a_{i_R}$. An (R, θ)-codon is thus a specification of the firing conditions for a cell whose afferent relations with its input fibres are simple, and anatomically and physiologically plausible.

Secondly, it has been observed in § 3.2 that suitable values of (R, θ) can be chosen to construct an (R, θ)-codon which will match any previously specified quality of evidence. Hence the second condition is fulfilled by the family of (R, θ)-codons. The various technical problems which arise when one tries to design a net which will produce (R, θ)-codons for a particular input can be solved, and will be discussed in the next section.

The above two arguments show that codon functions are sufficient to satisfy the two corresponding conditions: the next one shows that they are in some degree necessary for the third.

Let $\mathfrak{X}$ be the event space on $\{a_1, \ldots, a_N\}$ and let d be the natural metric of §1.3.2. Let $\{c_i\}_{i=1}^{M}$ be the evidence functions for a particular property Ω, and let Ω hold for a particular event $E \in \mathfrak{X}$, where

$$E(a_i) = 1 \quad (1 \leqslant i \leqslant L),$$
$$E(a_i) = 0 \quad (L < i \leqslant M).$$

Without loss of generality, suppose

$$c_i(E) = 1 \quad (1 \leqslant i \leqslant k),$$
$$c_i(E) = 0 \quad (k < i \leqslant M),$$

and choose $F \in \mathfrak{X}$ such that $d(E, F) = 1$. Then according to the Fundamental Hypothesis §1.6.4, the chance that F also has Ω is better than for an event arbitrarily selected from $\mathfrak{X}$. Hence most of the c_i with $c_i(E) = 1$ should have $c_i(F) = 1$ as well.

This argument applies to all F with $d(E, F) = 1$: so let $N_1(E) = \{F | d(E, F) \leqslant 1\}$. For each c_i, $1 \leqslant i \leqslant k$, define a subset C_i of $\{a_1, \ldots, a_N\}$ in the following way. Write $F_j =$ the event obtained from E by altering the value of the fibre a_j, i.e.

$$F_j(a_i) = E(a_i), \quad \text{all} \quad i \neq j,$$
$$F_j(a_j) = 1 \Leftrightarrow E(a_j) = 0.$$

The subset C_i is obtained thus:

$$C_i = \{a_j | c_i(F_j) \neq c_i(E)\}.$$

Then for $1 \leqslant i \leqslant k$,
$$c_i(F_j) = 1 \Leftrightarrow C_i \subseteq F_j.$$

That is, for $1 \leqslant i \leqslant k$, c_i *may be regarded within* $N_1(E)$ *as a detector of the subset* C_i *of the fibres* $\{a_1, \ldots, a_N\}$. Thus locally, (i.e. within $N_1(E)$), c_i behaves like the codon function with associated subset C_i.

But it has been observed that for an arbitrary change from E to F_j, some $1 \leqslant j \leqslant k$, the values of the majority of the functions c_i should remain unchanged. Hence, for most of the i, $1 \leqslant i \leqslant k$, it must be true that c_i takes the value 1 over most of $N_1(E)$, (assuming the c_i are not organized in any special way). This implies that the size of the subset C_i which c_i detects in $N_1(E)$ is *small*, for most i, $1 \leqslant i \leqslant k$.

This argument shows that if an evidence function is constructed for classifications in which the Fundamental Hypothesis is true, then such a function behaves locally like a codon function with a rather small associated subset.

This is the most that can be deduced about evidence functions from the necessarily imprecise considerations out of which the present theory is constructed. The case for (R, θ)-codons being the general form of evidence function is not logically established, but it would at present be impossible to make a rigorous argument for any family of functions. The three arguments presented above do constitute good evidence in favour of codons—evidence which it would require a strong and unexpected finding to disrupt.

Finally, in the particular case of the cerebellar cortex, where according to Marr

(1969) something analogous to the present theory actually occurs, the evidence cells are the granule cells, which are codon cells with $R \leqslant 7$. It will be pointed out in §6 that the cerebral neocortex contains cells which may be regarded as (R, θ)-codons with larger R. It is thought that the combined weight of these arguments constitutes sufficient grounds for studying in detail the setting up and performance of (R, θ)-codon cells, where the values of R and θ have various relations to the parameters of the code used on the set of input fibres $\{a_1, \ldots, a_N\}$.

4.3. *Codon neurotechnology*

4.3.0. *The possible need for codon formation*

At first sight, the use of codons virtually solves the problem of the neural representation of evidence functions. Provided the contact probability z from the afferent fibres $\{a_1, \ldots, a_N\}$ to the population $\mathfrak{P}$ of codon cells has the appropriate value, it remains only to set the thresholds of the codon cells in a suitable way (see §3.1).

The only possible problem with this scheme is that the evidence thus obtained may not have the required quality. The better the evidence required, the more specific the codon functions must be, and so the less frequently they take the value 1. If a roughly fixed number has to fire in order to provide an adequate representation of each input event, the size of the underlying population of codon cells has to be larger the better the evidence required. Unless special measures are taken, this might make it necessary in a particular case to provide a huge population of evidence cells, only a few of which are ever used. This difficulty can be avoided by using a special technique. It works by modifying just a few of the afferent synapses at a cell, so that a codon function of exactly the required sort is represented there. The process of determining to which codon a particular cell should respond is called *codon formation* at that cell.

The essence of codon formation is very simple. Let $\mathfrak{P}$ be a population of cells, each of which has R' afferent synapses. R' is such that a typical input event can expect to excite θ synapses at each cell of $\mathfrak{P}$, where θ is the θ of the (R, θ)-codons eventually required. The information which the codons have to represent arrives during a special *setting-up period* (§5.1.2), and only the synapses used during that time have any effective power later. This produces a population of codon cells such that only a few of the total number of afferent synapses have any power, but those few are the correct ones. The details are described fully in the following pages.

4.3.1. *Techniques for codon formation*

The three basic mechanisms for codon formation appear in figure 5. In (1) the afferent synapses are excitatory, and become ineffective if and only if there is post-without pre-synaptic activity. In (2), the synapses are composed of two parts: one excitatory and unmodifiable, and one initially ineffective, but which is facilitated by simultaneous pre- and post-synatic activity. The modifiable component is thus a

Hebb-modifiable synapse (Hebb 1949). The combination in one synapse of an unmodifiable excitatory component with a Hebb-modifiable component has an importance which was first noticed by Brindley (it appears at the *s*-cells in Brindley 1969). It is therefore proposed that such synapses be named *Brindley synapses*, to distinguish them from *Hebb synapses* which will taken be to possess the same modification conditions, but no unmodifiable excitatory component.

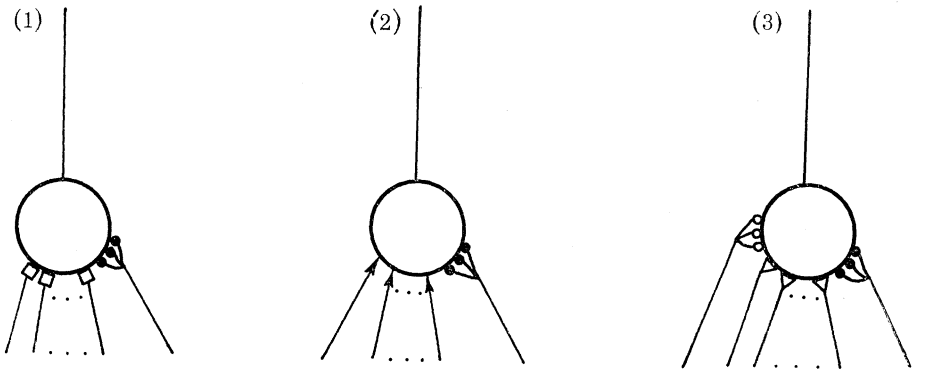

FIGURE 5. Three models for codon formation. (1) Uses synapses which are initially excitatory, but are modified to be ineffective by post- without pre-synaptic activity (open squares), (2) uses Brindley synapses (arrows), (3) uses Hebb synapses (open triangles) and a climbing fibre (open circles). All three have inhibitory synapses (filled circles) which set the cells' thresholds at an appropriate level.

In models (1) and (2), the cells also receive some inhibitory synapses which set their thresholds at the appropriate value. The equations governing the number of codons formed in any particular situation are those of §3.1: X is called the formation probability in those equations for this reason.

Case (3) is slightly different: this cell possesses an afferent fibre analogous to the cerebellar climbing fibre, and its ordinary afferent synapses are Hebb synapses, which are initially ineffective, and are modified by the conjunction of pre-synaptic and climbing fibre (or post-synaptic) activity. The climbing fibre is active only during the setting up period. The consequences of this model are slightly different from those of (1) and (2), for after setting up, *all* those synapses which were active during the setting up period will have been modified, not just those at a cell where a codon was successfully formed.

The conditions in which the codon cells may later be used are different for each of these models. In (1), there is no difficulty, since the irrelevant synapses have no power. In (3), the fact that all synapses active during the setting up period will have been modified may mean that an undesirably large number have been made excitatory. Methods (1) and (2) are in this sense more selective, and will tend to produce better evidence. In (2), during later use, the cell threshold has to be set so that activity in at least θ *modified* afferent synapses is required to discharge the cell. In all cases, the codon cell thresholds can be set at the appropriate level by using

sampling techniques—both of the afferent fibres and of the codon cell axons—in the same way as the cerebellar Golgi cells are thought to control the granule cell thresholds (Marr 1969).

4.3.2. *Model* (2) *preferred to model* (1)

Models (1) and (2) will produce evidence of the same quality in a given situation, but model (1) has an important disadvantage. If synaptic modification is an irreversible process, the process of codon formation in this model is a once and for all affair. The fact that all the synapses not involved in the first codon represented are thereby rendered ineffective means that the cell can never be used for more than one codon. This model essentially represented one codon by eliminating all other possibilities, and as such is unattractive. This is not true of model (2), where a synapse which is unused the first time could be used later on, if that became desirable. The model (2) needs slightly more complicated backing up by the inhibitory cells, since the level of inhibition necessary during codon formation both differs from that needed for recognition of codons already formed, and depends upon the number of codons already formed at that particular cell. This difficulty can be overcome if the inhibition level is set primarily by a count of the active codon cells, so it does not significantly affect the desirability of this model.

Model (3), like (2), does not suffer from the once-for-all disadvantage; but as pointed out in §4.3.1, is not strictly comparable with (1) since it forms evidence in a slightly different way.

4.3.3. *A problem with* (1) *and* (2)

In model (1), if synaptic modification is irreversible, each cell can represent only one codon. Hence the afferent synapses should not be modifiable all the time; the precious potential of a cell must be reserved for information for which it is worth being used. A similar point holds for model (2), since if the afferent synapses were permanently modifiable, any incoming information could cause the creation of codons. The point here is not that the first event rules out the rest, but that all are treated as indiscriminately valid. Since any input can create a codon if the anatomy allows it, the cell is no different in function from one where afferent synapses are unmodifiable excitatory. Therefore, for models (1) and (2), *the modifiable synapses involved must be modifiable only whilst that information for which codons are required is present in the afferent fibres.*

This difficulty arises in model (3) in a less acute form: the problem here is that something has anyway to specify when codon formation should take place. No difficulties arise with the hardware, since modification is geared to the climbing fibre activity; but climbing fibres cannot in general select the best cells.

4.3.4. *The solution using inhibition*

The only solution to this problem in models (1) and (2) which uses conventional ideas is to suppress the cells with inhibition until they are wanted. The alternative,

to excite them when they are wanted, is equivalent, but reduces (2) to an uninteresting variant of (3). This scheme would work until the first codon was formed, but would then fail in model (2): this is because inhibition cannot subsequently be maintained at these cells without their losing the ability to recognize the codons that have been formed at them. This defeats the object of the scheme.

4.3.5. *Another solution*

The alternative to this kind of solution is that the synapses genuinely should become modifible only at those times when codon formation is required. This is not as implausible an assumption as it might appear, since considerable organization has to take place before the formation of codons becomes necessary anyway. Codon formation takes place either when a new classificatory unit is formed, or when new evidence functions are added to an existing one. The decision about how to commit a piece of information to the neocortical store—whether as a new classificatory unit or as an association between existing ones—has to be taken on the basis of its relationship to other incoming events. It cannot in general be taken immediately: for example, it takes time for the mountainous structure of a probability distribution to become apparent.

This has the consequence that it is best to send all incoming information to a temporary associative store, where it is held and not altered. This is one point of Simple Memory theory (§5 and Marr 1971). When it becomes clear *how* a piece of information should be stored, it can be taken out and dealt with in the appropriate way. If, for example, it should be set up as a new classificatory unit, a location must be sought (the one with the most favourable pre-existing structure) and the information directed there for representation. The complete operation is so special and complex that the assumption, that a suitable delicate change in the chemical environment of the relevant codon cells accompanies the transmission there of the setting-up information, ceases to carry a special implausibility. The matter is discussed further in §5.1.2.

4.4. *Implementing the interpretation theorem*

4.4.0. *Preliminary assumptions*

The analysis of §4.2 suggested that codon functions are likely to be widely used as evidence functions. If they are, two conditions will hold, one about the input events, and one about the codons themselves. First, the input events for a particular output cell $\boldsymbol{\Omega}$ are likely to occupy a code of some fixed size L, say, on the input fibres $\{a_1, \ldots, a_N\}$. The reason for this is that if the input events have an arbitrary form, then codon functions of an arbitrary form have to be allowed. An arbitrary codon function is one which assigns the values 0 or 1 to a subset of $\{a_1, \ldots, a_N\}$: the codon functions we have met so far have assigned only the value 1. There is no objection in principle to the general codon function, but it is more difficult to build its neural representations, and much more difficult to model codon formation. It will therefore

be assumed, for the purposes of this section that the input events are events of size L over $\{a_1, \ldots, a_N\}$.

Secondly, all the codons associated with a given output cell $\boldsymbol{\Omega}$ are likely to be of about the same size. This is because only a small proportion of the codon cell population will be used for any single input event: these are chosen by selecting an appropriate codon cell threshold, and so come from the tail of a binomial distribution. The numbers of cells discovered in such a situation decreases sharply as the cells' thresholds rise, so that at any given threshold, the cells may to a first approximation be regarded as all having the same number of active afferents. Since the input events also will have the same size, all the codons connected with a given output cell $\boldsymbol{\Omega}$ may be regarded as having the same specifications. It will further be assumed that the actual codon cells which exist have been chosen randomly from the population of all such codon cells with those specifications.

These conditions are sensible also from another point of view, since the expected quality of evidence obtained from a codon depends upon its specifications. It was remarked in § 2 that the expected quality should be uniform for a given decision cell $\boldsymbol{\Omega}$, so this condition is likely to be fulfilled. Further, the randomness assumption means that problems about correlated evidence are avoided.

4.4.1. *Statement of the main result*

Suppose a set of (R, θ)-codons are chosen as evidence functions for diagnosis of the property Ω, and that these codons constitute a random sample from the set of all such codons. Suppose the input events have size L over $\{a_1, \ldots, a_N\}$: then an incomplete event specifies the values of less than L input fibres. It is shown that the interpretation of such an incomplete input may be carried out by taking a weighted sum of certain $P(\Omega|c_i)$ in a way analogous to the procedure for diagnosis of complete events. An estimate of this sum, for an incomplete input X, can be obtained in a real neural net by lowering the threshold of the codon cells until X causes activity in a significant number, and applying these signals to the output cell Ω in the usual way. Hence in a neural model where the codon cell thresholds are controlled by cells designed to maintain the number of active codon cells at a constant value, the interpretation of an incomplete event is a natural consequence of applying the event to the net.

There are two sources of error in this estimate: first, those codon cells with more active afferents than the current codon cell threshold will probably acquire an incorrect weighting of their corresponding value of $P(\Omega|c_i)$ at $\boldsymbol{\Omega}$; and secondly, the estimate is based on a sampling process. The first kind of error is alleviated by two facts: that most active codon cells have the same number of active afferents, only a very few having more (because the active cells come from the tail of a binomial distribution); and that those codon cells with more active afferents will be driven harder than the rest. This effect operates in the right direction to reduce the error. The inaccuracies from the second source are probably unimportant.

4.4.2. *Proof*

The interpretation theorem, §2.5, is concerned with the treatment of inputs in which the values of some of the fibres are undefined. In the present case, this corresponds to states where fewer than L of the input fibres $\{a_1, \ldots, a_N\}$ have the value 1. Let X be a subevent of the input event space $\mathfrak{X}$, and suppose that X specifies $X(a_i) = 1$, $1 \leqslant i \leqslant l < L$. Let $E_1, E_2, \ldots, E_J$ be the possible completions of X in $\mathfrak{X}$, so that each E_j $(1 \leqslant j \leqslant J)$ specifies that exactly L of the a_i have the value 1.

By the Interpretation Theorem,

$$\mathcal{P}(\Omega|X) = \sum_{j=1}^{J} P(E_j|X)\,\mathcal{P}(\Omega|E_j).$$

If nothing is known about $P(E_j|X)$, it must be assumed that $P(E_j|X) = 1/J$ all $1 \leqslant j \leqslant J$. Let $\mathfrak{C} = \{c_i|\ 1 \leqslant i \leqslant K\}$ be the set of all evidence functions for Ω over $\mathfrak{X}$. Then

$$\mathcal{P}(\Omega|E_j) = k^{-1}(E_j) \sum_{i=1}^{K} c_i(E_j) P(\Omega|c_i),$$

where $k(E_j)$ is the number of c_i with $c_i(E_j) = 1$, i.e. $k(E_j) = \sum_{i=1}^{K} c_i(E_j)$. Hence

$$\mathcal{P}(\Omega|X) = \sum_{j=1}^{J} J^{-1} \sum_{i=1}^{K} c_i(E_j) k^{-1}(E_j) P(\Omega|c_i).$$

Define the family of real-valued functions w_i, $1 \leqslant i \leqslant K$ on the set $\{E_1, \ldots, E_J\}$ by

$$\begin{aligned} \omega_i(E_j) &= 0 \quad \text{if} \quad c_i(E_j) = 0, \\ &= k^{-1}(E_j) \quad \text{if} \quad c_i(E_j) = 1. \end{aligned}$$

Then

$$\begin{aligned} \mathcal{P}(\Omega|X) &= \sum_{j=1}^{J} J^{-1} \sum_{i=1}^{K} w_i(E_j) P(\Omega|c_i) \\ &= J^{-1} \sum_{i=1}^{K} P(\Omega|c_i) \sum_{j=1}^{J} w_i(E_j). \end{aligned}$$

The operation of calculating $\mathcal{P}(\Omega|X)$ is thus equivalent to computing the weighted sum

$$\sum_{i=1}^{K} P(\Omega|c_i) \sum_{j=1}^{J} w_i(E_j);$$

the coefficient of $P(\Omega|c_i)$ is $\sum_{j=1}^{J} w_i(E_j)$, and we now study the value this takes. $\sum_{j=1}^{J} w_i(E_j)$ measures the weight with which $P(\Omega|c_i)$ contributes to the set of all possible completions of X in $\mathfrak{X}$. In a given completion, E_j, $P(\Omega|c_i)$ has a certain weight: it is zero if $c_i(E_j) = 0$ and if not, this weight is $1/k(E_j)$ where $k(E_j)$ is the size of the c_i-representation of E_j. Now the number $k(E_j)$ is a random variable obtained by adding the terms in the tail of a binomial distribution (see equation 3.1.1). Suppose k has distribution ν: then k^{-1} has distribution ν^{-1} say, with expectation $\overline{k^{-1}}$ ($\neq \bar{k}^{-1}$ in general), and variance σ (say). (Assume $k = 0$ with zero probability). The values of $k^{-1}(E_j)$ for different j are strictly speaking not independent, but if they were, the random variable $(1/n(c_i)) \sum_i c_i(E_j) k^{-1}(E_j)$ would have the same mean $\overline{k^{-1}}$, and variance $\sigma/\sqrt{\{n(c_i)\}}$, where $n(c_i)$ = the number of E_j with $c_i(E_j) = 1$.

The value of $\sigma/\sqrt{\{n(c_i)\}}$ does, however, give some guide to the variance of this random variable. It may be assumed that σ is small, since part of the function of the Golgi-type inhibitory cells which control the thresholds of the cells is to ensure a constant-sized representation for each input event E. The actual random variable described above will have a variance somewhere between σ and $\sigma/\sqrt{\{n(c_i)\}}$, but since σ is small, and the true value will be nearer $\sigma/\sqrt{\{n(c_i)\}}$, it may safely be assumed that its variance is small enough to be ignored.

Hence $\mathcal{P}(\Omega|X) = K^* \sum_i n(c_i) P(\Omega|c_i)$, where $n(c_i)$ = the number of E_j with $c_i(E_j) = 1$, and E_j completes X; K^* is some suitable normalizing constant.

Now $n(c_i)$ depends upon R, θ and r, where c_i is an (R, θ)-codon, and r is the number of afferent fibres active in X which are contained in $S(c_i)$, the support of c_i. In fact,

$$n(c_i) = \binom{N-W}{L-W} \sum_{x \geqslant 0} \binom{R-r}{\theta-r+x} \binom{N-R-W+r}{L-W-\theta+r-x},$$

the sum being taken until one factor reaches 0, and where

N = no. of input fibres,
L = no. of fibres active in each full sized input event,
W = no. of fibres active in X,
$\left.\begin{matrix} R \\ \theta \end{matrix}\right\}$ c_i is an (R, θ)-codon,
r = no. of fibres active in the support of c_i.

For $R = \theta$, $n(c_i)$ is primarily a function of r; call it $n(r)$.
Then

$$\frac{n(r+1)}{n(r)} = \frac{N-(W+R-r-1)}{L-(W+R-r-1)} > \frac{N-W}{L-W}.$$

For typical values, e.g.

$$N = 100, \quad L = 40, \quad W = 20, \qquad \frac{n(r+1)}{n(r)} > 4,$$

which illustrates the fact that those c_i with greater r have much more influence over $\mathcal{P}(\Omega|X)$ than those with smaller r.

The problem of estimating $\mathcal{P}(\Omega|X)$ from a family of (R, θ)-codons c_i is thus equivalent to taking the weighted average of $P(\Omega|c_i)$, where the weighting depends upon the number, r, of active input elements in the support of c_i. It will now be shown that this can be achieved by reducing the threshold of the cells for the (R, θ)-codon to some suitable lower value θ', which depends upon W, the size of X.

Two problems have to be solved when $\mathcal{P}(\Omega|X)$ is computed: first, enough c_i have to be used for the estimated answer to be reliable; and secondly, those c_i which are used have to be weighted in the correct way. It is assumed that the c_i are all (R, θ)-codons whose neural representation is effectively as shown in figure 4: it is immaterial whether this is achieved by models (1), (2) or (3) of figure 5. For an input X of size W, the probability of the cell's being active is

$$\pi(\theta') = 1 - \sum_{r=0}^{\theta'-1} \binom{R}{r} z^r (1-z)^{R-r},$$

where the cell has threshold θ', and $z = W/N$ (by analogy with 3.1.2). This is just the usual tail of a binomial distribution. Now as θ' decreases, the number of (R, θ)-codons which become active increases rapidly:

$$\frac{\pi(\theta')}{\pi(\theta'+1)} \doteqdot \frac{\theta'+1}{R-\theta'} \cdot \frac{N-W}{W};$$

while $\pi(\theta')$ is small, both $\theta'+1 > R-\theta'$ and $N > 2W$ will usually hold. Hence as the value of θ' is lowered, the number of c_i-cells which X fires increases very fast: so that the difference in θ' between having no cells active to having the usual number for a full event will only be of the order of 3 units of synaptic strength, and the great majority of the active c_i will have exactly θ' active afferent synapses.

The problem of the differential weighting of the $P(\Omega|c_i)$ can thus be alleviated as long as θ' does not lie far below the minimum number required to achieve the response of at least one c_i-cell. Provided the number of c_i-cells made active in this way is of the order of the number ordinarily excited by a full input event, enough evidence will be involved for the

estimate of $\mathcal{P}(\Omega|X)$ to be reliable. Strictly, all the c_i which could possibly take the value 1 on some completion of X should be consulted: but this number could be very large, and the problems of achieving the correct weighting become important. It is therefore much simpler to take an estimate using about the usual number of c_i.

Finally, it should be noted that if this is done, the $\boldsymbol{c}_i$-thresholds can be controlled by the same inhibitory cells as control their thresholds for normal input events, since it has already been shown that a circuit whose function is to keep the number of $\boldsymbol{c}_i$-cells active constant is adequate for this task. If this technique is used, those few $\boldsymbol{c}_i$-cells with more than θ' active afferents will have a higher firing rate than those with exactly θ'. Hence they will anyway be given greater weighting at the $\boldsymbol{c}_i$-cell. It would be optimistic to suppose this weighting would be exactly the correct amount, since the factor involved depends on the parameters N, L, W, R, θ, r; but the effect will certainly reduce the errors involved.

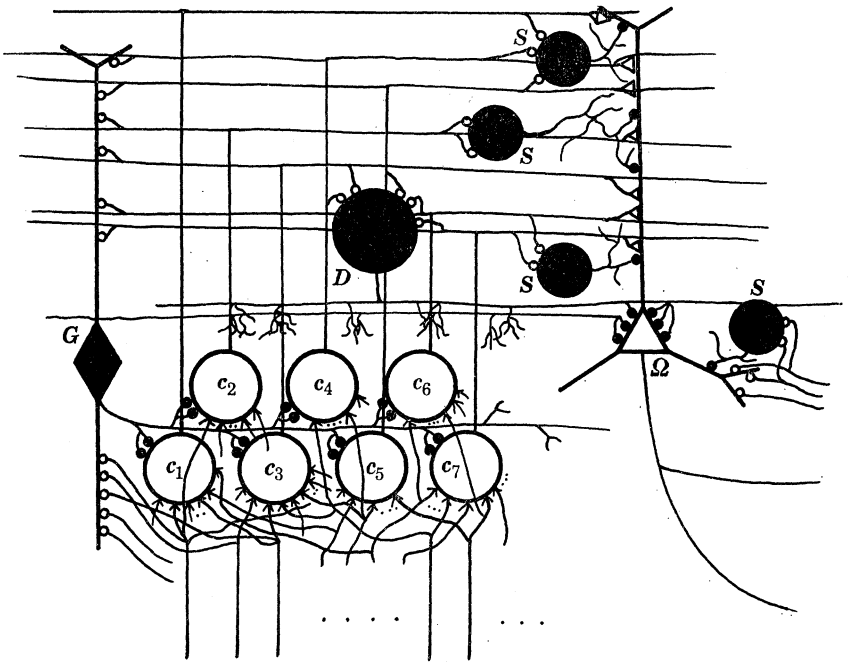

FIGURE 6. The basic neural model for diagnosis and interpretation. The evidence cells $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_7$ are codon cells with Brindley afferent synapses. The $\boldsymbol{G}$-cell controls the codon cell threshold: it uses negative feedback through its ascending dendrite to keep the number of codon cells active roughly constant. Its descending dendrite samples the input fibres directly, thus providing a fast pathway through which an initial estimate is made. The other cells and synapses are as in figures 3 and 5(2).

4.5. *The full neural model for diagnosis and interpretation*

The arguments of §§4.1 to 4.4 lead to the design of figure 6 for the basic diagnostic model for a classificatory unit. The afferent synapses to the $\boldsymbol{c}_i$-cells are excitatory, and may have been achieved by some suitable codon formation process: model (2) of figure 4 has been chosen for figure 6. The inhibitory cells $\boldsymbol{G}$ control the thresholds of the $\boldsymbol{c}_i$-cells, and their function is to keep the number of active $\boldsymbol{c}_i$-cells roughly constant. If they do this, the model automatically interprets input events which are

incomplete as well as those which are full-sized. The ***G***-cells are analogous to the Golgi cells of the cerebellum, and it is therefore natural to assume that, as in the case of those cells, the ***G***-cells can be driven both by the input fibres $\boldsymbol{a}_j$, and by the $\boldsymbol{c}_i$-cell axons. The final control should be exercised by the number of $\boldsymbol{c}_i$-cell axons active, but a direct input from the $\boldsymbol{a}_j$ axons would provide a fast route for dealing with a sudden increase in the size of the input event.

The $\boldsymbol{c}_i$ axons and the output cell **Ω** have been dealt with at length in §4.1. The cells ***S*** are the subtracting inhibitory cells, and the cells ***D*** provide the final division. The cell **Ω** is shown with two types of evidence cell afferent: one, through the $\boldsymbol{c}_i$-cells to the apical dendrites, and one (whose origin is not shown) to a basal dendrite.

In practice, the distribution of the $\boldsymbol{a}_j$ terminals, and the ***G***, ***D*** and ***S***-cell axons and dendrites will all be related. The kind of factor which arises has already been met in the cerebellar cortex for the Golgi and stellate cell axons and dendrites. Roughly, the more regular and widespread the input fibre terminals, the smaller the dendrites of the interneurons may be, and the further their axons may extend. Little more of value can be added to this in general, except that the exact most economical distributions for a particular case depend on many factors, and their calculation is not an easy problem.

§5. The discovery and refinement of classes

5.0. *Introduction*

There are three principal categories of problem associated with the discovery and refinement of classificatory units. They are the selection of the information over which a new unit is to be defined; the selection of a suitable location for its representation, together with the formation there of the appropriate evidence and output cells (formation in the information sense, not their physical creation); and the later refinement of the classificatory unit in the light of its performance.

The selection of information over which a new classificatory unit is to be defined depends, according to the Fundamental Hypothesis, upon the discovery of a collection of frequent, similar subevents in the existing coding of the environment. The difficulty of this task depends mainly on two factors: the *a priori* expectation that the fibres eventually decided upon would be chosen; and the time for which records have to be kept in order to pick out the subevents. The three basic techniques available are simple storage in a temporary associative memory, which allows collection of information over long periods; the associative access, which allows recall from small subevents, and hence eventually the selection of the appropriate fibres for a new unit; and the mountain climbing idea, which discovers the class once the population of fibres has been roughly determined. Only the third technique can be dealt with here.

The selection of a location for a new classificatory unit is simply a question of choosing a place where the relevant fibres distribute with an adequate contact

probability. The formation of evidence cells there is a problem which has already been discussed in §4: the formation of output cells is dealt with here.

Finally, the refinement problem arises because part of the hazard surrounding the formation of a new classificatory unit is that it is known in advance neither why it is going to be useful, nor of exactly what events it should be composed. When first created, therefore, the new classificatory unit is a highly speculative object, whose boundaries and properties have yet to be determined. The subsequent discovery of the appropriate boundaries (if such exist) is the refinement of the classificatory unit.

5.1. *Setting up the neural representation: sleep*

5.1.0. *Introduction*

It is convenient to begin with the second problem, of selecting a location and forming there a suitable neural structure. The reason is that the other two problems are best dealt with in the context of explicit neural models, and these are not complete enough until the apparatus necessary for the setting up problem has been incorporated. For the purposes of this section, it will therefore be assumed that the subevents which are to make up the new classificatory unit have been decided upon in advance, and are held in a store. The problem then reduces to that of discovering a suitable location, and creating there the appropriate evidence and output cells.

5.1.1. *Selecting a location*

The natural method of discovering a suitable location is to form a representation in *all* those places which are suitable. For this, the whole cortex is, so to speak, placed in a suitably receptive state, and in those regions where enough information is received, a representation is automatically set up. Later refinement will select for the most successful, and not all of the representations initially set up will survive.

This method has two important advantages: first, it removes the difficulties which arise in computing where the appropriate fibres gather together with a large enough contact probability. The discovery of these special locations is better left to the method suggested, whereby it is a natural consequence of their existence.

Secondly, the method allows the multiple formation of representations, which means that a single input can generate many different classes. There are often excellent grounds for categorizing information, and dealing with each category separately. For example, information about shape can profitably be classified separately from information about colour, and this could be implicit in the way the connexions are originally arranged. An area of cortex which received only information of a particular category would classify within that category. If many such areas existed, one piece of information could simultaneously cause classes in several categories to form. This is probably an important aspect of the solution to the partition problem §1.3.3, but one which relies on the rough genetic specification of the categories.

5.1.2. *Codon formation and sleep*

The problems of what evidence functions to form, and how to form them, have been discussed in §4. It may turn out never to be necessary to use codon formation, since this technique is essential only where a standard codon transformation, with unmodifiable excitatory synapses (Marr 1969), does not produce evidence of sufficient quality. The finer the classifications required, however, the better the quality of the evidence must be; and the more sophisticated they are, the less certain it becomes that genetic information can provide pre-formed codons of the right type: so if codon formation is used at all, it will be used more in higher than in lower animals.

In §4.3.5, it was decided that the most likely technique for codon formation used Brindley synapses which become modifiable only at those times when codon formation takes place. Arguments were set out there for the view that this assumption does not have a complexity which is disproportionate to those concerning the other operations which must take place at these times.

It was pointed out in §4.3.3 that when the afferent synapses to codon cells are modifiable, only that information for which new evidence functions are required should be allowed to reach these cells. In §4.3.5, it was shown that information from which a new classificatory unit is to be formed will often come from a simple associative store, not directly from the environment. In §5.1.1 it was argued that the most natural way of selecting a location for a new classificatory unit was to allow one to form wherever enough of the relevant fibres converge. This requires that potential codon cells over the whole cerebral cortex should simultaneously allow their afferent synapses to become modifiable. Hence, at such times, ordinary sensory information must be rigorously excluded. The only time when this exclusion condition is satisfied is during certain phases of sleep.

The tentative conclusion of the theory is therefore that some cerebral codon cells have Brindley afferent modifiable synapses, which only become modifiable during sleep. The firm conclusion of the theory is that if the locations for new classificatory units are selected by the method of §5.1.1; if there exist plastic codon cells in the cerebral cortex; and if they use Brindley afferent modifiable synapses; then these synapses are modifiable only during the correct phases of sleep. A consequence of this phenomenon for the learning characteristics of the animal as a whole is set out in §7.6.

5.1.3. *Output cell selection: generalities*

No methods have so far been proposed for the selection of output cells for classificatory units. The question was raised in §4.1 of whether more than one physical cell could profitably be used as the output for a single classificatory unit: it was concluded impracticable unless such cells formed independent representations.

The problem of output cell selection is therefore that of finding a single, hitherto unused cell whose dendrites are favourably placed to receive synapses from most of

the evidence cells created for the classificatory unit concerned. These codon cells will be clustered round the projection region of the relevant fibres, so the selection process has to work to choose a cell in the middle of that region. The methods available for cell selection are essentially the same as those described in §4.3 for codon formation (figure 5), but the arguments for and against each method are different in the present context. The methods are discussed separately.

5.1.4. *Output cell selection: particularities*

The final state of the output cell afferent synapses has been defined by the preceding theory: they must have strength which varies with $P(\Omega|c_i)$, each c_i. There is therefore not the distinction between different models for output cell selection that there was between models (1) and (2) of figure 4 for codon formation. If some model of this kind is used, the synapses must initially all have some standard excitatory power, which gradually adjusts to become $P(\Omega|c_i)$. The exact details of the way this happens will be the subject of §5.2, but the outline can be given here. First, the cell will fire only when a significant number of afferent synapses are active: so it will only be selected for a set of events most of which it can receive. If there exists a single collection of common, overlapping subevents in its input, this collection will tend to drive the cell most often, and those synapses not involved in this collection will decay relative to those which are. Hence the cell will perform a kind of mountain climbing of its own accord.

There are two possible arguments against this scheme: first, such a system can only work successfully if there is just one significant mountain in the probability space over the events it can receive. This makes it rather bad at selecting a particular mountain from several, and responding only to events in that; so the cell will not be very adept at forming a specialized classificatory unit unless it is fed data in a very careful manner. Secondly, some disquiet naturally arises over the conditions required for synaptic modification—that modification is sensitive to simultaneous pre- and post-synaptic activity. The Ω-cell dendrite will need to collect from a wide range of c_i-cell axons, and will therefore be much larger than the c_i-cell dendrites. In such circumstances, it is far from clear that these conditions are realizable. The most reasonable kinds of hypothesis for synaptic modification by a combination of activities in pre- and post-synaptic cells concern activities in *adjacent* structures, not elements up to 1 mm apart. There are therefore some grounds for being dissatisfied with model (1) of figure 7, even supposing the mountain-climbing details turn out in a favourable way.

The second model (figure 7 (2)) is based on some kind of climbing fibre analogue. It is of course not a direct copy of the cerebellar situation, since there can exist no cerebral analogue of the inferior olivary nucleus. It works thus: suppose there exists a single collection of common, overlapping input events in the input space of $\mathbf{\Omega}$, and let a_1 be one of the input fibres involved. Then most of the c_i used for such events will occur frequently with a_1, since a_1 is itself frequently involved in such events. Now suppose a_1, as well as reaching $\mathbf{\Omega}$ through orthodox evidence cells, also

drives a climbing fibre to **Ω**: then this will cause the modification of most of the c_i-cell synapses used in the collection of frequent events. The cell **Ω** will then be found to have roughly the correct values of $P(\Omega|c_i)$ for most of the c_i, and the final adjustments can be made by the same methods as were used in model (1).

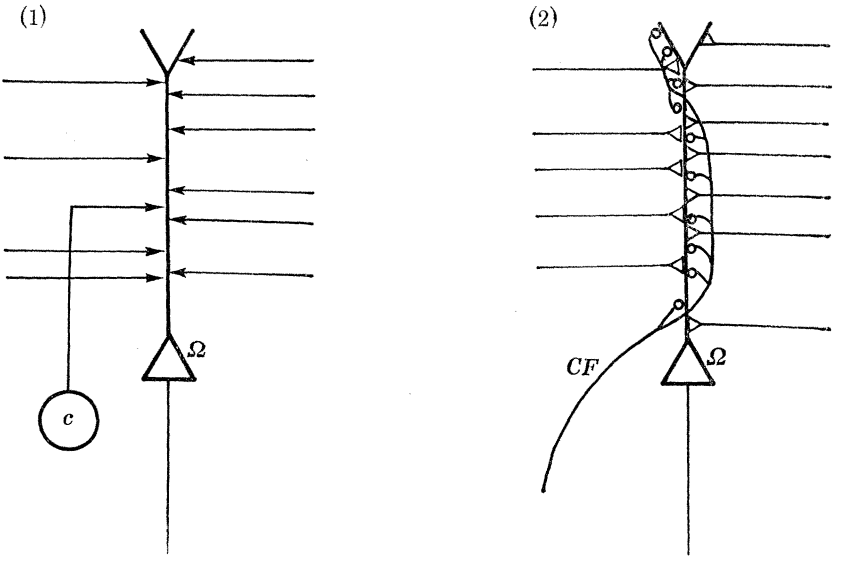

FIGURE 7. Two models for output cell selection. (1) Uses Brindley synapses, (2) uses Hebb synapses and a climbing fibre (*CF*).

In other words, the effect of tying modification conditions initially to a climbing fibre driven by something known to be correlated with the events of a mountain is to point the output cell **Ω** at that mountain. The use of a climbing fibre therefore, as well as eliminating difficulties about the implementation of synaptic modification, also removes the condition needed in model (1) that there should exist just one mountain in the event space to which **Ω** is exposed. With the climbing fibre acting as a pointer, there can be as many as you like: the only condition is that the more there are, the more specific the pointer has to be.

5.1.5. *Driving the climbing fibre*

The exact details of both these techniques will be analysed in §5.2, but before leaving this section, it is worth discussing the kind of way in which the climbing fibres may be driven. One possibility is the method already mentioned, where the climbing fibre is driven by one of the input fibres of the event space of **Ω**. This will do for many purposes, but it may not always provide a specific enough pointer.

The alternative method is to drive the **Ω**-cell by a climbing fibre whose action is more localized in the event space $\mathfrak{X}$ for **Ω** than the simple fibre a_1. In this scheme, the climbing fibre is driven by a cell near the **Ω**-cell, and one which consequently

fires only when there is considerable evidence-cell activity near **Ω**. This cell then acts as a more specific pointer than a simple fibre would, and is called an *output selector cell* (see figure 8).

It is an elementary refinement of this idea to have more than one climbing fibre attached to a given cell **Ω**, which then requires activity in several to be effective in causing synaptic modification. The crucial thing about the climbing fibre input is that it should provide a good enough rough guide to the events at which **Ω** should

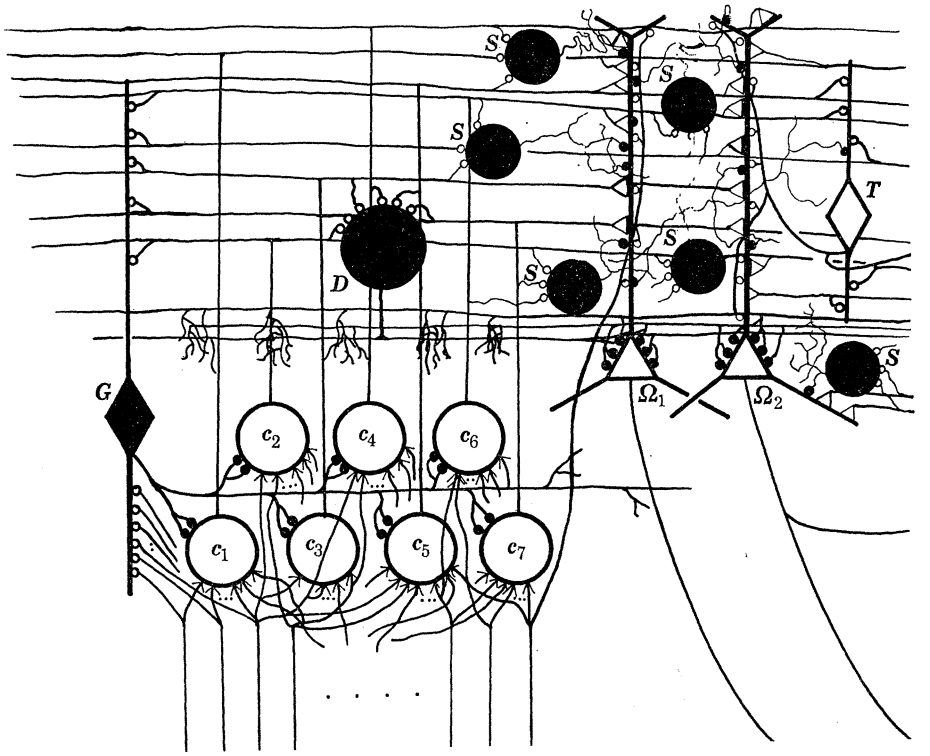

FIGURE 8. The fundamental neural model, obtained by combining the models of figures 6 and 7(2). Two climbing fibres are shown; one from an input fibre, and one from a nearby output selector cell ***T***.

look for **Ω** eventually to be able to discriminate a single mountain from the rest of its event space. It is important also to note that this kind of system can be used directly to discover new classificatory units. As long as no codon formation is required, climbing fibres can cause the discovery of mountains—i.e. new classificatory units—directly on the incoming information. Provided that the connectivity is suitable (i.e. that information gets brought together in roughly the correct way), new classificatory units will form without the need for any intermediate storage.

5.2. *The spatial recognizer effect*

5.2.0. *Introduction*

The process central to the formation of new classificatory units is the discovery that events often occur that are similar to a given event over a suitable collection of fibres. This was split in § 1.4 into the partition problem, which concerns the choice of roughly the correct collection of fibres; and the problem of selecting the appropriate collection of events over those fibres. The second part of this problem has been discussed in connexion with ideas about mountain climbing, and an informal description of the solution has been given in § 5.1. The essence of this solution is that an output cell performs the mountain climbing process naturally, and if started by a suitably driven climbing fibre in roughly the correct region of the event space to which it is sensitive, it will ultimately respond to the events in the nearby mountain. In this section, a closer examination of this process is made.

5.2.1. *Notation: the standard* (k, M)*-plateau*

The notation for this section will be slightly different from usual, since the output cell $\mathbf{\Omega}$ is sensitive to events E over $\mathfrak{X}$ only in terms of the evidence functions $c_i\,(1 \leqslant i \leqslant K)$. It is therefore convenient to construct the space $\mathfrak{Y}$ of all events of size k over the set $\{c_1, \ldots, c_K\}$. Each input event E over $\mathfrak{X}$ is translated into an event $Y = Y(E)$ over $\mathfrak{Y}$, and for the sake of simplicity, it will be assumed that each input event E causes exactly k of the c_i $(1 \leqslant i \leqslant K)$ to take the value 1. As far as $\mathbf{\Omega}$ is concerned, the events with which it has to deal occupy a code of size k over $\{c_1, \ldots, c_K\}$.

The c_i are imagined to be active in translating input events for many output cells other than $\mathbf{\Omega}$, and this allows the further simplifying assumption that all the c_i are 1 about equally often: that is $P(c_i) = P(c_j)$, all $1 \leqslant i \leqslant K$. Only those events which occupy k fibres concern $\mathbf{\Omega}$, and the relative frequencies of these are described by the probability distribution λ^* (say) over $\mathfrak{Y}$. λ^* is the probability distribution the environment induces over $\mathfrak{Y}$, and is derived from the input distribution λ over $\mathfrak{X}$. Both λ and λ^* have mountainous structure, but if $\mathfrak{Y}$ is obtained from $\mathfrak{X}$ by a codon transformation, the mountains in $\mathfrak{Y}$ are more separated than their parents in $\mathfrak{X}$.

The term 'mountain' has hitherto had no precise definition. It is not known exactly what kinds of distribution are to be expected, so some kind of general function has to be set up out of which all sensible mountains may be built. This is what motivates the following

Definition. Let μ be the probability distribution over $\mathfrak{Y}$ defined thus: let $M < K$, and for each $Y \in \mathfrak{Y}$

$$\mu(Y) = \binom{M}{k}^{-1} \quad \text{if} \quad Y \subseteq \{c_1, \ldots, c_M\},$$

$$\mu(Y) = 0 \quad \text{otherwise.}$$

Then μ is a *standard* (k, M)*-plateau* over $c_1, \ldots, c_M$.

That is, μ ascribes a constant value to the probability of every event which gives $c_i = 0$ for all fibres outside some chosen collection $\{c_1, \ldots, c_M\}$. The collection $\{c_1, \ldots, c_M\}$ is called the *support* of the plateau, and is written $S(\mu)$. A *simple mountain* μ^* is one that can be built up out of plateaux μ_i with nested supports: i.e.

$$\text{s.t. } \mu^* = w_1\mu_1 + w_2\mu_2 + \ldots + w_\rho\mu_\rho,$$

where $$\sum_{i=1}^{\rho} w_i = 1 \quad \text{and} \quad S(\mu_1) \supset S(\mu_2) \supset S(\mu_\rho).$$

In the absence of any better guesses about what kind of distributions should be studied, this section will deal with simple mountains. The fact that they can so simply be constructed from standard plateaux means that it is in fact enough to study the properties of standard plateaux. Further, we shall consider plateaux over the event space generated by the codon functions for a given classificatory unit, rather than plateaux over the event space generated by the input fibres. This is because the crucial operations occur at the output cell, which receives only evidence fibres.

5.2.2. *Climbing fibres and modification conditions*

Without loss of generality, it may be assumed that the output cell $\boldsymbol{\Omega}$ receives only one climbing fibre, which will be represented by the function $\phi(t)$ of time. ϕ cannot in general be regarded as a function from $\mathfrak{Y}$ to $\{0, 1\}$ since ϕ may take the value 1 at a time when there is no event in $\mathfrak{Y}$. Some kind of relation between ϕ and the events of $\mathfrak{Y}$ has to be assumed; it is that the conditional probability $P(\phi|c_i)$ is well-defined and independent of time.

The climbing fibre input to $\boldsymbol{\Omega}$ is closely related to the conditions for synaptic modification at $\boldsymbol{\Omega}$, but there are two possible views about the exact nature of this relation. One is that the climbing fibre is all-important in determining the strength of the synapse from $\boldsymbol{c}_i$ to $\boldsymbol{\Omega}$, and on this view, the strength varies with $P(\phi|c_i)$. The cell $\boldsymbol{\Omega}$ really diagnoses ϕ if this is so, but it will be shown in § 5.2.3 that if the structure of λ^* over $\mathfrak{Y}$ is appropriate, this will be adequate.

The other possible view is that ϕ acts as a pointer for $\boldsymbol{\Omega}$. On this model, the effect of ϕ is to set the values of the synaptic strengths at $P(\phi|c_i)$ initially. The true conditions for synaptic modification are simultaneous pre- and post-synaptic activity. It is a little difficult to see how the climbing fibre should be dealt with after it has set up the initial synaptic strengths, so in the theory of § 5.2.4, it is regarded simply as doing this, and is then ignored. This is an approximation, but seems the best one available. The true situation probably lies somewhere between those described in §§ 5.2.3 and 5.2.4.

5.2.3. *Mountain selection with* $P(\Omega|c_i) = P(\phi|c_i)$

Let $[p, q]$ denote the plausibility range of $\boldsymbol{\Omega}$. The state of $\boldsymbol{\Omega}$'s afferent synapses can be represented by the vector $\boldsymbol{\omega} = (\omega^1, \ldots, \omega^K)$ where $\omega^i = P(\phi|c_i)$, and it is assumed for this model that $\boldsymbol{\omega}$ is fixed—that the climbing fibre is the supreme determinant of

the synaptic strengths. Let $X \in \mathfrak{X}$. Then X has a representation as a vector $Y = (Y^1, \ldots, Y^K) \in \mathfrak{Y}$ with exactly k of the $Y^i = 1$, and all the rest zero. Let . denote the scalar product of vectors in the usual way: that is $\boldsymbol{\omega} . Y = \sum_i \omega^i Y^i$. Then the cell $\boldsymbol{\Omega}$ responds to X iff $\Sigma c_i(X) P(\phi|c_i) \geqslant kp$, i.e. iff $\boldsymbol{\omega} . Y \geqslant kp$. Hence $N_{\boldsymbol{\Omega}}$, the set of events to which $\boldsymbol{\Omega}$ responds, is given by

$$N_{\boldsymbol{\Omega}} = \{X | \boldsymbol{\omega} . Y \geqslant kp\}. \tag{1}$$

The following example shows how this may work adequately in practice. Let μ denote the standard (k, M)-plateau on $\{c_1, \ldots, c_M\}$, $M < K$, and let ν denote the standard (k, N)-plateau on $\{c_{S+1}, \ldots, c_{S+N}\}$ where $1 < S < M < S+N \leqslant K$. Suppose $\phi = c_1$. If the input distribution $\lambda^* = \mu$ we have

$$\begin{aligned} P(\phi|c_i) &= 1 \quad (i = 1) \\ &= (k-1)/(M-1) \quad (1 < i \leqslant M) \\ &= 0 \quad (M < i \leqslant K). \end{aligned}$$

If $\lambda^* = \nu$, we have $P(\phi|c_i) = 0$, all $i > 1$.

If $\lambda^* = \frac{1}{2}(\mu+\nu)$: $P(\phi|c_i) = 1 \quad (i = 1)$

$$\begin{aligned} &= (k-1)/(M-1) = \alpha(\text{say}) \quad (1 < i \leqslant S) \\ &= \frac{k(k-1)}{M(M-1)} \left(\frac{k}{M} + \frac{k}{N}\right)^{-1} = \beta(\text{say}) \quad (S < i \leqslant M) \\ &= 0 \quad (M < i \leqslant K). \end{aligned}$$

Hence if the lower limit of the plausibility range $[p, q]$ of $\boldsymbol{\Omega}$ is $p = k^{-1}(S\alpha + (k-S)\beta)$, the cell $\boldsymbol{\Omega}$ will respond to E if and only if $\mu(E) \neq 0$. Thus the output cell $\boldsymbol{\Omega}$ has selected the mountain μ from the distribution $\lambda^* = \frac{1}{2}(\mu+\nu)$ even though the climbing fibre ϕ did not. This is the crucial property which the system possesses.

In general, if $\phi = c_1$, ϕ will select the events of any plateau containing c_1 in its support, and can therefore be made (by suitable choice of p) to reject all events of other plateaux which do not fall into such a plateau.

The relation (1) can be used to construct the explicit condition that a climbing fibre ϕ can induce $\boldsymbol{\Omega}$ to respond to a particular set of events. If $\boldsymbol{\omega}$ is the climbing fibre vector $\boldsymbol{\omega} = (P(\phi|c_1), \ldots, P(\phi|c_K))$ and $N_{\boldsymbol{\Omega}} = \{X | \boldsymbol{\omega} . Y \geqslant kp\}$, then $\boldsymbol{\Omega}$ can select the events N out of $\{\mathfrak{X}, \lambda\}$ iff $\lambda(N_{\boldsymbol{\Omega}} \triangle N) = 0$; i.e. the probability under the input distribution λ that an event occurs which is in exactly one of $N_{\boldsymbol{\Omega}}$, N is zero.

5.2.4. *The spatial recognizer effect*

In the more general case, ϕ acts as a starting condition rather than permanently defining the strength of the synapse from c_i to $\boldsymbol{\Omega}$ $(1 \leqslant i \leqslant K)$. The subsequent strengths of these synapses depend on and only on $P(\Omega|c_i)$.

Write $P(\phi \,|\, c_i) = \omega_0^i$, $1 \leqslant i \leqslant K$ and let $\theta = kpP(c_i)$. Since $P(c_i) = P(c_j)$, all $1 \leqslant i, j \leqslant K$ (§5.2.1), the initial firing condition for $\boldsymbol{\Omega}$ is simply $\sum_i \omega_0^i c_i(X) \geqslant \theta$. As before write $\boldsymbol{\omega}_0 = (\omega_0^1, \ldots, \omega_0^K)$ as a vector: $\boldsymbol{\omega}_0$ defines the state of the afferent

synapses to $\mathbf{\Omega}$. If Y is the usual vector (consisting of 0's and 1's) which represents the event X over $\{c_1, \ldots, c_K\}$, the firing condition for $\mathbf{\Omega}$ is

$$\boldsymbol{\omega}_0 . Y \geqslant \theta. \tag{2}$$

The difference here is that $\boldsymbol{\omega}$ is now a variable. The point is that the vector $\boldsymbol{\omega}$ depends on the input distribution λ, *and* on those events to which (by (2)) $\mathbf{\Omega}$ responds. Define $N_\theta(\boldsymbol{\omega}_0) \subseteq \mathfrak{X}$ by $N_\theta(\boldsymbol{\omega}_0) = \{X | \boldsymbol{\omega}_0 . Y \geqslant \theta\}$. Define the new vector

$$\boldsymbol{\omega}_1 = (\omega_1^1, \ldots, \omega_1^K) \quad \text{by} \quad \omega_1^i = \sum_{X \in N_\theta(\boldsymbol{\omega}_0)} c_i(X)\lambda(X) \quad (1 \leqslant i \leqslant K). \tag{3}$$

That is, the co-ordinates ω_1^i of $\boldsymbol{\omega}_1$ are simply the projections onto the c_i of the restriction $\lambda | N_\theta(\boldsymbol{\omega}_0)$ of λ to $N_\theta(\boldsymbol{\omega}_0)$. Then $\boldsymbol{\omega}_1$ represents the state of the synapses from the c_i to $\mathbf{\Omega}$ if $\mathbf{\Omega}$ responds only to the events in $N_\theta(\boldsymbol{\omega}_0)$.

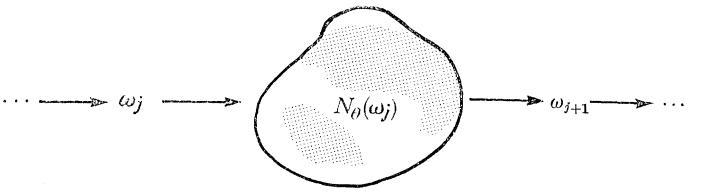

FIGURE 9. The state vector $\boldsymbol{\omega}_j$, which describes the strengths of the afferent synapses to the output cell $\mathbf{\Omega}$, determines the set $N_\theta(\boldsymbol{\omega}_j)$ of events to which $\mathbf{\Omega}$ will respond. This in turn determines a new state vector $\boldsymbol{\omega}_{j+1}$. Equilibrium occurs when $\boldsymbol{\omega}_j = \boldsymbol{\omega}_{j+1}$.

The situation is thus that in the state $\boldsymbol{\omega}_0$, the cell $\mathbf{\Omega}$ responds only to events in $N_\theta(\boldsymbol{\omega}_0)$: exposure to such events may be expected to change the state vector $\boldsymbol{\omega}_0$ into $\boldsymbol{\omega}_1$, from where the process is repeated. This generates a series of successive transformations of the state vector $\boldsymbol{\omega}$ for $\mathbf{\Omega}$, and this is called the *spatial recognizer effect* (see figure 9).

Theorem. The state vector achieves equilibrium iff there exists a j such that $\boldsymbol{\omega}_j = \boldsymbol{\omega}_{j+1}$.

Proof. In equilibrium, the set of events $N_\theta(\boldsymbol{\omega}_j)$ to which the cell $\mathbf{\Omega}$ responds specifies a state vector $\boldsymbol{\omega}_{j+1}$ such that $\lambda(N_\theta(\boldsymbol{\omega}_j) \triangle N_\theta(\boldsymbol{\omega}_{j+1})) = 0$: hence each co-ordinate of $(\boldsymbol{\omega}_j - \boldsymbol{\omega}_{j+1})$ is the projection onto a c_i of $\lambda | N_\theta(\boldsymbol{\omega}_j) \triangle N_\theta(\boldsymbol{\omega}_{j+1})$, and so is zero. Thus $\boldsymbol{\omega}_j - \boldsymbol{\omega}_{j+1} = \mathbf{0}$, and $\boldsymbol{\omega}_j = \boldsymbol{\omega}_{j+1}$.

In the simple example $\lambda^* = \frac{1}{2}(\mu + \nu)$ of §5.2.3, equilibrium is achieved in exactly one step. As already observed, $\boldsymbol{\omega}_0$ is defined by

$$\left.\begin{aligned} \pi\omega_0^i &= 1 \quad (i = 1) \\ &= \alpha \quad (1 < i \leqslant S) \\ &= \beta \quad (S < i \leqslant M) \\ &= 0 \quad (M < i \leqslant K) \end{aligned}\right\} \quad \text{where } \pi^{-1} = P(c_i), \text{ and is constant.}$$

For $p = k^{-1}(S\alpha + (k-S)\beta)$, the cell $\mathbf{\Omega}$ responds only to those events X with $\mu = 0$ which also have $\nu = 0$ so that $\boldsymbol{\omega}_1$ has the following specification.

$$\left.\begin{aligned} \pi\omega_1^i &= 1 \quad (1 \leqslant i \leqslant S) \\ &= \frac{1}{M}\left(\frac{1}{M} + \frac{1}{N}\right)^{-1} \quad (S < i \leqslant M) \\ &= 0 \quad (M < i \leqslant K) \end{aligned}\right\}$$

and $\boldsymbol{\omega}_1 = \boldsymbol{\omega}_2$. This result extends to any simple mountain μ^*, $\mu^* = w_1\mu_1 + \ldots + w_\rho\mu_\rho$, where $\phi = c_i \in S(\mu^*) = S(\mu_1)$ is an element in its support.

5.2.5. *A general characterization of the recognizer effect*

It is natural to seek some elegant way of describing the spatial recognizer effect. In the following informal argument, a characterization is given in terms of a search for steepest ascents in $\mathfrak{Y}$ under λ^*. This effectively puts a stop to any attempt to produce a necessary and sufficient condition that the starting state $\boldsymbol{\omega}_0$ should lead to a particular final state $\boldsymbol{\omega}^*$, since the general question depends upon the detailed structure of λ. The answer that it does if and only if a line of steepest ascent leads

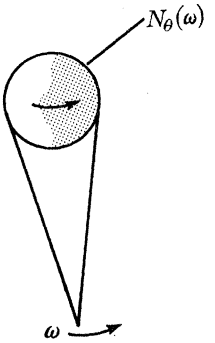

FIGURE 10. The state vector $\boldsymbol{\omega}$ determines the set $N_\theta(\boldsymbol{\omega})$ of events to which $\boldsymbol{\Omega}$ responds. The environmental probability distribution over $N_\theta(\boldsymbol{\omega})$ is stippled where it has non-zero values. $\boldsymbol{\omega}$ changes so as to tend to make the centre of gravity of this distribution coincide with the centre of $N_\theta(\boldsymbol{\omega})$. This is the principle behind $\boldsymbol{\Omega}$'s ability to perform a mountain-climbing operation.

there is probably its own neatest characterization. It is convenient to make the restriction that p, the lower limit to the plausibility range for $\boldsymbol{\Omega}$, is variable, and varies to keep the average amount of activity of $\boldsymbol{\Omega}$ constant; i.e. p is such that $\int_{N_\theta(\boldsymbol{\omega}_i)} \mathrm{d}\lambda$ is constant, for all response neighbourhoods $N_\theta(\boldsymbol{\omega}_i)$ (defined by equation (2) of §5.2.4). Write $N_\theta(\boldsymbol{\omega}) = \{X \mid Y.\boldsymbol{\omega} \geqslant \theta\}$ (see figure 10). $\boldsymbol{\omega}$ moves to $\boldsymbol{\omega}_1$ given by the projections onto the c_i of the restriction $\lambda|N_\theta(\boldsymbol{\omega})$ of λ to $N_\theta(\boldsymbol{\omega})$. (Compare (3) of §5.2.4.) Now $\boldsymbol{\omega}_1$ effectively measures the centre of gravity so to speak of the events in $N_\theta(\boldsymbol{\omega})$ since if $\boldsymbol{\omega}_1 = (\omega_1^1, \ldots, \omega_1^K)$, ω_1^i varies with the expected probability that $c_i = 1$ in $N_\theta(\boldsymbol{\omega})$ under λ. Since, in each event X of $N_\theta(\boldsymbol{\omega})$ exactly k of the c_i have the value 1, this means that the response area of $\boldsymbol{\Omega}$ moves towards that region of $N_\theta(\boldsymbol{\omega})$ which contains the closest, most common events. $\boldsymbol{\Omega}$ is attracted by both commoness, and by having many events close to one another all having non-zero probability. The way these two kinds of merit compete is approximately that the movement which maximizes the expectation of $\boldsymbol{\omega}$. Y over $N_\theta(\boldsymbol{\omega})$ is the one which is actually made: but the full result along these lines is complicated. In fact, the move is the one which has the best chance of maximizing this expectation.

Thus $\boldsymbol{\omega}$ moves to climb gradients in the scalar function $\boldsymbol{E}(\boldsymbol{\omega}.\, Y)$ taken over the response area defined by $\boldsymbol{\omega}$. A proof of this result will appear elsewhere.

5.3. *The refinement of a classificatory unit*

The refinement of a classificatory unit is the discovery of such appropriate boundaries as it might have. There are two kinds of information on which this process can be based: they are the frequencies of the subevents on which the unit is defined, and the correlation of instances of the unit with properties not included in its definition (i.e. support). The modification of a classificatory unit on the basis of its subevent distribution is called its *intrinsic refinement*, and has essentially been dealt with in §5.2: alteration made as a result of comparison with external properties is called *extrinsic refinement*, and will be discussed briefly here.

General extrinsic refinement requires a simple memory; but it basically consists of the same kind of mountain climbing techniques as intrinsic refinement. The only piece of the problem that can be discussed at the moment is the hardware needed for it. It is appropriate to deal with this now, since the necessary machinery must appear in the fundamental neural model.

There exist three main strategies for the extrinsic refinement problem: they are characterized by the change during refinement of the number of subevents to which the output cell $\boldsymbol{\Omega}$ will give a positive diagnosis. This number can increase, decrease, or remain about the same. The basic point is that the strategy which requires the number to decrease is the one which is easiest to implement, since it is easier to remove events from the response area of $\boldsymbol{\Omega}$ than to add them. This is because the only way of adding an event to $\boldsymbol{\Omega}$'s response area is by stimulating the climbing fibre. This needs some way of gaining access to the correct climbing fibre cell. The models of §5.1 for output cell selection make this difficult, since one of the key points in their design was the absence of a special climbing fibre for each output cell, and alternative schemes are unacceptably complicated.

The other possibility for adding events to Ω is to use an associational path to $\boldsymbol{\Omega}$ itself (for example, the basilar dendrite afferents of figure 5): but it was thought (§4.1.8) that the associational activity of the $\boldsymbol{\Omega}$-cell should not have this kind of ability to influence the strengths of the synapses arising from more direct inputs. Finally, there can be no guarantee that the existing evidence functions for Ω can cope with a new event.

Given these difficulties, it is natural to examine the possibility of refining a classificatory unit by eliminating inappropriate events from its response field. The main advantage such a method produces is that a general inhibitory influence acting over all output cells (in a particular region) can be used to alter values of $P(\Omega|c_i)$ for one particular $\boldsymbol{\Omega}$ in a way in which a general excitatory influence cannot. For suppose the event E is to be cut out: this must be achieved by allowing E to enter the c_i-cells for $\boldsymbol{\Omega}$ while preventing the formation of modification conditions at $\boldsymbol{\Omega}$ itself. If the chance that E should be interpreted in a cell near $\boldsymbol{\Omega}$ is small, this effect can be achieved by applying a general inhibitory signal to all the output cells in the region

containing **Ω**. Hence the only additional hardware this method requires is a fairly non-specific inhibitory input to all output cells. This does not appear in figure 8, since its derivation from the theory is less firm than that of the other elements which appear there.

§ 6. Notes on the cerebral neocortex

6.0. *Introduction*

The present theory receives its most concrete form in the neural model of figure 8. In this section, the fine structure of the cerebral cortex is reviewed in the light of that model. Anyone familiar with the present state of knowledge of the cerebral cortex will anticipate the sketchy nature of the discussion, but enough is probably known to enable one to grasp some at least of the basic patterns in the cortical design.

It need scarcely be said that cerebral cortex is much more complicated than that found in the cerebellum. Nothing of note has been added to the researches of Cajal (1911) until comparatively recently (Sholl 1956; Szentágothai 1962, 1965, 1967; Colonnier 1968; Valverde 1968), because Cajal's work was probably a contribution to knowledge to which significant additions could be made only by using new techniques. Degeneration methods have since been developed, and the electron microscope has been invented; so there is now no reason in principle why our knowledge of the cerebral cortex should not grow to be as detailed as that we now possess of the cerebellar cortex. It is, as Szentágothai (1967) has remarked, a Herculean undertaking; but it is within the range of existing techniques.

6.1. *Codon cells in the cerebral cortex*

6.1.1. *The ascending-axon cells of Martinotti*

The main source of information for this section is the description by Cajal (1911) of the general structure of the human cerebral cortex. The codon cells of the cerebellum are, according to Marr (1969), the granule cells, whose axons form the parallel fibres. The basic neural unit of figure 8 has analogies with the basic cerebellar unit (one Purkinje cell, 200000 granule cells, and the relevant stellate and Golgi cells, in the cat), so it is natural to look for a similar kind of arrangement in the cerebral cortex.

The first point to note is that cerebral cortex, like cerebellar cortex, has a molecular layer. According to Cajal (p. 521) this has few cells, and consists mostly of fibres. The dendrites there are the terminal bouquets of the apical dendrites originating from pyramidal cells at various depths. Most pyramidal cells, and some other kinds, send dendrites to layer I, so there is a clear hint in this combination that some such cells may act as output cells. The great need is for the axons of the molecular layer to arise mainly from cells which may be interpreted as codon cells. Cajal himself was unable to discover the origins of the axons of the molecular layer, and probably believed they came mainly from the stellate cells there. The problem

was unresolved until Szentágothai (1962) invented a technique for making small local cortical ablations without damaging the blood supply, and was at last able to determine the true origin of these mysterious fibres. It is the ascending-axon cells of Martinotti, which are situated mainly in layer VI in man.

This fundamental discovery showed that the analogy with cerebellar cortex is not empty, for the similarity of the ascending-axon cells of Martinotti to cerebellar granule cells is an obvious one. There are, however, notable differences; for example, the Martinotti cells are much larger than the cerebellar granules; and in sensory cortex, primary afferents do not terminate in layer VI.

The interpretation of Martinotti cells as cerebral codon cells raises five principal points, which will be taken separately. The first is the cells of origin of their excitatory afferent synapses. There is unfortunately rather little information available about this, but it appears from Cajal's description that the following sources could contribute fibres:

(i) The collaterals of the pyramidal cells of layers V, VI and VII.
(ii) Descending axons from the pyramids of IV.
(iii) Collaterals of fibres entering from the white matter.
(iv) Local stellate cells.

It would best fit the present theory if intercortical association fibres formed their main terminal synapses with these cells, and the collaterals of the pyramidal cells in layers V to VII were relatively unimportant. There is some evidence that association fibres tend to form a dense plexus in the lower layers of the cortex (Nauta 1954; and Cajal 1911, pp. 584–5).

The second point is that the Martinotti cells would have to have inhibitory afferent synapses driven by the equivalent of the ***G***-cells which appear in figure 8. The effect of these synapses should be subtractive rather than divisive, so that to be consistent with the ideas about inhibition expressed in §4.1 on output cell theory, the synapses from the ***G***-cells should be distributed more or less all over the dendrites of the Martinotti cells. (There is some evidence that this is so for certain cells of layer IV in the visual cortex of cat (Colonnier 1968), but it rests upon an as yet unproved morphological diagnosis of excitatory and inhibitory synapses.) This is in direct contrast to what the theory predicts for output cells, a distinct fraction of whose afferent inhibitory synapses should be concentrated at the soma.

The third point concerns the possible independence of the dendrites of Martinotti cells. These cells commonly have a quite large dendritic expansion, and it may be unreasonable to expect much interaction between synapses on widely separated branches. The effect, if their afferent synapses are unmodifiable, is to enable the cell as a whole to operate as the logical union of $m\,(R, \theta)$-codons (where m is the number of independent dendrites) instead of as a single (mR, θ)-codon: the advantage of this is a better quality of evidence function.

The fourth point concerns the possibility that the excitatory afferent synapses to Martinotti cells may be modifiable: this has been discussed in §5.1.2. If these synapses are Brindley synapses, then the dendrites may be independent from the

point of view of synaptic modification, as well as in the way described in point three. If there is some kind of climbing fibre arrangement, the fibres must be driven from some external source, and must be allowed to operate only when codon formation is required. The second possibility could allow the modification condition to operate simultaneously over the whole cell. It has been seen, however, that climbing fibres are unlikely to be used. If location selection proceeds as described in §5.1.1, the Martinotti afferent synapses are modifible only during the correct phase of sleep.

Fifth and last, it is a simple consequence of the present theory that Martinotti cells should be excitatory, and should send axons to synapse with five types of cell: the output cells, whose ordinary excitatory afferent synapses are modifiable; the two types (***S*** and ***D***) of inhibitory cell; the Martinotti threshold controlling cells, the ***G***-cells; and perhaps output cell selector cells, whose axons terminate as climbing fibres on output cells. A Martinotti axon may itself under certain circumstances terminate as a climbing fibre as well as making crossing-over synapses with output cells; but this possibility may be excluded for developmental reasons.

6.1.2. *The cerebral granule cells*

In layer IV of granular cerebral cortex, there are found a large number of small stellate cells, 9 to 13 μm in diameter, whose fine axons end locally. This layer is especially well developed in primary sensory cortex, where it sees the termination of the majority of the afferent sensory fibres. It has long been believed that such fibres synapsed mainly with the granule cells (Cajal 1911). Szentágothai (1967) has, however, pointed out that many sensory afferents in fact terminate as climbing fibres on the dendritic shafts passing through IV, and believes this may be an important method of termination.

Valverde (1968) has made a quantitative study of the amount of terminal degeneration in the different cortical layers of area 17 of mouse after enucleation of the contralateral eye, and has demonstrated that about 64 % occurs in layer IV, the other principal contributions being from the adjacent layers III and V. In view of the abundance of granule cells in layer IV, it is difficult to imagine that the afferent fibres never synapse with them, and so likely that the traditional view is correct. There can be no doubt that afferents also terminate as climbing fibres, and the possibility that both these things happen fits very neatly with the predictions of the present theory.

These views support the interpretation of the granule cells as codon cells, in which case the remarks of §6.1.1 about Martinotti cells may be applied to them. An interesting characteristic of granule cells is that they are often very close to raw sensory information, in a way in which the Martinotti cells are not. They will therefore not support classificatory units which rest on much preceding cortical analysis—that is, classificatory units for which, if it occurs at all, codon formation is most likely to be used. The theory therefore contains the slight hint that the Martinotti cells may be the plastic codon cells, and the granule cells the pre-formed

codon cells. The consequence of this would be that the Martinotti cells have modifiable afferent excitatory synapses, while the granule cells have unmodifiable afferent synapses.

6.2. *The cerebral output cells*

The present theory requires that candidates for output cells should possess the following properties:

(i) A dendritic tree extending to layer I and arborizing there to receive synapses from Martinotti cells.

(ii) An axon to the white matter, perhaps giving off collaterals.

(iii) Inhibitory afferent synapses of two general kinds: one, fairly scattered over the main dendrites, and performing the subtractive function; the other clustered over the soma, performing the division.

(iv) Climbing fibres over their main dendritic trees.

(v) A mixture of modifiable and unmodifiable afferent synapses. Those synapses from codon cells—Martinotti and granule cells—should initially be ineffective (or have some fixed constant strength), but should be facilitated by the conjunction of pre-synaptic and post-synaptic (or possibly just climbing fibre) activity, so that the final strength of the synapse from $\boldsymbol{c}$ to $\boldsymbol{\Omega}$ varies with $P(\Omega|c)$. These synapses should certainly be modifiable during the course of ordinary waking life, and should probably be permanently modifiable.

The cortical pyramidal cells of layers III and V are the most obvious candidates for this rôle. According to Cajal (1911), they satisfy (i), and (iv), and (ii) (Szentágothai 1962). The evidence for (iii) is indirect, but these cells receive axosomatic synapses of the basket type, and these have been shown to be inhibitory wherever their action has been discovered, (in the hippocampus (Anderson, Eccles & Løyning 1963), and the cerebellum (Anderson, Eccles & Voorhoeve 1963)). Various kinds of short axon cell exist in the cortex; there are probably enough to perform the subtraction function (§6.4).

The axon collaterals of these pyramidal cells could perform two functions. Either they can themselves act as input fibres to nearby Martinotti cells; this would enable two successive classifications to be performed in the same region of cortex. Or they could act as association fibres, synapsing with the basilar dendrites of neighbouring output cells. This would be useful if nearby cells dealt with similar information, but not necessarily useful otherwise (Marr 1971).

6.3. *Cerebral climbing fibres*

One of the crucial points about the output cells is that they should possess climbing fibres. The various possible sources of these were discussed in §5, where it was stated that there might be two origins—afferent fibres themselves, and cells with a local dendritic field.

The first observation of cerebral climbing fibre cells was made by Cajal (1911), who describes certain cells with double protoplasmic bouquet, as follows. 'The axon filaments [of these cells] are so long that they can extend over the whole thickness of

the cortex, including the molecular layer.... If one examines closely one of the small, parallel bundles produced by the axons of these cells, one notices between its tendrils an empty, vertical space which seems to correspond in extent to the dendritic stem of a large or medium pyramidal cell. Since the axon of one of these double-dendrite neurons can supply several of these small bundles, it follows that it can come into contact with several pyramidal cells,' (pp. 540–541).

Cajal saw these fibres only in man, but Valverde (1968) has beautiful photomicrographs of some coursing up the apical dendrite of a cortical pyramidal cell of the mouse, so they clearly exist in other animals. Szentágothai (1967) has found that various types of cell can give rise to such fibres, and remarks that specific sensory afferents often terminate in this way.

The cortical cells which give rise to climbing fibres have been called output cell selectors. The theory requires that they possess a rather nonspecific set of afferents, so that those cells in the centre of an active region of the cortex receive most stimulation. Such cells may also possess afferent inhibitory synapses to prevent their responding to small amounts of activity.

The present theory does not favour the view that cells other than output cells should possess climbing fibres, but it does not absolutely prohibit it.

6.4. *Inhibitory cells*

The basic theoretical requirements for inhibition in the cerebral cortex would be satisfied by having three types of inhibitory cell. Two should act upon the output cells, one synapsing on the dendrites, and one on the soma; and one, the analogue of the cerebellar Golgi cells, on the codon cells.

6.4.1. *The subtractor cells*

The first place in which to look for inhibitory cells for the subtraction function is the molecular layer I, where the Martinotti axons meet the pyramidal cell dendrites. This layer does contain some cells: it is wrong to believe that it consists of nothing but axons and dendrites. Cajal remarks upon the abundance of short axon cells there, stating that in number and diversity they achieve their maximum in man. He distinguishes (pp. 524–525) four main types; ordinary, voluminous, reduced, and neurogliaform. The last are like the dwarf stellate cells which appear frequently in other cortical layers.

The short axon cells can be interpreted as performing the role of subtraction on the output cell dendrite. They and their homologues are common throughout the cortex. The small size and great complexity of many of their axons and dendrites enable them to assess accurately the amount of fibre activity in their neighbourhood, so it does not require undue optimism to imagine that they can provide about the correct amount of inhibition. For this purpose, the more there are of such cells, the smaller and more complex their axonal and dendritic arborization, the more accurate will be their estimates of the amount of inhibition required. The neurogliaform cells therefore seem most suited to this task.

6.4.2. *The division cells*

The requirements of cells providing inhibition at a pyramidal cell soma for the function of division are different. Their action is concentrated in one place, and does not need to be accurately balanced over the dendritic field in the way that the subtraction inhibition must. The division inhibition can therefore be provided by a sampling process with convergence at the soma. The details of this sampling must depend on the distribution to the Martinotti and granule cells of the afferent fibres, and are based on the same principles as govern the distribution of the cerebellar basket cell axons.

There is no doubt that the pyramidal cells of layers III and V possess basket synapses (Cajal 1911), but Cajal does not describe them for those of layer II, which otherwise look like output cells. Colonnier (1968) has however studied the pyramids of II in area 17 in some detail, and has shown that, while synapses on the somas of these cells are not densely packed, they do exist, and are exclusively of the symmetrical type with flattened vesicles. It would be interesting to have some comparative quantitative data about somatic synapses on pyramidal cells of different layers in the cortex.

6.4.3. *The codon cell threshold controls*

The control of the Martinotti and granule cell thresholds requires an inhibitory cell which, like the cerebellar Golgi cell, is designed to produce a roughly constant amount of codon cell activity. There are various short axon cells in layers IV and VI which might perform this rôle, but no evidence available about the cells to which they send synapses. The obvious candidates in IV are the dwarf cells (Cajal 1911, p. 565) and perhaps the horizontal cells; and in VI, the dwarf cells and stellate cells with locally ramifying axon. For the control of Martinotti cell thresholds, it seems probable that the device of an ascending dendrite should be used to assess the amount of activity in the molecular layer. This could be done, for example, by an inhibitory pyramidal or fusiform cell with basilar and ascending dendrites, and locally arborizing axon. Such a cell would possess no climbing fibre, nor any modifiable afferent synapses. There exist various fusiform cells in layers VI and VII which might do this, but there is too little data available to know for certain.

6.5. *Generalities*

The theory expects output cells to fire at different frequencies, and it expects output cells at one level to form the input fibres for the next. It is therefore implicit in the theory that input fibres $a_i(t)$ should take values in the range [0,1], and should not be restricted simply to the values 0 and 1. The theory has been developed here only for the simple case of binary-valued fibres. Its extension to the more general case is a technical matter, and will be carried out elsewhere.

Finally, it is unprofitable to attempt a comprehensive survey of cortical cells at this stage: neither the theory nor the available facts permit more than the barest

sketch. It is most unsatisfying to have to give such an incomplete series of notes, and I write these reluctantly. It does, however, seem essential to say something here. It both illustrates how the theory may eventually be of use, and indicates the kind of information which it is now essential to acquire. More notes on the cerebral cortex will accompany the Simple Memory paper, but until then, it seems better to err on the side of reticence than of temerity.

§ 7. Neurophysiological predictions of the theory

7.0. *Introduction*

In this section are summarized the results which are to be expected to hold if the theory is correct, together with an assessment of the firmness with which the individual predictions are made. The firmness is indicated by superscripted stars accompanying the prediction, the number of stars increasing with the certainty of the statement they decorate. Three stars*** indicates a prediction which, if shown to be false, would disprove the theory: two stars** indicates that a prediction is strongly suggested, but that remnants of the theory would survive its disproof: one star* indicates that a prediction is clear, but that its disproof would not be a serious embarrassment, since other factors may be involved; and no stars indicates a prediction which is strictly outside the range of the theory, but about which the theory provides a strong hint.

7.1. *Martinotti cells*

Each Martinotti cell should have many inputs***, mainly from intercortical association fibres**, which should terminate by means of excitatory synapses***. Each should also have inhibitory inputs***, subtractive in effect** and therefore widely distributed over the dendrites**. These should be driven by local cells*** with locally arborizing axon***, designed to keep the amount of Martinotti cell activity evoked by different inputs roughly constant**.

Excitatory Martinotti cell afferent synapses are probably modifiable*, and if they are modifiable, they are probably Brindley synapses*, becoming modifiable only during the correct phases of sleep*. If location selection proceeds as in § 5.1.1, and if these synapses are modifiable, then they are modifiable only during the correct phases of sleep***. Martinotti cell dendrites are probably independent.

The output from these cells is excitatory***, and goes to output (pyramidal) cells*** through modifiable synapses***, three** kinds of inhibitory cells*** through unmodifiable synapses***, and to output selector cells** through unmodifiable synapses.

7.2. *Cerebral granule cells*

These cells fall broadly into the same class as Martinotti cells, and the predictions concerning them are the same, with the following exceptions. Their input is mainly more direct than that of the Martinotti cells, and should (because of their smaller

size) come from thalamo-cortical rather than cortico-cortical projections. They probably do not have modifiable afferent synapses. In the sensory projection areas, where afferents are known to terminate in layer IV, these afferents should form the main source of excitatory synapses on the granule cells*.

7.3. *Pyramidal cells*

The pyramidal cells of layers III and V, and probably also those of layer II, are interpreted as output cells, in the sense of the theory. On the assumption that this is correct, they receive two kinds of excitatory synapses**, and two kinds of inhibitory synapses**. The majority of afferent synapses comes from Martinotti and granule cells**, almost all such cells making not more than one synapse with any given pyramidal cell**. These synapses are either Hebb or Brindley type modifiable synapses***. The strength of the synapse form the codon cell $\boldsymbol{c}$ to the output cell $\boldsymbol{\Omega}$ stabilizes at the value $P(\Omega|c)$**. (This receives only two stars, since there may be a workable all-or-none approximation to this value.) These synapses should be modifiable during the course of ordinary waking life***, and probably during sleep as well*. All other afferent synapses described here are unmodifiable***.

If the dendrite is large, there exists a second excitatory input in the form of a climbing fibre**. If there is no climbing fibre present, the other excitatory afferent synapses must be Brindley synapses***. The climbing fibre input, if it exists, can produce the conditions for synaptic modification in the whole dendrite simultaneously***, but it is subsequently not the only input able to do this*.

There are two kinds of inhibitory input to the cell**: one scattered, which has the effect of performing a subtraction**, and one clustered at the soma, performing the division**. At least one of these functions is performed***, but the all-or-none approximation would require only one. Both essentially estimate the number of afferent synapses from codon cells active at the cell***.

The output from these cells is excitatory if it forms the input to a subsequent piece of cortex**. Their axon collaterals synapse with neighbouring output and Martinotti cells.

7.4. *Climbing fibres*

These are present only on output cells*. The climbing fibre at a given pyramidal cell provides an accurate enough pointer for that cell for the spatial recognizer effect to take over and make the cell a receptor for a classificatory unit***. Climbing fibres are excitatory***, if used for this purpose.

7.5. *Other short axon cells*

Many of the short axon cells which are not codon or climbing fibre cells are inhibitory***. The theory distinguishes three principal kinds**. Subtractor cells sample the activity of codon cell axons near local regions of dendrite**, and send inhibitory synapses to those regions**. These have a subtractive effect**. Division cells, the basket cells, are inhibitory**; and so are cerebellar Golgi cell analogues, which keep the amount of codon cell activity about constant**.

The granule cell threshold controls receive excitatory*** synapses from either the granule cell excitatory afferents, or the granule cell axons***, and perhaps from both*. They send inhibitory synapses to the granules themselves***, and these synapses are scattered over the granule cell soma and dendrites**. The Martinotti cell threshold controls receive excitatory*** synapses either from the Martinotti afferents, or from the Martinotti axons***. In view of the length of the Martinotti axons, they probably receive from both**, and therefore have an ascending dendritic shaft**. Layers VI and VII contain fusiform cells which could be Martinotti cell threshold controllers.

The axonal and dendritic distributions of the inhibitory cells of the cortex depend on the distributions of the afferents, and of the codon cell axons, in a complicated way.

7.6. *Learning and sleep**

This section as a whole receives one star, but if location selection proceeds as in §5.1.1, and if there exist plastic codon cells, then it receives three stars. The truth of these conditional propositions cannot be deduced from the available data. Star ratings within the section are based on the assumption that both propositions are true.

Sleep is a prerequisite for the formation of some new classificatory units***. The construction of new codon functions for high level units***, and perhaps the selection of new output cells, takes place then, though the latter can** occur, and probably usually does*, during waking.

Let $\mathfrak{I}_1$ and $\mathfrak{I}_2$ be two collections of pieces of information such that many of the spatial relations present in $\mathfrak{I}_2$ appear frequently in $\mathfrak{I}_1$, and have not previously appeared in the experience of an animal. The animal is exposed to $\mathfrak{I}_1$, and then to $\mathfrak{I}_2$. If the exposures are separated by a period including sleep, the amount of information the animal has to store in order to learn $\mathfrak{I}_2$ is less than the amount he would have to store if the exposures had been separated by a period of waking***. This is because the internal language is made more suitable during the sleep, by the construction of new classificatory units to represent the spatial redundancies in $\mathfrak{I}_1$. The recall of $\mathfrak{I}_1$ itself is not improved by sleep**.

Conversely, if this effect is found to occur, some codon cells have modifiable synapses**.

I wish to thank especially Professor G. S. Brindley, F.R.S., to whom I owe more than can be briefly described; Mr S. J. W. Blomfield, who made a number of points in discussion, and who proposed an idea in §1.5; Professor A. F. Huxley, F.R.S., for some helpful comments; and Mr H. P. F. Swinnerton-Dyer, F.R.S., for various pieces of wisdom. The embryos of many of the ideas developed here appeared in a Fellowship Dissertation offered to Trinity College, Cambridge, in August 1968: that work was supported by an MRC research studentship. The work since then has been supported by a grant from the Trinity College Research Fund.

References

Anderson, P., Eccles, J. C. & Løyning, Y. 1963 Recurrent inhibition in the hippocampus with identification of the inhibitory cell and its synapses. *Nature, Lond.* **197**, 540–542.

Anderson, P., Eccles, J. C. & Voorhoeve, P. E. 1963 Inhibitory synapses on somas of Purkinje cells in the cerebellum. *Nature, Lond.* **199**, 655–656.

Barlow, H. B. 1961 Possible principles underlying the transformations of sensory messages. In *Sensory Communication* (Ed. W. A. Rosenblith), pp. 217–234. MIT and Wiley.

Blomfield, Stephen & Marr, David 1970 How the cerebellum may be used. *Nature, Lond.* **227**, 1224–1228.

Brindley, G. S. 1969 Nerve net models of plausible size that perform many simple learning tasks. *Proc. Roy. Soc. Lond.* B **174**, 173–191.

Cajal, S. R. 1911 *Histologie du Système Nerveux* **2**. Madrid: CSIC.

Colonnier, M. 1968 Synaptic patterns on different cell types in the different laminae of the cat visual cortex. An electron microscope study. *Brain Res.* **9**, 268–287.

Eccles, J. C., Ito, M. & Szentágothai, J. 1967 *The cerebellum as a neuronal machine*. Berlin: Springer-Verlag.

Hebb, D. O. 1949 *The organisation of behaviour*, pp. 62–66. New York: Wiley.

Hubel, D. H. & Wiesel, T. N. 1962 Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* **160**, 106–154.

Jardine, N. & Sibson, R. 1968 A model for taxonomy. *Math. Biosci.* **2**, 465–482.

Jardine, N. & Sibson, R. 1970 The measurement of dissimilarity. (Submitted for publication.)

Kendall, D. G. 1969 Some problems and methods in statistical archaeology. *World Archaeology* **1**, 68–76.

Kingman, J. F. C. & Taylor, S. J. 1966 *Introduction to measure and probability*. Cambridge University Press.

Kruskal, J. B. 1964 Multidimensional scaling. *Psychometrika.* **29**, 1–27; 28–42.

Marr, David 1969 A theory of cerebellar cortex. *J. Physiol.* **202**, 437–470.

Marr, David 1971 Simple Memory: a theory for archicortex. (Submitted for publication.)

Nauta, W. J. H. 1954 Terminal distributions of some afferent fibre systems in the cerebral cortex. *Anat. Rec.* **118**, 333.

Petrie, W. M. Flinders, 1899 Sequences in prehistoric remains. *J. Anthrop. Inst.* **29**, 295–301.

Rēnyi, A. 1961 On measures of entropy and information. In: *4th Berkely Symposium on Mathematical Statistics and Probability* (Ed. J. Neyman), pp. 547–561. Berkeley: Univ. of California Press.

Shannon, C. E. 1949 In *The mathematical theory of communication*, C. E. Shannon & W. Weaver. Urbana: Univ. of Illinois Press.

Sholl, D. A. 1956 *The organisation of the cerebral cortex*. London: Methuen.

Sibson, R. 1969 Information radius. *Z. Wahrscheinlichkeitstheorie* **14**, 149–160.

Sibson, R. 1970 A model for taxonomy. II. (Submitted for publication.)

Spencer, W. A. & Kandel, E. R. 1961 Electrophysiology of hippocampal neurons IV. Fast prepotentials. *J. Neurophysiol.* **24**, 274–285.

Szentágothai, J. 1962 On the synaptology of the cerebral cortex. In: *Structure and functions of the nervous system* (Ed. S. A. Sarkisov). Moscow: Medgiz.

Szentágothai, J. 1965 The use of degeneration methods in the investigation of short neuronal connections. In: *Degeneration patterns in the nervous system, Progr. in Brain Research* **14** (Eds. M. Singer & J. P. Schadé), 1–32. Amsterdam: Elsevier.

Szentágothai, J. 1967 The anatomy of complex integrative units in the nervous system. *Recent development of neurobiology in Hungary* **1**, 9–45. Budapest: Akadémiai Kiadó.

Valverde, F. 1968 Structural changes in the area striata of the mouse after enucleation. *Exp. Brain Res.* **5**, 274–292.