David Marr
1970 – Cambridge, England

# From the Retina to the Neocortex

## Selected Papers of David Marr

Edited by Lucia Vaina

With Commentaries by

Jack D. Cowan
W. Eric L. Grimson
Norberto M. Grzywacz
Ellen C. Hildreth
Bruce McNaughton
Terrence J. Sejnowski
W. Thomas Thach
David Willshaw

Birkhäuser
Boston · Basel · Berlin 1991

Lucia Vaina
Intelligent Systems Laboratory
College of Engineering and Department of Neurology
Boston University
and
Harvard - Massachusetts Institute of Technology
Division of Health Sciences and Technology

**Library of Congress Cataloging-in-Publication Data**

Marr, David, 1945–1980.
From the retina to the neocortex : selected papers of David Marr /
Lucia M. Vaina. editor.
p. cm.
Consists of reprints of selected papers of David Marr, with commentaries and personal memories by specialists in his field.
Includes bibliographical references.
ISBN-13: 978-1-4684-6777-2 e-ISBN-13: 978-1-4684-6775-8
DOI: 10.1007/978-1-4684-6775-8
1. Neural networks. 2. Brain–Computer simulation. 3. Marr, David, 1945–1980. I. Vaina, Lucia, 1946– . II. Title.
[DNLM: 1. Marr, David, 1945–1980. 2. Cerebellum–physiology–collected works. 3. Computer Simulation–collected works. 4. Depth Perception–physiology–collected works. 5. Models, Neurological–collected works. 6. Nerve Net–physiology–collected works. 7. Vision. Binocular–physiology–collected works. WL 7 M3580]
QP363.3.M37 1991
006.3–dc20
DLC

Printed on acid-free paper.



Original material prepared by the editor using TeX.

ISBN-13: 978-1-4684-6777-2

9 8 7 6 5 4 3 2 1

# Contents

*To Madge and Doug with love.*

# Introduction

David Courtnay Marr was born on January 19, 1945 in Essex, England. He went to the English public school, Rugby, on scholarship and between 1963 and 1966 studied mathematics at Trinity College, Cambridge University where he obtained his B.S. and M.S. degrees. Rather than pursue a Ph.D. in mathematics he preferred to switch to neurophysiology under Giles Brindley. His education involved training in neuroanatomy, neurophysiology, biochemistry, and molecular biology. Marr's Ph.D. work resulted in a theory of the cerebellar cortex, the essence of which became "A Theory of the Cerebellar Cortex," reproduced in Chapter 1 of this volume with a commentary by Thomas Thach. He wrote a short paper subsequently with Stephen Blomfield, "How the Cerebellum May Be Used," (Chapter 2 in this volume with commentary by Jack Cowan). After obtaining his Ph.D., David Marr accepted an appointment to the scientific staff of the MRC Laboratory of Molecular Biology in Cambridge in the division of Cell Biology under Sydney Brenner and Francis Crick.

Two other major studies, "Simple Memory: A Theory of the Archicortex" (Chapter 3 in this volume, commented on by Bruce McNaughton and David Willshaw) and "A Theory for Cerebral Neocortex" (Chapter 4 in this volume and commented on by Jack Cowan) followed the cerebellum study.

"Truth, I believed, was basically neuronal, and the central aim of research was a thorough analysis of the structure of the nervous system" (Marr, 1982). This view, combined with his initial training in mathematics, shaped the methodology that Marr applied in these three studies:

> For a mathematician, understanding (or explanation) is all, yet in science, proof is, of course, what counts. In the case of Information-Processing devices, understanding is very important; one can know a fact about a device for years without really understanding it, and part of the theoretician's job is to place into a comprehensible framework the facts that one already knows. I still think that the cerebellum is a good example. For sure, the idea that the parallel fibre — Purkinje cell synapses — might be modifiable may not have been very difficult to arrive at, and other theories have since incorporated it; but that surely is only a part of the story. I found the real impact of that story to lie in the combinatorial trick. That is, this granule cell arrangement, with associated inhibitory interneurones, had been right in front of people's eyes ever since Cajal (modulo inhibition and excitation) but its significance had not been appreciated. Of course my theory might yet be wrong, but if it is right, then I would regard a major part of its contribution as being explanatory. And also, that that is almost inevitable.
>
> from a letter to Francis Crick, 1977

Marr's early work was aimed at understanding cortical structures in functional terms, and the mathematical framework allowed him to make several predictions that, especially for the cerebellum theory, inspired many experimentalists over the years. For brain theorists, Marr's models of the cerebellum, archicortex, and neocortex remain models of simplicity, mathematical rigor and explanatory power.

In 1973 David Marr came to the Artificial Intelligence Laboratory at MIT, first as a visiting scientist for a few months, but since "the facilities and the people were really impressive" he decided to stay on for "a year or two." At MIT he began working on vision. So, he writes to Giles Brindley in October 1973:

> I turned to vision when I arrived here [MIT], hoping that insight into the functions you had to perform to recognize something, together with the detailed neurophysiological knowledge and an unexcitable disposition, would be capable of illuminating many questions that are surely not vulnerable to the microelectrode.

In December of the same year, his decision to break with the previous research was stated clearly in a short letter to Brindley:

> I do not expect to write any more papers in theoretical neurophysiology — at least not for a long time: but I do not regard the achievements of yours 1969, or my papers as negligible. At the very least, they contain techniques that anyone concerned with biological computer architecture should be aware of, and I shall be very surprised if my 1969 or 1971 papers turn out to be very wrong.

Influenced by Horn's algorithm for computing lightness and by Land's retinex theory, Marr began thinking about the functions of the retina. His work in vision took a fresh approach influenced both by the enthusiasm in the then new field of artificial intelligence and in neuroscience. Cambridge, Massachusetts was already an intellectual Mecca where things were happening, where communication was fast and the work was first rate:

> I have just spent a week with Jack Pettigrew, who is a very bright and exciting person! He is studying the development of the visual cortex, and has the most extraordinary results! The features coded for really do depend on what the kittens see. He was full of the results you mentioned, and especially those of Zeki. Apparently there is a stereo area, a movement area, as well as a colour one. I am writing a short summary of the computations performed by the visual cortex.
>
> Marr wrote to a Cambridge friend, May 1973

The same year in September he wrote a long and thoughtful letter to Sydney Brenner, his intellectual mentor and friend:

> I have been thinking about the future. Presumably, as a result of the Lighthill report, AI must change its name. I suggest BI (Biological Intelligence!). I am more and more impressed by the need for a functional approach to

> the CNS and view it as significant that the crucial steps for the retina were taken by Land, the only scientist in the field actually concerned with handling real pictures (in his case on colour film). The moral is that if you wish to do vision research, you must have the facilities for taking, recording and processing real live pictures, to see if what you think gets results actually does. I see a bright future for vision in the next few years, and am anxious to stay in the subject, doing my own AI research as well as acting as interface with neurophysiology.

He began thinking of a computational approach to vision. The motivation and the essence of the new approach were clear to him already in 1973, as he replies to Dunin-Barkovski's request for permission to translate his earlier papers into Russian:

> It would be fun to have some of it translated into Russian. My present opinion of my earlier work is, however, that even if it is correct, it does not take one much further in the study of how the brain works than, for example, the study of more obviously physical phenomena like synaptic transmission, or the conduction of nervous impulses. The reason why I believe this is that this part of my work has to do more with computer architecture than with biological computer *programs*! I have studied how some basic "machine-code" instructions can be implemented in nervous tissue; but these studies tell you rather little about how the rest of the brain uses these facilities — e.g., what is the overall structure of a particular motor program for picking an object up, or for throwing a ball. *It is the second kind of question that I am now interested in.*

Neural net theories, fashionable then in theoretical biology, had severe limitations that Marr clearly expressed in a review of approaches to biological information processing:

> The neural net theory states that the brain is made of neurons, connected either specifically (for small structures) or randomly (for larger ones). Hence, in order to understand the brain we need to understand the behavior of these assemblies of neurons. Here there are two problems. First, the brain is large but it is certainly not wired up randomly. The more we learn about it, the more specific the details of its construction appear to be. Hoping that random neural net studies will elucidate the operation of the brain is therefore like waiting for the monkey to type Hamlet. Second, given a specific function of inevitable importance like a hash-coded associative memory, it is not too difficult to design a neural network that implements it with tolerable efficiency. Again, the primary unresolved issue is *what* functions you want to implement and *why*. In the absence of this knowledge, a neural net theory, unless it is closely tied to the known anatomy and physiology of some part of the brain and makes some unexpected predictions, is of no value.
>
> *Science*, 1975, vol. 190 pp. 875-876

The reactions to this harsh view were mixed as we read in this computer mail message from Kanerva:

> I admire your courage in submitting to print your considered, and critical, views on theories of biological information processing (*Science*, 1975) and on AI. You probably mentioned on the phone of their getting you into some trouble, but I just wonder who really is in trouble. As I see it, you are not inclined to build your house on sand, and that you question whether researchers by and large consider what foundations they are building on or what the structure, if finished, is supposed to accomplish. My feelings are with you, but I find justifying anything — beyond justifying it to myself — extremely difficult. And yet that is what one has to do.
>
> Pentti Kanerva, (15 March 1977)

In the meantime Marr's work on the retina was progressing very nicely:

> For the retina, I am not wholly responsible. Nick Horn, co-director of the vision mini-robot project, came up with a beautiful algorithm for computing Land's Retinex function [see *J. Opt. Soc. Am.* 61 (1971) pp. 1-11]. It is not quite the actual one actually used, but was near enough to enable one to take the last steps. I am busy tying up all the detailed anatomy and physiology now, and am very hopeful that the whole thing will turn out to be very pretty. But the retinex is the real secret. We haven't decided yet how to publish it: perhaps two separate papers. If so, mine will show how almost everything that needs to be said is in the literature somewhere, but scattered over about 200 papers. It is great fun, even if not as original as my earlier work. One of our wholly new findings is that the so called center-surround organization of the retinal ganglion cells is all a hoax! It is nothing but a by-product of showing silly little spot stimuli to a clever piece of machinery designed for looking at complete views. That will put the cat among the pigeons in a very satisfying manner!
>
> from a letter to Sydney Brenner, July 1973

Two papers were published in 1974: Horn's paper entitled "On Lightness" and Marr's entitled "The Computation of Lightness by the Primate Retina." The latter is reproduced in Chapter 5 in this volume with a commentary by Norberto Grzywacz.

The retina paper was followed by a computational theory of stereopsis outlined first in an internal AI lab memo, "A Note on the Computation of Binocular Disparity in a Symbolic Low-level Visual Processor." (Chapter 1 in Part II of this volume). This paper marked the beginning of the famous collaboration with Tommy Poggio, who was then at the Max Planck Institute in Tubingen. They first published the "Cooperative Computation of Stereo Disparity." (Chapter 2 in Part II here), and subsequently, in "A Computational Theory of Human Stereo Vision", they proposed an algorithm thought to be used by the human visual system for solving the stereo problem (Chapter 3 in Part II). These three papers are commented on by Ellen Hildreth and Eric Grimson, and the extraordinary excitement of the work is most vividly described by Poggio in the Epilogue (Part IV) of this volume.

Marr advocated and practiced a program for research into brain functions that required focusing on the study of the information processing problems inherent in the tasks themselves, rather than structural details of the mechanism

that performs them. He stressed, however, that the study of the information processing problems was not sufficient. In vision, for example, once we know how to compute a description of a scene from an image intensity array it will be possible to design neural implementations of the computation. Essentially, useful contributions had to be made at the computational level, and this requires working on real problems (as opposed to idealized blocksworld problems), and powerful and flexible computational facilities that were available at the MIT Artificial Intelligence Laboratory were making this work possible. He wrote to a friend in the Spring of 1975, "I left the cozy and comfortably decadent confines of the British Isles to confront the harsher realities of this abrasive and invigorating climate, and am now studying vision." In 1977 he joined the faculty of the MIT Psychology Department and in 1980 was promoted to a permanent position and full professor.

At MIT, David Marr spent years of incredibly intense and fruitful collaborations with Poggio, Ullman, Grimson, Hildreth, Nishihara, Richards and Stevens, among the closest. The results of these collaborations were presented in a series of papers and in his book Vision which presents "A computational investigation into the human representation and processing of visual information" (the subtitle of the book). The new and original approach of this book has made it into a classic textbook and reference for anybody working in vision, no matter what approach he takes.

In the closing chapter of this volume (Part III) Sejnowski presents an elegant consistency proof of Marr's approaches in the early and the later studies, and demonstrates that together, these constitute an important framework for those working in computational neuroscience.

The book ends with an Epilogue, which through letters from friends, students, and colleagues, vividly portrays David Marr's complex personality and his zest for living. He lived with the same intensity and commitment to life with which he carried out his research. Life was to be enjoyed, discovered, and conquered in all its beauty and complexity. And those who were close to him will always remember that, until the last day, November 17, 1980, David remained faithful to his commitment to life and work.

Lucia M. Vaina

*Cambridge, September 1990*

# Acknowledgments

I wish to thank the Editors of *Nature*, The Physiological Society, Pergamon Press, Inc., Springer-Verlag, the American Association for the Advancement of Science, and the Royal Society for permission to reprint the articles in this volume. The preparation of this volume has been supported, in part, by the National Institutes of Health and the National Eye Institute, grant no. 5RO1 EY7861-2.

This book would not have been possible without the thoughtful contributions—the commentaries and personal reminiscences—of David's many friends and colleagues. I thank them all. I would also like to express my gratitude to George Adelman and Jim Doran from Birkhäuser Boston for their patience and encouragement during the preparation of this book. Thanks are due to my student, Norman Stratton for his kind help in preparing some of the figures. Finally, my warmest thanks to JoAnn Sorrento, who has coped gracefully with the many versions, and all the tedious chores associated with producing this book.

# Early Papers

*J. Physiol.* (1969), **202**, *pp.* 437–470
*With 1 plate and 2 text-figures*



# A THEORY OF CEREBELLAR CORTEX

BY DAVID MARR*
*From Trinity College, Cambridge*

(*Received* 2 *December* 1968)

SUMMARY

1. A detailed theory of cerebellar cortex is proposed whose consequence is that the cerebellum learns to perform motor skills. Two forms of input–output relation are described, both consistent with the cortical theory. One is suitable for learning movements (actions), and the other for learning to maintain posture and balance (maintenance reflexes).

2. It is known that the cells of the inferior olive and the cerebellar Purkinje cells have a special one-to-one relationship induced by the climbing fibre input. For learning actions, it is assumed that:

(*a*) each olivary cell responds to a cerebral instruction for an elemental movement. Any action has a defining representation in terms of elemental movements, and this representation has a neural expression as a sequence of firing patterns in the inferior olive; and

(*b*) in the correct state of the nervous system, a Purkinje cell can initiate the elemental movement to which its corresponding olivary cell responds.

3. Whenever an olivary cell fires, it sends an impulse (via the climbing fibre input) to its corresponding Purkinje cell. This Purkinje cell is also exposed (via the mossy fibre input) to information about the context in which its olivary cell fired; and it is shown how, during rehearsal of an action, each Purkinje cell can learn to recognize such contexts. Later, when the action has been learnt, occurrence of the context alone is enough to fire the Purkinje cell, which then causes the next elemental movement. The action thus progresses as it did during rehearsal.

4. It is shown that an interpretation of cerebellar cortex as a structure which allows each Purkinje cell to learn a number of contexts is consistent both with the distributions of the various types of cell, and with their known excitatory or inhibitory natures. It is demonstrated that the mossy fibre-granule cell arrangement provides the required pattern discrimination capability.

5. The following predictions are made.

(*a*) The synapses from parallel fibres to Purkinje cells are facilitated by the conjunction of presynaptic and climbing fibre (or post-synaptic) activity.

* Now at the Institute of Psychiatry, London, S.E. 5

(*b*) No other cerebellar synapses are modifiable.

(*c*) Golgi cells are driven by the greater of the inputs from their upper and lower dendritic fields.

6. For learning maintenance reflexes, 2(*a*) and 2(*b*) are replaced by

2′. Each olivary cell is stimulated by one or more receptors, all of whose activities are usually reduced by the results of stimulating the corresponding Purkinje cell.

7. It is shown that if (2′) is satisfied, the circuit receptor → olivary cell → Purkinje cell → effector may be regarded as a stabilizing reflex circuit which is activated by learned mossy fibre inputs. This type of reflex has been called a learned conditional reflex, and it is shown how such reflexes can solve problems of maintaining posture and balance.

8. 5(*a*), and either (2) or (2′) are essential to the theory: 5(*b*) and 5(*c*) are not absolutely essential, and parts of the theory could survive the disproof of either.

## §0. INTRODUCTION

The cortex of the vertebrate cerebellum has a simple and extremely regular fine structure. This happy combination has made detailed experimental investigations possible, with the result that the arrangement and connexions of the cerebellar cells, together with the excitatory or inhibitory nature of the various synapses, are now clear (see Eccles, Ito & Szentagothai, 1967).

The structure of cerebellar cortex, though well understood, has as yet received no plausible interpretation. In the present paper, a theory of the cortex is proposed which explains what is known about it, and makes certain definite and testable predictions. The implication of the cortical theory is that the purpose of the cerebellum is to learn motor skills, so that when they have been learned a simple or incomplete message from the cerebrum will suffice to provoke their execution. Brindley (1964) suggested this was the function of the cerebellum.

The exposition is divided into various sections. In the first, an outline of the theory is presented: this is intended to provide a framework within which the reader may fit the details. The next five sections contain a cell by cell account of the cortex, and these are followed by a closer look at the input–output relations consistent with the theory.

## §1. OUTLINES

The axons of the Purkinje cells form the only output from the cortex of the cerebellum (see Fig. 1); and these cells are driven by two essentially different kinds of input, one direct, the other indirect. The first is the climbing fibre input, and the second the mossy fibres, whose influence on the Purkinje cells may be complicated.

The inferior olive is the only known source of climbing fibres: every cell in the inferior olivary nuclei projects to the cerebellum, and every part of the cerebellum possesses climbing fibres (Eccles *et al.* 1967). Each of the rather small olivary cells sends out an axon which terminates in one climbing fibre on just one Purkinje cell: there are very few exceptions. The climbing fibre completely dominates the dendritic tree of the Purkinje cell, and its action has been shown to be powerfully excitatory (Eccles *et al.* 1967). Thus every olivary cell has a unique representational cell in the cerebellum which can be acted upon by all the influences mediated by the parallel fibres. In the present theory, it is suggested that each olivary cell corresponds to a 'piece of output' which it is necessary to have under control during movements. This 'piece of output' could take many forms: it might be a limb movement, or a fine digit movement, or an instruction to read vestibular output in a particular way to set up an appropriate control loop. Such 'pieces of output' will be called *elemental movements*; and each olivary cell may for the moment be supposed to correspond to one elemental movement in the sense that it is driven by an instruction for that movement to take place.

It is imagined that the olivary dictionary of elemental movements is complete: that is, every possible action can be represented as an ordered pattern of elemental movements each of which has a special olivary cell. Every action therefore has a defining representation as a sequence of firing patterns in the olive.

The final assumption, which relates the olivo-cerebellar system to the execution of motor actions, is that the nervous system has a way of converting the (inhibitory) output of a Purkinje cell into an instruction which provokes the precise movement to which its uniquely related olivary cell responds.

It will be argued that the reason for the special and in a sense substitutive relationship between a cell of the inferior olive and a Purkinje cell of the cerebellum is that the Purkinje cell can learn all the 'situations' in which the olive cell movement is required, and later, when such a situation occurs again, can implement that movement itself. If this were true of enough Purkinje cells (at least one for every elemental movement), the cerebellum could learn to carry out any previously rehearsed action which the cerebrum chose to initiate, for as that action progressed, the context for the next part of it would form, would be recognized by the appropriate Purkinje cells, and these would turn on the next set of muscles, allowing further development of the action. In this way, each muscle would be turned on and off at the correct moment, and the action would be automatically performed.

Information defining the context for each Purkinje cell is provided by

the mossy fibre input: and to establish that the Purkinje cells can learn contexts in the appropriate way it is necessary to demonstrate that the mossy fibre-granule cell–Purkinje cell arrangement could operate as a pattern recognition device. The notion fundamental to this is that the mossy fibre-granule cell articulation is essentially a pattern separator. That is, it amplifies discrepancies between patterns that are rather similar, translating two overlapping collections of mossy fibres into bundles of parallel fibres that overlap proportionately much less, if at all. One Purkinje cell can be made to store different contexts quite reliably by facilitating the relevant parallel fibre–Purkinje cell synapses: and this will work as long as the Purkinje cell does not try to learn too much. Evidently, the cue for synaptic modification is that the relevant climbing fibre be also active, and it is this which leads to the modification hypothesis.

These ideas lead to the notion that a mossy fibre input has been learnt by a given Purkinje cell if, and only if, the input is transformed into impulses in a bundle of parallel fibres all of whose synapses with that Purkinje cell have been facilitated. Two crucial points now arise. First, the number of parallel fibres into which a mossy fibre input is translated increases very sharply with the number of active mossy fibres unless the threshold of the granule cells also increases. The number of patterns each Purkinje cell can learn depends on the number of synapses which are facilitated in each: so economy arguments suggest that the granule cell threshold should be controlled in a suitable way. An inhibitory interneurone could achieve this, and the Golgi cells are interpreted as fulfilling this role.

The second point is that although the effect of the Golgi cells is to decrease the variation in the amount of parallel fibre activity, such variation will still exist. Whether or not a Purkinje cell should respond to a given mossy fibre input cannot therefore be decided by a fixed threshold mechanism. The Purkinje cell threshold must vary directly with the number of active parallel fibres running through its dendritic tree, and its actual value must be such that the cell emits a signal when and only when all the active parallel fibres have facilitated synapses with its dendrite. The natural way to implement this is to allow the parallel fibres to drive an interneurone which inhibits the Purkinje cell: and it will be shown that the various stellate inhibitory cells can be associated with this function, although their dendritic and axonal distributions are at first sight unsuitable.

### 1.1. *Data*

The anatomical and physiological information used in this paper concerns the cerebellum of cat, and is mostly derived from Eccles *et al.* (1967). Facts which are well known will not usually be given a reference:

information which is less well known is given a page reference in Eccles *et al.* (1967) if it appears there; otherwise an external reference is given.

A diagram of the general cerebellar cortical structure appears in Fig. 1. The cortex has two types of afferent fibre, the climbing fibres (*Cl*) and the mossy fibres (*Mo*). Each climbing fibre makes extensive synaptic contact with the dendritic tree of a single Purkinje cell (*p*), and its effect there is powerfully excitatory. The axons of the Purkinje cells leave the cortex (they form the only cortical output) and synapse with cells of the cerebellar nuclei.

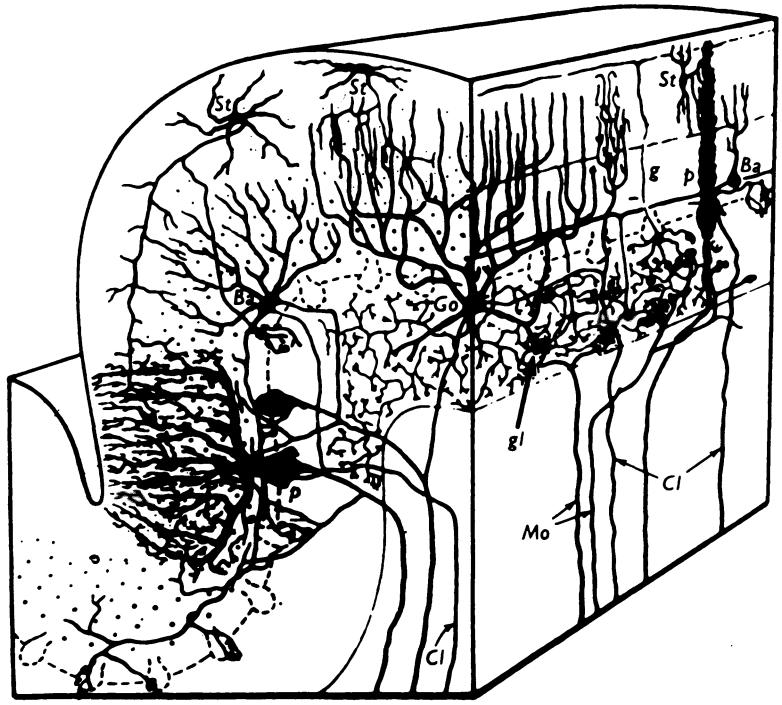



Fig. 1. Diagram of cerebellar cortex (from Eccles *et al.* 1967, Fig. 1). The afferents are the climbing fibres (*Cl*) and the mossy fibres (*Mo*). Each climbing fibre synapses with one Purkinje cell (*p*), and sends weak collaterals to other cells of the cortex. The mossy fibres synapse in the cerebellar glomeruli (*gl*) with the granule cells, whose axons (*g*) form the parallel fibres. The parallel fibres are excitatory and run longitudinally down the folium: they synapse with the Purkinje cells and with the various inhibitory interneurones, stellate (*St*), basket (*Ba*) and Golgi cells (*Go*). The stellate and basket cell axons synapse with the Purkinje cells, and the Golgi cell axons synapse in the glomeruli with the granule cells. As well as their ascending dendrites, the Golgi cells possess a system of descending dendrites, with which the mossy fibres synapse in the glomeruli. The Purkinje cell axons form the only output from the cortex, and give off many fine collaterals to the various inhibitory interneurones.

The second input, the mossy fibres, synapse in the cerebellar glomeruli (*gl*) with the granule cells. Each glomerulus contains one mossy fibre terminal (called a rosette), and dendrites (called claws) from many granule cells. The glomerulus thus achieves a considerable divergence, and each mossy fibre has many rosettes.

The axons of the granule cells rise (*g*) and become the parallel fibres, which synapse in particular with the Purkinje cells whose dendritic trees they cross. Where the granule cell axons (i.e. the parallel fibres) make synapses, they are excitatory.

The remaining cells of the cortex are inhibitory interneurones. The Golgi cells (*Go*) are large, and have two dendritic trees. The upper tree extends through the molecular layer, and is driven by the parallel fibres. The lower dendrites terminate in the glomeruli, and so are driven by the mossy fibres. The Golgi axon descends and ramifies profusely: it terminates in the glomeruli, thereby inhibiting the granule cells. Every glomerulus receives a Golgi axon, almost always from just one Golgi cell: and each Golgi cell sends an axon to all the glomeruli in its region of the cortex.

The other inhibitory neurones are stellate cells, the basket (*Ba*) and outer stellate (*St*) cells. These have dendrites in the molecular layer, and are driven by the parallel fibres. Both types of cell synapse exclusively with Purkinje cells, and are powerfully inhibitory.

Finally, the cortex contains various axon collaterals. The climbing fibres give off weak excitatory collaterals which make synapses with the inhibitory interneurones situated near the parent climbing fibre. The Purkinje cell axons give off collaterals which make weak inhibitory synapses with the cortical inhibitory interneurones, and perhaps also very weak inhibitory synapses with other Purkinje cells. These collaterals have a rather widespread ramification.

Behind this general structure lie some relatively fixed numerical relations. These all appear in Eccles *et al.* (1967), but are dispersed therein. It is therefore convenient to set them down here.

Each Purkinje cell has about 200,000 (spine) synapses with the parallel fibres crossing its dendritic tree, and almost every such parallel fibre makes a synaptic contact. The length of each parallel fibre is 2–3 mm ($1\frac{1}{2}$ mm each way), and in 1 mm down a folium, a parallel fibre passes about 150 Purkinje cells. Eccles *et al.* (1967) are certain each fibre makes at least 300 (of the possible 450) synaptic contacts with Purkinje cells, and think the true number is nearer 450. There is one Golgi cell per 9 or 10 Purkinje cells, and its axon synapses (in glomeruli) with all the granule cells in that region, i.e. around 4500. There are many granule cells ($2{\cdot}4 \times 10^6$ per mm$^3$ of granule cell layer), each with (usually) 3–5 dendrites (called claws): the average is 4·5 and the range 1–7. Each dendrite goes to one and only one

glomerulus, where it meets one mossy fibre rosette. It is, however, not alone: each glomerulus sees the termination of about 20 granule cell dendrites, possibly a Golgi cell descending dendrite, and certainly some Golgi axon terminals, all from the same Golgi cell. Within each folium, each mossy fibre forms 20–30 rosettes, giving a divergence of 1 mossy fibre to 400–600 granule cells within a folium. The mossy fibre often has branches running to other folia, and in Fig. 2 below one can count 44 rosettes on one fibre.

Just below the Purkinje cells are the Golgi cell bodies, and just above them are the basket cell bodies. There are 10–12 % more basket cells than Purkinje cells, and about the same number of outer stellate cells. Each basket cell axon runs for about 1 mm transversely, which is about the distance of 10 Purkinje cells. The basket axon is liable to form baskets round cells up to three away from its principal axis, so its influence is confined to a sort of box of Purkinje cells about 10 long and 7 across. The distribution of the outer stellate axons is similar except that it has a box about $9 \times 7$, since its axon only travels about 0·9 mm transfolially. The outer stellates inhabit the outer half of the molecular layer, and the basket cells the inner third. There are intermediate forms in the missing sixth. None of these cells has a dendritic tree as magnificent as that of the Purkinje cell, and Eccles *et al.* (1967) do not venture any comparative figures. Some outer stellates are small, with a local axonal distribution. A lot of the synapses of parallel fibres with this last group of cells are directly axo-dendritic, but all other parallel fibre synapses are via spines, though these are of different shapes on the different sorts of cell. Calculations based on slightly tenuous assumptions (in which Fig. 2 is an essential link) suggest that each Purkinje cell receives connexions from about 7000 mossy fibres: this will be explained in 3.1.

## §2. CLIMBING FIBRES

The climbing fibre input has already been discussed at some length, and a formal statement of its part in the modification hypothesis will be made in 5.1. It is important to note that the fibre climbs like a creeper all over the dendritic tree of its chosen Purkinje cell, and forms synaptic contact almost everywhere. Each climbing fibre also sends terminals to other types of cell (basket, stellate and Golgi) in the vicinity of its Purkinje cell. These terminals seem to be excitatory, but only weakly so (Eccles *et al.* 1967, Table 1, p. 63). The climbing fibre collaterals and the Purkinje axon collaterals will be discussed together in 5.5.

## §3. MOSSY FIBRES AND GRANULE CELLS

### 3.0. *The codon representation*

The synaptic arrangement of the mossy fibres and the granule cells may be regarded as a device to represent activity in a collection of mossy fibres by elements each of which corresponds to a small subset of active mossy fibres. It is convenient to introduce the following terms: a *codon* is a subset of a collection of active mossy fibres. The representation of a mossy fibre input by a sample of such subsets is called the *codon representation* of that input: and a *codon cell* is a cell which is fired by a codon. The granule cells will be identified as codon cells, so these two terms will to some extent be interchangeable. The size of codon that can fire a given granule cell depends upon the threshold of that cell, and may vary: and the mossy fibres which synapse with the granule cell determine the codons which may fire that cell.

There are exactly

$$\binom{L}{R} = \frac{L!}{R!(L-R)!}$$

codons of size $R$ associated with a collection of $L$ active mossy fibres. If two mossy fibre inputs each involve activity in $L$ fibres of which $M$ were common to the two, the two inputs are said to *overlap* by $W$ elements; and they may be expected to have some codons in common. In fact the number they share is precisely $\binom{W}{R}$. The ratio $X$ of the number of shared codons to the number of codons each possesses is given by

$$X = \binom{W}{R} \Big/ \binom{L}{R} = \frac{W(W-1)\ldots(W-R+1)}{L(L-1)\ldots(L-R+1)} \tag{1}$$

which tends to $(W/L)^R$ as $W$ increases. The limiting values of $X$ for relevant values of $R$ appear in Table 1. It will be observed that the effect of the subset coding is to separate patterns, because similar inputs have markedly less similar codons.

TABLE 1. Overlap Table, i.e. values of $(W/L)^R$

| $(W/L)$ | $R = 2$ | 3 | 4 | 5 |
|---|---|---|---|---|
| 0·5 | 0·25 | 0·12 | 0·06 | 0·03 |
| 0·6 | 0·36 | 0·22 | 0·13 | 0·08 |
| 0·7 | 0·49 | 0·34 | 0·24 | 0·17 |
| 0·8 | 0·64 | 0·51 | 0·41 | 0·33 |
| 0·9 | 0·81 | 0·73 | 0·66 | 0·59 |

The mossy fibre-granule cell relay effectively takes a sample of the codon distribution of an input: the sample is small enough to be manageable, but large enough for the input event to be recoverable from it with high probability.

### 3.1. *The mossy fibre–Purkinje cell convergence*

A knowledge of the number of mossy fibres which may influence a given Purkinje cell is a prerequisite of a discussion of the codon sampling statistics: and this number may be estimated as follows. Let **P** be an arbitrary but henceforth fixed Purkinje cell: and assume that 200,000 parallel fibres synapse with **P**. Each granule cell has (on average) 4·5 claws,

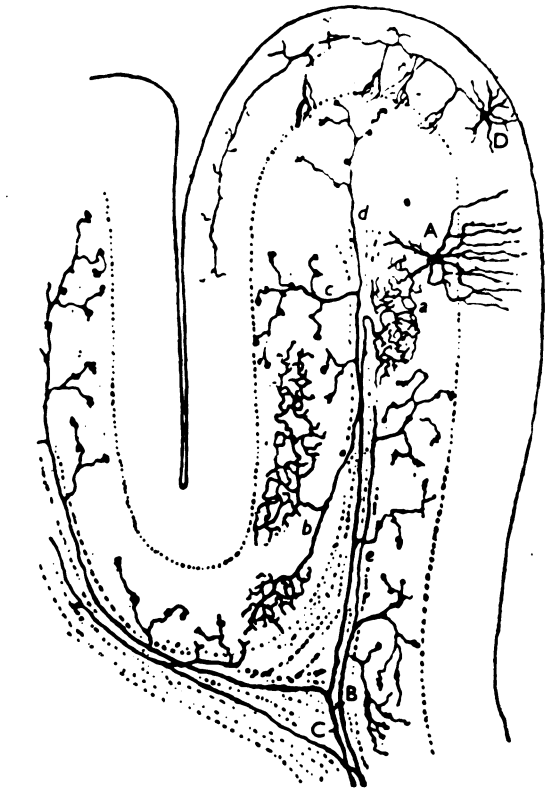



Fig. 2. Mossy fibres (B and C) terminating in two neighbouring folia (from Cajal, 1911, Fig. 41). The distribution of the terminals from each mossy fibre lies in the same plane as the axon of the basket cell (D). A is a Golgi cell.

so not more than 900,000 mossy fibres can influence **P** through the parallel fibres. Since the mossy fibre-granule cell divergence is 400–600 within a folium, the minimum figure for the mossy fibre–**P** convergence is 1500. It is apparent from Fig. 2 that the mossy fibre terminals occur in clumps of 4–10 rosettes, the average being 7 or 8, all of which might be expected to lead to granule cells most of which will contact any nearby Purkinje cell. If a mossy fibre leads to **P**, it may therefore expect to do so by 140–160 different paths (allowing for the divergence factor of 20 due to each

glomerulus). An average of 150 implies that about 6000 different mossy fibres lead to **P**. The edge effects will increase this figure, so 7000 would probably be a reasonable guess. This estimate will be used in the subsequent calculations.

3.2. *The assumption of randomness*

In the investigation that follows, it is assumed that the terminals of the 7000 mossy fibres are distributed randomly among the 200,000 granule cells leading to **P**. It is regrettable that no data exist to suggest a better model: and evidence will be produced (3.3.3 and 4.3) for the view that this assumption is actually false. Its value is that it enables computation whose results are at least illustrative: and one has the comfort of knowing that the capacity of a real cerebellum will anyway not be less than the result of calculations which assume a random distribution.

3.3. *The granule cell claws*

3.3.1. *Boundary conditions.* It will be assumed first that the claw arrangement of the granule cell dendrites is, as suggested by Eccles *et al.* (1967), a device to secure a high mossy fibre-granule cell divergence with minimal physical structure. But why do the granule cells have 4 or 5 claws and not more? These cells are extremely small and densely packed: and the parallel fibre synapses on **P** are extremely numerous. It is therefore reasonable to assume that the figure of 200,000 (or thereabouts) is the maximum physically realizable number of cells of this sort which can all send axons to **P**.

Secondly, it will be assumed that the synapses at the granule cells are not modifiable: that is, an excited mossy fibre will add a contribution to the excitatory post-synaptic potential (EPSP) of any granule cell with which it synapses, and this contribution has to be considered in determining whether or not that cell fires. This is justified below (3.3.3). Thirdly, the number of mossy fibres leading to **P** is of the order of 7000 in number: and fourthly, it is assumed that the system is to be used under conditions in which the number of active input fibres varies from around 20 to around 2000, if that is possible. There is clearly a need to allow considerable variation; some actions involve many more muscles and much more information from receptor organs than others. These figures are proposed as outer bounds, in the absence of any relevant evidence. Fortunately, it turns out not to matter crucially: the essential point is that the numbers are all nearer 0 than 7000 (on an arithmetical scale).

3.3.2. *Codon sampling.* The following rough model is used to calculate the number of granule cells per Purkinje cell that a given input can expect to stimulate. Suppose the number of active mossy fibres among the 7000

connected to the Purkinje cell **P** is $L$: then the number of possible codons is $\binom{L}{R}$, where $R$ is the size of the codon. The number of codons which could be generated by the 7000 possible mossy fibres is $\binom{7000}{R}$; and if we assume that the granule cells have $R$ claws and a threshold of $R$, then they represent a collection of 200,000 codons, supposed chosen randomly from the possible $\binom{7000}{R}$. Hence the number of granule cells per Purkinje cell stimulated by a given input of $L$ active mossy fibres follows approximately binomial statistics with expectation

$$200{,}000 \binom{L}{R} \Big/ \binom{7000}{R}. \tag{2}$$

The calculations that follow are concerned only with expectations: the numbers have in fact to be large enough for the distribution to be rather tightly clumped round the expectation. This is discussed in 5.2.3.

Suppose now that the granule cell has $C$ claws and threshold $R \leqslant C$. That granule cell now has a catchment area of exactly $\binom{C}{R}$ codons of size $R$: and expression (2) becomes

$$200{,}000 \binom{C}{R} \binom{L}{R} \Big/ \binom{7000}{R} \tag{3}$$

which is valid for expectations small compared with 200,000. (3) becomes (2) when $C = R$. The approximation (3) may be used, since it will be shown in 4.4 that situations will probably never occur in which the expectation is greater than 10,000. The values of (3) have been calculated for a selection of values of $L$, $C$, and $R$, and some of the results appear in Tables 2–4. Numbers greater than 20,000 have been replaced by an asterisk.

3.3.3. *Conclusions.* The conditions of 3.3.1 may be used to discover limits on the expected values of $C$ and $R$. First, it is apparent from Table 4 that no codon size above 9 can ever be used when there are fewer than 13 claws per granule cell because too few granule cells would be activated. It is also evident and unsurprising that the maximum codon size used depends critically on the number of claws to each cell. Given this, the factor that will determine the number of claws to each cell will be economy of structure; and the relevant question is what is the least number of claws such that:

(i) The system is not swamped by large inputs: i.e. what is the least number of claws which still allows a small granule cell response to large inputs. Table 4 shows that a small response (less than 500) can be assured by 6 claws; so we expect to find at most 6.

(ii) The system remains sensitive to small inputs, if necessary by using

TABLE 2. Values of $200{,}000\binom{C}{R}\binom{L}{R}/\binom{7000}{R}$: i.e. the number of the 200,000 granule cells synapsing with one Purkinje cell that a mossy fibre input involving $L$ active fibres (out of 7000) can expect to stimulate. The granule cells have $C$ claws, and threshold $R$

$L = 20$

| $R$ | $C=2$ | 4 | 6 | 8 | 10 | 12 |
|---|---|---|---|---|---|---|
| 1 | 1134 | 2286 | 3429 | 4571 | 5714 | 6857 |
| 2 | 2 | 9 | 23 | 43 | 70 | 102 |
| 3 | — | 0 | 0 | 0 | 0 | 1 |
| 4 | — | 0 | 0 | 0 | 0 | 0 |
| 5 | — | — | 0 | 0 | 0 | 0 |
| 6 | — | — | 0 | 0 | 0 | 0 |
| 7 | — | — | — | 0 | 0 | 0 |
| 8 | — | — | — | 0 | 0 | 0 |
| 9 | — | — | — | — | 0 | 0 |
| 10 | — | — | — | — | 0 | 0 |
| 11 | — | — | — | — | — | 0 |
| 12 | — | — | — | — | — | 0 |

TABLE 3. Values of $200{,}000\binom{C}{R}\binom{L}{R}/\binom{7000}{R}$: see legend to Table 2

$L = 100$

| $R$ | $C = 2$ | 4 | 6 | 8 | 10 | 12 |
|---|---|---|---|---|---|---|
| 1 | 5,714 | 11,429 | 17,143 | * | * | * |
| 2 | 40 | 242 | 606 | 1,132 | 1,819 | 2,667 |
| 3 | — | 2 | 11 | 32 | 68 | 125 |
| 4 | — | 0 | 0 | 1 | 2 | 4 |
| 5 | — | — | 0 | 0 | 0 | 0 |
| 6 | — | — | 0 | 0 | 0 | 0 |
| 7 | — | — | — | 0 | 0 | 0 |
| 8 | — | — | — | 0 | 0 | 0 |
| 9 | — | — | — | — | 0 | 0 |
| 10 | — | — | — | — | 0 | 0 |
| 11 | — | — | — | — | — | 0 |
| 12 | — | — | — | — | — | 0 |

TABLE 4. Values of $200{,}000\binom{C}{R}\binom{L}{R}/\binom{7000}{R}$: see legend to Table 2

$L = 2300$

| $R$ | $C = 2$ | 4 | 6 | 8 | 10 | 12 |
|---|---|---|---|---|---|---|
| 1 | * | * | * | * | * | * |
| 2 | * | * | * | * | * | * |
| 3 | — | * | * | * | * | * |
| 4 | — | 2,327 | * | * | * | * |
| 5 | — | — | 4,582 | * | * | * |
| 6 | — | — | 251 | 7,016 | * | * |
| 7 | — | — | — | 657 | 9,862 | * |
| 8 | — | — | — | 27 | 1,213 | 13,339 |
| 9 | — | — | — | — | 88 | 1,943 |
| 10 | — | — | — | — | 3 | 191 |
| 11 | — | — | — | — | — | 11 |
| 12 | — | — | — | — | — | 0 |

an $R$ less than $C$. The Table for information about this is Table 2, where for $L = 20$, we have to use $R = 1$ for all tabulated values of $C$, and it is not until $L = 100$ (see Table 3) that one can use $R = 2$ with $C = 6$. It

would therefore appear that to store inputs concerning fewer than 100 active mossy fibres, systems with $C = 6$ or less have to use codon size 1. This means a loss of discrimination between overlapping inputs of fewer than 100 active fibres. Provided however there are not many such small inputs, this will not be too serious. The number that must be kept negligible is the probability that a small, unlearnt mossy fibre input will occur all of whose active fibres have previously been involved in small learnt inputs.

This difficulty can to some extent be avoided if the mossy fibres which are active together in small input events have some tendency to grow near each other. The expected granule cell responses at codon sizes $R > 1$ will then be substantial at localized spots. This can be used, because it turns out that it is best to set the codon size on a local basis, rather than setting it uniformly over all the granule cells synapsing with a given Purkinje cell. The result for the animal will be greater reliability in its cerebellar responses, so mossy fibres which are correlated in this way could be drawn together by selection.

These arguments suggest that the arrangement of 4–5 claws per granule cell is consistent with structural economy and the conditions of 3.3.1. One point remains to be discussed: it is the assumption of 3.3.1 that the mossy fibre–granule cell synapses are unmodifiable. The most straightforward argument is this: every granule cell has a synapse with at least 300 Purkinje cells, each of which probably learns about 200 mossy fibre inputs (5.3). The chance that a given mossy fibre–granule cell synapse is used in none of these is extremely small (a generous estimate is $10^{-20}$); whether or not it was initially facilitated, it almost certainly will be at some time. There is therefore no advantage in its being modifiable originally.

## §4. THE GOLGI CELLS

### 4.0. *The need for variable codon size*

It became apparent in 3.3 that if the number of active parallel fibres was to remain reasonably small over quite large variation in the number of active mossy fibres, the thresholds of the granule cells had to vary appropriately. It will be shown (5.3) that the number of patterns a Purkinje cell can learn decreases sharply as the number of active parallel fibres involved in each increases. It is therefore essential to the efficient functioning of the system that the codon size should depend on the amount of mossy fibre activity.

### 4.1. *Requirements of a codon size regulator*

In the simple model containing one output cell **P**, 200,000 associated granule cells each making (possibly ineffective) synapses with **P**, and 7000

mossy fibres making contact in a random way with the granule cells, the task of a codon size regulator is in principle simple. It must count the number of these 7000 mossy fibres which are active, and set the threshold of the granule cells so that the following conditions are satisfied.

4.1.1. The number of active granule cells must be large enough to allow adequate representation of the mossy fibre input: that is, every active mossy fibre must with high probability be included in at least one codon.

This condition may be relaxed a little, since one factor on which the discriminatory power of the cerebellum depends is the accuracy with which the decision threshold at the Purkinje cell is set (5.2.3). There is no advantage in guaranteeing representation of the whole mossy fibre input if events slightly different from a learned event will anyway be responded to because of errors introduced later.

4.1.2. The number of active parallel fibres must exceed some lower bound $N$, where $N$ will be taken as 500. This arises because the Purkinje cell threshold is not set directly from the parallel fibres with which it synapses, but from the results of sampling a number of different but closely related parallel fibres. The sampling is more reliable the more parallel fibres are active. This is explained in 5.2.3, where the figure of 500 is derived.

4.1.3. The codon size set for a particular mossy fibre input must depend only on that input; so that the same input is always translated into the same parallel fibres.

4.1.4. The codon size must be maximal, subject to conditions 4.1.1 to 4.1.3. This ensures that the number of modifiable synapses used for each learned event is minimal, and hence that the capacity is maximal (5.4).

It will be assumed that a signal in a mossy fibre is represented by a burst of impulses lasting many tens of milliseconds; and that a signal from a Purkinje cell is represented by a prolonged increase in its firing rate. This is discussed later (5.0); for the moment, it is needed only to justify the fifth condition.

4.1.5. The codon size regulating cell need not have set the granule cell threshold before the very first impulse in a signal arrives, but it must act very fast in response to such an impulse. It is essential that very little activity should be allowed into the parallel fibres while the granule cells are set at an inappropriately low threshold.

A mechanism to vary the threshold subject to these conditions could work in one of two ways: the threshold of the granule cells could be intrinsically high, and the mechanism provide excitation decreasing with increasing size of input; or the threshold could be low, and the mechanism provide inhibition increasing with increasing size of input.

4.2. *Properties of the Golgi cells*

The Golgi cells are inhibitory, can be driven by mossy fibres (through their descending dendrites) and synapse exclusively with granule cells. Further, they are particularly notable for the speed of their response (Eccles *et al.* 1967, p. 141). If the Golgi cells can be interpreted as a codon size setting device, it will therefore be as a mechanism of the second type described above.

There are, however, certain difficulties inherent in such an interpretation: first, each Golgi cell is driven by only a small number of the mossy fibre afferents to a single Purkinje cell, and sends an axon terminal to a relatively small number of granule cells; and second, the Golgi cells possess a large ascending dendrite system (Fig. 1), which on the present naive model is unexpected. The idea which the model lacks and which accounts for these various anomalies is the notion that Purkinje cells may share granule cells. Such sharing could clearly lead to great economies where two Purkinje cells needed codons from similar underlying subset distributions; but it is not obvious that sharing can be made to work, since two Purkinje cells may simultaneously require two different codon sizes.

4.3. *The effects of sharing granule cells among Purkinje cells*

If Purkinje cells are to be allowed to share granule cells, the assumption that the granule cell threshold should be constant over all cells synapsing with a given Purkinje cell must be abandoned. The most important single condition on the mossy fibre-granule cell transformation is (4.1.3) that it should be one-valued: a given mossy fibre input to a Purkinje cell should be carried there by parallel fibre activity which is determined by that input alone, and is independent of the simultaneous inputs to nearby Purkinje cells. This condition determines (in principle) the number and distribution of granule cells whose thresholds can be controlled together: for consider two adjacent Purkinje cells, $\mathbf{P}_1$ and $\mathbf{P}_2$. The collection of granule cells which synapse with $\mathbf{P}_1$ but not with $\mathbf{P}_2$ must be free to act as an independent unit, since it must be able to assume a threshold value different from the $\mathbf{P}_2$ cells. If each parallel fibre is 3 mm long, and synapses with each of the 450 Purkinje cells that grow in 3 mm along a folium, the number of granule cells that synapse with a given Purkinje cell but not with its neighbour is about $200{,}000/450 = 444$.

The conclusion that may be drawn from these arguments is that the codon size should be set independently over blocks of about 450 granule cells. If this were done by an inhibitory cell, it should possess an axon distribution like that of the existing Golgi cells but limited to 450 granule cells, and a dendrite system like the descending Golgi dendrites: further there should be one such cell per Purkinje cell.

The fact that there are fewer and bigger Golgi cells than these arguments suggest must depend on certain information not incorporated in the model. This information concerns the distribution of the mossy fibre terminals which, if it were random and 4.1.3 were satisfied, would necessitate an arrangement near the expected one. In fact, one can see from Fig. 2 that the mossy fibres have a strong tendency to course transfolially, and in any case, given one mossy fibre rosette, there is a high probability that there will be another from the same fibre nearby. The effect of this, even apart from the considerations of 3.3.3, is to make nearby granule cells more related than they would be on the random hypothesis; and it is this which allows the larger axon and basal dendrite distribution which the Golgi cells are found to possess.

### 4.4. *The ascending Golgi dendrites*

The parallel fibre activity evoked by a mossy fibre input should be unique but perhaps more important even than that, it should involve rather few fibres, since the storage capacity of a Purkinje cell depends crucially on the number of parallel fibres active in each learned event (5.3). Some idea of the numbers of parallel fibres needed for various amounts of mossy fibre activity may be gained by using the simple random model. In Table 5, the expected number of active granule cells has been computed

TABLE 5. Possible codon size transitions (underlined); $L$ is the number of active mossy fibres; $R$ is the codon size

| $L$ | $R = 1$ | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 100 | 12,857 | 323 | 4 | 0 | 0 |
| 300 | * | 2,929 | 109 | 2 | 0 |
| 500 | * | 8,148 | 507 | 15 | 0 |
| 700 | * | 15,979 | 1,395 | 60 | 1 |
| 900 | * | * | 2,967 | 163 | 3 |
| 1,100 | * | * | 5,420 | 364 | 10 |
| 1,300 | * | * | 8,950 | 711 | 22 |
| 1,500 | * | * | 13,754 | 1,261 | 45 |

for inputs with $L$ active mossy fibres, these $L$ chosen at random from a population of size 7000. The calculation has been performed on the assumption that 100,000 granules have 4 claws, and 100,000 have 5: for a threshold of $R$, the approximation used was

$$\text{expected number} = 100{,}000 \left( \binom{5}{R} + \binom{4}{R} \right) \binom{L}{R} \Big/ \binom{7000}{R} \qquad (4)$$

which is derived the same way as expression (3), and is valid only for answers small compared with 100,000. The codon size transition regions have been underlined. It will be observed that on this rather crude model, each input arouses between 500 and 9000 granule cells: (500 is the lower

bound justified in 5.2.3). If as many as 9000 are used, the capacity of each Purkinje cell will be drastically reduced. For large values of $L$, around 1000 (if such are ever used) the figure of 9000 is not unreasonable. Indeed some number of that order will be necessary to cover all the active incoming fibres. But if not, and it is questionable whether such large inputs ever occur, then to use so many would be wasteful. Provided that the total number of active parallel fibres is greater than 500, it is possible to use condition 4.1.1 to submit this number to an upper bound which depends on the amount of mossy fibre activity. For example, the number of parallel fibres active should only exceed 5000 if the number of active mossy fibres exceeds 500.

The upper dendritic tree of a Golgi cell may be interpreted as a mechanism to superimpose this upper bound; and it may be expected to work as follows. A mossy fibre signal arrives which may be quite different from what was going on before. The descending Golgi dendrites sample it and quickly set new thresholds at the relevant granule cells: this setting amounts to a first guess based on local sampling. Rather a long time later, the signals appear in the parallel fibres, and the Golgi cell, by examining the activity in a large number of these, can tell whether or not its initial assessment was the most economical solution. If it was, its behaviour should not alter: if not it should; but this will always entail shifting to a higher codon size. One cannot say that the local or global sample will always give the best solution; for example, it might happen that the mossy fibre input is sufficiently localized that it can support a high codon size for just one or two Golgi cells.

In general, therefore, a Golgi cell should be driven by that dendritic system from which it receives most excitation. This suggests that the upper and lower dendritic fields should have rather a peculiar relationship. The synaptic influences among the upper dendrites should summate, and so should the effects among the lower ones: but the summed contributions should interact so that the output from the cell is driven by the maximum of the two, not the sum. There is no firm evidence to support this prediction, but Eccles *et al.* (1967) mention that the two dendritic fields are probably too far apart to allow summation.

A proper investigation of the action of the Golgi cells would be difficult for two reasons. First, one cannot use a random model for the way the granule cells are distributed over the possible subsets of the mossy fibres, for as well as the objections of 4.3, it is likely that mossy fibres whose activities are correlated will grow near one another. This is because input events would then tend to need fewer granule cells to cover them, and could therefore be more economically stored. Secondly, an analytic model of the relationships between neighbouring Golgi cells under various input

conditions needs unrealistic simplification before it can be handled. The correct approach is probably to use a simulation programme; and the kind of result to which it will lead is that an action $A$, learnt in isolation, may have to be relearnt to some extent if it is to be performed immediately after some closely related action $B$. This will arise because the hangover in the parallel fibres of action $B$ could cause temporarily different codon sizes in certain Golgi cell blocks. The parallel fibres for $A$ in this situation are slightly different from those for $A$ performed in isolation. In other words, the price of economy is probably a not too serious loss of uniqueness for the mossy fibre–parallel fibre transformation.

### 4.5. *The Golgi cell afferent synapses*

It will be clear that within the present theory, no advantage would be gained by having the mossy fibre–Golgi cell synapses modifiable: but it is not so clear whether this is also true of the parallel fibre–Golgi cell synapses. Although there is no very simple way in which it would be useful to have these synapses modifiable, it is conceivable that there might be fringe benefits. Suppose, for example, that activity in a particular set of mossy fibres always preceded a large volley: then such advance information could be used by the Golgi cell if the conditions under which modification took place were arranged suitably.

On the other hand, modifying a synapse on a Golgi cell implies that the parallel fibre has a special relationship with the granule cells below that Golgi cell. Leaving aside the case that it came from one of them (not a special relationship of the relevant kind) there is no reason why, even if such a relationship were to hold over a number of inputs, it should hold over a majority, since one Golgi cell can expect to serve a huge number of different facilitated responses. And, in contrast to the Purkinje cells, there are no inhibitory cells of any power acting upon the Golgi cell, so there is no mechanism for deciding whether or not a majority of the currently active fibres have or have had such a special relationship. (The absolute size of a 'majority' is variable: so the Golgi cells would need a variable threshold to make such a decision, for the same reasons as do the Purkinje cells.) This argument suggests rather strongly that these synapses are not modifiable.

The other afferent Golgi synapses come from the Purkinje cell collaterals and the climbing fibre collaterals: these will be discussed in 5.5.

## §5. THE PURKINJE CELLS

### 5.0. *The Purkinje cell output*

The main branch of a Purkinje cell axon goes to one of the cerebellar nuclei and forms the only output from the cortex: its effect in the nucleus is inhibitory. It is an assumption of the present theory that the central nervous system has a means of converting a signal in a Purkinje cell axon into an order to provoke the elemental movement to which its corresponding olivary cell responds. The inhibitory nature of the Purkinje cell output suggests that it may require a positive effort to read from the cerebellum, since excitation must be fed in somewhere to the effector circuit. This would be useful (though not essential) if it were required that cerebellar output should often be ignored: and indeed it is likely that such occasions will frequently occur during waking life and possibly also during sleep. The fact that the cortical output is inhibitory can therefore be interpreted as a convenience for easy ignoring, though this is neither the only nor a necessary view.

The second point arises from the fact that Purkinje cells have a high resting discharge of 20–50 impulses/sec. (Eccles *et al.* 1967, p. 306). It was assumed above (4.1.5) that a signal in a mossy fibre was represented by a burst of impulses: and the codon size setting function of the Golgi cells depended upon this. It is a necessary consequence that efferent cortical signals should also be represented by a train of impulses rather than a single one, since the delays involved in turning on the inhibitory interneurones could make the initial response of a Purkinje cell to a mossy fibre input inappropriate. This may occur frequently, and would conveniently be hidden by a high resting discharge. Purkinje cells can sustain high rates of firing (greater than 400/sec, according to Eccles *et al.* 1967, p. 308): it is therefore reasonable to assume that a signal in a Purkinje cell axon is represented by a large increase in the firing rate, and that the effector systems are only sensitive to such messages. This assumption would have to be made for almost any theory of the cortex, since the Purkinje cells form the only output.

The input–output relations for the cortex as a whole receive attention in §7, and the Purkinje axon collaterals in 5.5.

### 5.1. *The hypothesis of modifiable synapses*

The fundamental hypothesis for the mechanism of the change of effectiveness of a parallel fibre–Purkinje cell synapse is that *if a parallel fibre is active at about the same time as the climbing fibre to a Purkinje cell with which that parallel fibre makes synaptic contact, then the efficacy of that synapse is*

*increased towards some fixed maximum value.* ('At about the same time' is an intentionally inexact phrase: the period of sensitivity needs to be something like 50–100 msec.)

If this hypothesis is true, it may have implications about the physiological conditions for synaptic modification. The most striking fact about the climbing fibres is that they extend over the whole Purkinje dendritic tree. Two of the possible reasons for this seem plausible: first, that the climbing fibre releases some sort of 'change' factor which modifies the active synapses; or second, that the fundamental condition for modification is simultaneous pre- and post-synaptic depolarization. Hebb (1949) suggested that the nervous system might contain synapses with modification conditions of the second sort.

The other and rather dangerous place one might look for implications of the modification hypothesis is in the comparison of electron-micrographs of cells supposed to have modifiable synapses with those supposed not to. This will not be attempted, but it may be relevant that the Purkinje cells seem to be the only ones in the cerebellum whose dendrites carry the characteristic tubular system which terminates 'abruptly' at the base of the spines (Eccles *et al.* 1967, p. 9).

### 5.2. *Simplifying assumptions*

The calculation of the learning capacity of a single Purkinje cell requires that certain simplifying assumptions be made.

5.2.1. It will be assumed that a synapse is either totally modified or totally unmodified: and that stimulation of a totally unmodified synapse has no effect on the post-synaptic membrane.

This is equivalent to allowing modification to increase synaptic efficacy from some fixed minimum to some fixed maximum value in one step: since the two situations can be identified by subtracting any 'ground' excitation of an unmodified synapse. Such a subtraction has a linear dependence on the number of parallel fibres active at any moment, and could easily be performed by an unmodifiable inhibitory interneurone such as the basket or outer stellate cells. This may indeed be one of the functions of these cells: it is a matter of no importance to the present theory, since such an effect would be constant throughout the life of the cerebellum. The phrase 'in one step' is merely a conceptual convenience: the matter will be discussed in § 7.

5.2.2. Secondly, it will be assumed that each learned event occupies a set of parallel fibres which may be regarded as having been chosen at random from the 200,000 which synapse with the Purkinje cell. This assumption can be justified on the grounds first that any estimation of capacity arrived at by using it is likely to be too low; and, secondly, that

almost any other assumption would involve a computational effort out of all proportion either to the probable truth or to the value of any results thereby achieved. The assumption is likely to be false for two reasons: first, the mossy fibre–granule cell relay is probably not randomly constructed; and, secondly, the learned events are unlikely to have a random structure over the space of possible mossy fibre inputs. Some types of structure on the learned inputs will positively confuse the system into giving false responses.

This topic will receive a full and more general analysis in a later paper: two remarks however are not out of place here. If there are many learned mossy fibre inputs which all overlap each other by a considerable amount, the cerebellum may not be able to discriminate inputs which have been learned from inputs which have not when an unlearned input has much overlapping with learned inputs. The fact that no granule cells have more than 7 claws introduces an absolute upper bound to the discriminatory power of the cerebellum: and, when this is inadequate, control must revert to the cerebrum. It should also be noted that any subset of a learned mossy fibre input will behave as a learned input if it causes the same codon size range to be selected as did the learned input. In the full model, the condition is more restrictive in that the codon size range must be the same for most of the 150 or so Golgi cells concerned.

5.2.3. The maximum desirable number of facilitated synapses on any one Purkinje cell will be taken as 140,000, and the minimum number of parallel fibres active in any learned event as 500.

These figures are related by the way the Purkinje cell threshold is set (see 6.1). It turns out that the most economical way of doing this is by sampling a population of parallel fibres closely related to and including the ones passing through the Purkinje cell dendritic tree. Let $T(E)$ be the threshold set in response to the stimulation of $M(E)$ parallel fibres by the mossy fibre input $E$ (regarded as an input to a particular Purkinje cell **P**). Let $f$ be the fraction of the (200,000) parallel fibre synapses which have been facilitated at **P**. If $E$ has been learnt, all $M(E)$ of the active parallel fibres will have facilitated synapses at **P**. Hence if $E$ is to be recognized as learnt, $T(E) \leqslant M(E)$ (i).

If $E$ is not a learned event, and $F(E)$ is the number of the active parallel fibres which have facilitated synapses at **P**, then $E$ will be ignored only if $T(E) > F(E)$ (ii).

If recognition is a reliable process, both (i) and (ii) must be true with high probability.

The randomness assumption 5.2.2 allows us to assume that $F(E)$. taken over events $E$ with constant $M(E)$, has a binomial distribution with expectation $fM(E)$. It is unlikely that $T$ (taken over the same event population)

has worse than a binomial distribution, since the binomial assumption is equivalent to regarding the value of $T$ as being set on the basis of a measurement of the number of active mossy fibres involved in $E$. If one assumes $T$ has a binomial distribution, conditions (i) and (ii) are satisfied with probability rather greater than 0·99 if $M \geqslant 500$, $f = 0{\cdot}7$. These values for $f$ and the minimum value of $M$ will be used despite the rather low confidence level because (*a*) the threshold setting mechanism is certain to be better than a binomial process, and (*b*) few input events will use the minimum number of parallel fibres allowed.

### 5.3. *The storage capacity of a Purkinje cell*

The capacity of a Purkinje cell may be calculated very simply from the assumptions 5.2. Suppose the fraction of facilitated parallel fibre synapses is 0·7, and each learned event occupies $n$ parallel fibres. Then $x$, the expected number of events which may be learned before the total proportion of synapses used exceeds 0·7, is the largest integer for which

$$(1 - n/200{,}000)^x > 0{\cdot}3.$$

$x$ has been computed for various values of $n$, and the results appear in Table 6. It will be seen that the advantage of having a small number of fibres active in each learned event is an enormous increase in capacity: the Golgi cell arrangement of local as well as global constraints on the codon size begins to make good sense. If the minimum number of parallel fibres active in learned event is 500, the average number of responses stored by each Purkinje cell is probably in excess of 200.

TABLE 6. $x$ is the number of events each occupying $n$ parallel fibres that can be learned by one Purkinje cell, i.e. $x$ is the largest integer for which

$$\left(1 - \frac{n}{200{,}000}\right)^x > 0{\cdot}3$$

| $n$ | 500 | 1,000 | 2,000 | 5,000 | 10,000 | 20,000 |
|---|---|---|---|---|---|---|
| $x$ | 480 | 240 | 119 | 47 | 23 | 11 |

### 5.4. *The Purkinje cell threshold*

The inhibition of the basket and outer stellate cells can be a powerful influence on the behaviour of a Purkinje cell. The present theory requires that the Purkinje cell should fire if and only if more than a proportion $p$ of the active parallel fibres have facilitated synapses with it, where $p$ is close to 1. It is proposed that the purpose of these stellate cells is to provide the appropriate inhibition, and that their peculiar axonal distribution is a device to secure an economy of dendrite by a factor of up to 20 (see §6).

It is made possible by the distribution of the mossy fibre terminals beneath them.

If $M$ of the parallel fibres synapsing with a particular Purkinje cell are active, and $Ms$ of these have been facilitated, then the cell must fire (or produce a burst of firing) if and only if $Ms \geqslant Mp$. The Purkinje cell thus has the superficially simple task of summing $Ms$ (represented by excitation) and $Mp$ (represented by inhibition). If, however, one reflects upon the enormous expanse of the Purkinje cell dendritic tree, it becomes apparent that to arrange such a summation might not be an easy problem of dendritic engineering. The example in Plate 1 makes it difficult to imagine how the junction of a spiny branchlet with the rest of the dendrite could carry accurate information about the number of active spines if this were large, for if the 100 nearest the junction were active, it is hard to see how (say) 10 at the end of the branchlet could make much difference, at least on any simple view of dendritic function. Such a system can only provide accurate summation for numbers of active synapses rather small compared with the total population.

This overload effect can be overcome locally if the number of active fibres is kept small: but it is bound to recur on a larger scale unless the numbers are kept very small. A further trick seems necessary, and the right one is probably to do the subtraction piecemeal: add up the outer contributions to $Ms$, subtract the outer $Mp$, and transmit the result to be added to the contribution to $Ms$ of the next region. This is the only way of subtracting $B$ from $A$ with large $A$ and $B$ but small $(A-B)$ without ever handling large numbers.

The distribution of the axon terminals of the basket and outer stellate cells is peculiarly well suited to this interpretation. The outer stellate cells effectively sample the activity in the outer half of the molecular layer and send their (inhibitory) contribution to $-Mp$ to a region quite high up the dendritic tree of the Purkinje cell. The basket cells sample about the inner third, sending their contribution to the soma; and the intermediate cells perform an intermediate task. The basket cell action represents the last stage in computing $(Ms-Mp)$, and one may assume that the numbers are then small enough for the coding from dendrite to soma to be adequate.

Interpretation of the function of simple summation within any reasonable theory of dendrites would be made easier by two hypotheses: first that the number $M$ of active parallel fibres was both small and reasonably constant; and secondly, that the excitation due to a facilitated synapse differed very little between synapses. In view of the existing Golgi cell arrangement and the great increase in capacity which is a consequence of having $M$ small, there are strong reasons why the first hypothesis should be true. And the

second may be a consequence of the fact that these synapses are all spine synapses, which possess a definite morphological uniformity.

### 5.5. *The Purkinje axon collaterals and the climbing fibre collaterals*

The axons of the Purkinje cells give off numerous fine collaterals which form two plexuses. The infraganglionic plexus lies below the Purkinje cell bodies, and its fibres run in a predominantly transverse direction. The supraganglionic plexus, which is fed both directly and by branches of fibres rising from the infraganglionic plexus, lies above the Purkinje cell bodies, and its fibres run in a predominantly longitudinal direction (Eccles *et al.* 1967, p. 178). Not a great deal is known about the distribution of the Purkinje cell collaterals, but it seems that at least in the vermis the spread of the collaterals in the longitudinal direction is small, whereas in the transverse direction it may be quite large; and the longer collaterals tend to join points of cortex to their corresponding contralateral points. (See Eccles *et al.* 1967, pp. 178ff., for a discussion and references.)

These collaterals have weak inhibitory synapses with basket and Golgi cells, and perhaps also very weak inhibitory synapses with other Purkinje cells (Eccles *et al.* 1967, pp. 184ff.). Their effect through the basket cells is to release Purkinje cells from inhibition, but their influence through the Golgi cells is more complicated. It is likely that this influence will ultimately be excitatory at a given Purkinje cell **P** only if most of the granule cells thereby released from Golgi inhibition have modified synapses at **P**: and this will be true only if **P** has already learned a number of mossy fibre inputs all quite similar to the current input.

The only obviously reasonable interpretation of the effect of these collaterals is that they tend to excite the Purkinje cells in the cortex to which they distribute; and in certain circumstances can loosen the discrimination exercised by those cells. The fact that a Purkinje cell $\mathbf{P}_1$ has just fired may be relevant in a borderline firing decision for **P** if **P** and $\mathbf{P}_1$ lie in closely related pieces of cortex: and the Purkinje axon collaterals provide a suitable means of distributing this information. They can help **P** overcome inhibition due, perhaps, to an unlearned mossy fibre input which it has previously received, or they can make **P** more likely to accept the current input even though it may not be exactly one which has been learned.

This view is not entirely satisfying, but it does provide an interpretation of the climbing fibre collaterals. These make weak excitatory synapses with the inhibitory interneurones of the cortex (Eccles *et al.* 1967, Table 1, p. 63), and perhaps with Purkinje cells. Their distribution is more local than that of the Purkinje axon collaterals (Eccles *et al.* 1967, p. 215), but their effect locally probably roughly balances them. Hence it could be

argued that when a climbing fibre is active, that is when synaptic modification is taking place, the effect of Purkinje axon collaterals is at least partly annulled, and so something nearer a true representation of the mossy fibre input is stored.

## §6. THE OUTER STELLATE AND BASKET CELLS

### 6.0. *Justification of their joint treatment*

The outer stellate and basket cells will be taken together under the general heading of stellate cells for the following reasons.

6.0.1. They are both inhibitory.

6.0.2. They both send axons to the Purkinje cells only.

6.0.3. They are both driven mainly by parallel fibres, and have analogous dendritic fields, the outer stellates being further out in the molecular layer.

6.0.4. They have very similar axon distributions; the outer cells synapse further up the Purkinje cell dendritic tree, and reach a little less far across the folium than the inner ones.

6.0.5. There exist many intermediate forms.

The discussion will also include in a general way the apparently rather weak 'on-beam' outer stellate cells whose axons terminate locally, though these will receive special mention.

### 6.1. *The function of the stellate cells*

The stellate cells together have the task of controlling the threshold of the Purkinje cells: they are powerful, and have to be, since if the overload ideas 5.4 are correct they have to be able to contain almost the maximal excitation that parallel fibre activity can evoke in the Purkinje cell dendritic tree. (This, it was argued, is achieved long before all the parallel fibres are active.) The quantitative relations between the number of parallel fibres active and the strength of the inhibition necessary have been discussed in 5.2.3, and reasons for the distribution of the terminals on the Purkinje dendritic trees have been proposed in 5.4. It remains only to sort out two points: the size, shape and position of the dendritic tree, and the distribution of the Purkinje cells to which the stellate cells send axon terminals.

If one naively set about constructing a threshold-setting cell to perform the function required by the present theory, one would propose one inhibitory cell per one or two Purkinje cells. Its axon would synapse with just the one (or two adjacent) Purkinje cells, and its dendritic field would at least be very close to that of its corresponding Purkinje cell. If there were such cells, however, their dendrites would have to be not only very

close to those of the relevant Purkinje cell, but also very nearly as extensive: this would be necessary in order to obtain a reliable measurement of the usually sparse parallel fibre activity.

The reason why the stellate cells are not arranged like this is that since such a dendritic tree would necessarily take up roughly as much room as does a Purkinje cell dendrite, the number of Purkinje cells that could be packed in any given length of folium would be about halved. The key to the success of the existing solution is that the rosettes of each mossy fibre are numerous and on the whole distributed transfolially in the granular layer. The actual mossy fibres that drive the cortex therefore change quite slowly across a folium, and they can be watched efficiently and economically by sampling the parallel fibre activity across it (Fig. 2).

There is no quantitative evidence available from which one might investigate the tenability of this hypothesis: one can only estimate the economies to which the proposed sampling technique may lead. Each Purkinje cell receives inhibition from about 40 stellate cells: the inhibition to the Purkinje cell is therefore driven by a dendritic field about 40 times as large as that of a single stellate cell. If these 40 were distributed randomly just next to the Purkinje cell, a good sample (approximately $1 - 1/e^2 = 0{\cdot}86$) would be obtained if each tree had even 1/20 of the synapses that a Purkinje cell has. If the mossy fibre distribution alters slowly (which it has to do anyway for the system to work), the saving in dendrite could therefore be a factor of up to 20; and, in practice, the sampling is certainly not random.

### 6.2. *The stellate cells with local axonal distribution*

It is convenient to complete the review of the cells of the cortex with some remarks about the time courses of the excitatory and inhibitory synaptic actions. It is evident that the time course of transmitter action at a Purkinje cell is the ultimate factor determining the temporal extent of the influence on that cell of information from that fibre.

At a normal sort of synapse, such influence would not be expected to continue more than 20 msec after activity in the afferent axon had ceased: but so short a period would seem inappropriate for real-time analysis of events with characteristic times rarely less than 100 msec. The observed time course at a parallel fibre–Purkinje cell synapse is of the order of 100 msec, and Eccles *et al.* (1967) mention (p. 70) that this may be one function of spines. Now the connectivity of cerebellar cortex is such that the onset of the various post-synaptic effects at a Purkinje cell due to a mossy fibre signal is likely to be both slow and patchy: the various components arrive along paths with different latencies, and there may be build-up effects in the synapses themselves. Similar factors will affect the way the post-

synaptic effects decay at the end of a mossy fibre signal. All these effects can only be disruptive, at least as far as the present theory is concerned. The fact that the various effector circuits in the rest of the nervous system are geared only to recognizing bursts from Purkinje cells will minimize the effect of any stray impulses which might for a number of reasons leak out: but it is conceivable that the effects of an input during a 'turning on' or 'turning off' period could cause a false response from a Purkinje cell, and that response could last up to 20 msec.

In order that false outputs of this sort should not occur, it is necessary that the build up of inhibition at a Purkinje cell should occur faster than the build up of excitation, and that the IPSP should last longer than the EPSP. The latter is an observed phenomenon, with IPSP time courses up to 500 msec, EPSP ones up to about 100 msec; and it is possible that one function of the 'on-beam' stellate cells is to ensure the former. These cells have a local axonal distribution, so their axons are relatively very short; and many of their synapses with the parallel fibres are direct (i.e. not spine synapses). The first factor must, and the second may favour a fast production of IPSP at the Purkinje cell dendrite, and this IPSP could well arrive early enough to counteract the initial build up of EPSP from the Purkinje spines. The IPSP induced by these cells is weak, but by the time the Purkinje cell is turned on to any appreciable extent, the other stellate cells will also be active. It is therefore proposed that the weak on-beam stellate cells be interpreted as a device to prevent a false initial response by the Purkinje cell.

## §7. CEREBELLAR INPUT–OUTPUT RELATIONS

There are two main types of cerebellar input–output relation which are compatible with the present cortical theory and they are described separately.

### 7.1. *Learned movements*

The first possibility is the one suggested in §1, and concerns the learning of that particular sort of motor skill which may be described as a movement. During learning, the cerebrum organizes the movement, and in so doing, causes the appropriate olivary cells to fire in a particular sequence. This causes the Purkinje cells to learn the contexts within which their corresponding elemental movements are required, so that next time such a context occurs the mossy fibre activity stimulates the Purkinje cell, which evokes the relevant elemental movement.

This scheme imposes severe restraints upon the nature of the stimulus that may drive an olivary cell: indeed, almost the only permissible case is that in which each olivary cell is driven by a collateral of a cerebral command

fibre for some elemental movement. This statement may be justified by the following argument. During execution of a learned movement, the mossy fibre activity is responsible for the initiation of the various elemental movements: and it is therefore essential that, during learning, the Purkinje cell is associated with the context occurring just *before* its elemental movement. The present theory suggests that the granule cells and Golgi cells together provide extremely effective pattern discrimination: so the mossy fibre activity must be virtually the same during cerebellar execution of a movement as it was while that movement was being learnt. Hence, for the cerebellum to be able to learn a movement in which the contexts change rapidly, the olivary activity during learning has to be driven by impulses effectively synchronized with the commands. This conclusion can only be avoided in one of the two following ways: either some delay is specially introduced into the mossy fibre afferents, or the olivary cells are driven by the elemental movement just preceding the current one. The first assumption is unlikely on grounds of efficiency, and the second would require a probably unacceptable number of olive cell-Purkinje cell pairs, one for each sequence of two elemental movements.

The above argument, however, cannot be applied to those situations where the contexts are changing very slowly: and in such cases it is at least logically possible for an olivary cell to be driven by a signal which was slightly later than the command signal during learning, since the relevant context will scarcely have altered. It is therefore not impossible for an olivary cell to be driven by a receptor which is sensitive to the movement initiated by its corresponding Purkinje cell: although, if the contexts do change slowly, a context driven system will not reproduce the timings of the stages in a movement at all accurately, and so cerebellar learning will anyway be rather bad.

It can therefore be concluded that an olive cell–Purkinje cell pair, whose olive cell is driven by a receptor, is unlikely to be used for learning motor skills involving much movement. It is however well known that the inferior olive is divided into two portions, one driven by descending fibres (Walberg, 1954) and one by ascending fibres (Brodal 1954). Further, it is known that at least some cells in the 'ascending' or 'spinal' part of the olive are driven by receptors (Armstrong, Eccles, Harvey & Matthews, 1968), and these authors also demonstrate the convergence at some cells of impulses from receptors of quite different types.

If the present cortical theory is correct, and the cerebellum does learn motor skills, there is only one situation in which it is not absurd to drive the olivary cells by receptors rather than by cerebral command fibre collaterals, and that is when the cerebellum is required to carry out an action in a different language from that in which the cerebrum originally

set it up. This condition is likely to be fulfilled in the cerebellar control of balance or posture, where one may reasonably expect the cerebrum to deal in a language oriented towards the problem of changing postures, while the cerebellum is concerned primarily with maintaining an achieved posture. It is to this kind of control that the second form of input–output relation is particularly well suited, and it will be discussed in detail in 7.2.

If it is assumed that such situations are best dealt with by the methods described in 7.2, the following conclusion may be drawn. Where the cerebellum is required to learn a motor skill consisting of a movement, the cells of the inferior olive should be driven by the equivalent of a collateral of the cerebral command fibre for a particular elemental movement; and the Purkinje cell corresponding to that olivary cell should be able to provoke that same elemental movement. The particular elemental movement associated with an olivary cell–Purkinje cell pair need not be fixed, but it presumably is: and the elemental movements associated with this kind of input–output relation are probably mostly small movements.

To complete the study of this kind of input–output relation, four further points must be discussed. The first concerns a possible variant in the way information is read out of the cerebellum. It was assumed in §5 that the level of inhibition at a Purkinje cell was generally rather low, and that mossy fibre activity involved in a learned context was enough to produce a signal in the Purkinje cell axon. There is another possibility, in which the level of inhibition at a Purkinje cell is generally rather high, and the rest of the brain decides whether the current context has been learned by observing the results of a climbing fibre impulse. If the mossy fibre input has been learned, the Purkinje cell gives a large response; if not, it gives a small one and the effector circuits respond accordingly. This form of output may be described as *inhibition sampling*, and has effectively been suggested by Eccles *et al.* (1967, p. 177), though not in the context of modifiable synapses.

The second point concerns the command circuit used by the cerebrum while setting up a movement. It is possible that the olivary cells are literally driven by collaterals of the cerebral command fibres: but it is also possible that the command circuit actually is the cortico-olivo–Purkinje cell–effector circuit path. This hypothesis involves no difficulties and is especially attractive if Purkinje cell output is obtained by inhibition sampling: for this could then easily be achieved by uniform weak descending activity to the inferior olive, arriving by the same pathways as are used for the cerebral organization of movements. One extra hypothesis is also needed if this system is postulated, namely, that the mechanism of synaptic modification at Purkinje cells is sensitive only to intense climbing fibre activity.

The third question arises from the possibility that synaptic modification may be subject to gradual decay. This might be necessary, in view of the limited learning capacity of a Purkinje cell: and one might imagine that repetition of a movement carried out even under cerebellar control should have some reinforcing effect. If a Purkinje cell command were somehow fed back to excite the relevant climbing fibre, a reinforcing effect would certainly be obtained, but lack of this feed-back would not rule out the possibility that a reinforcing effect exists, since this depends on the details of the synaptic modification mechanism.

The final point to be raised in this discussion of the cerebellar control of movements is the question of the speed with which such movements are executed. There is no reason why a context dependent system should not be run at different speeds, and if the extra postulate were made of some general intensity control acting uniformly over the effector circuits, a movement learnt at one speed could be performed at another. (This idea would fit nicely with the suggestion made above that during cerebellar control of a movement, the olive receives uniform weak descending activity, for the strength of the Purkinje cell output would then depend upon the strength of this uniform activity.) It is, however, likely that if the time course of a movement were changed substantially, some relearning would be necessary.

### 7.2. *Learned conditional reflexes*

The explanation of the second type of input–output relation compatible with the cortical theory is much simplified by the introduction of a new idea, which extends the classical notion of reflex.

*Definition. A conditional reflex* is a reflex which operates when, and only when, certain conditions outside the reflex arc are satisfied. These conditions are the *context* of the conditional reflex, and a *learned conditional reflex* is a conditional reflex whose context is learned.

An ability to acquire learned conditional reflexes would make the task of maintaining balance and posture very much easier for the nervous system. For example, consider the problems which confront a child as he learns to stand. It would greatly aid him if he could form a reflex circuit which connected a vestibular signal indicating some imbalance directly to an order for the appropriate compensating movement: this, however, could not be a true reflex since the child will not always wish to stand. The appropriate form of control is a conditional reflex whose context is the state of standing, and which therefore only operates while the child is standing. In order to suspend the standing reflexes, the child has only to disrupt the 'standing' context, and this could be done, for example, by his suddenly wishing to stand no longer.

A particular olivary cell–Purkinje cell pair may be interpreted as a storage unit for a conditional reflex if and only if the circuit environment → receptor → olivary cell → Purkinje cell → effector → environment is a stabilizing negative feed-back loop when activated by a learned mossy fibre input. The context represented by the learned mossy fibre input is the context of the conditional reflex. Explicitly, the conditions for storage of a conditional reflex are as follows:

(i) Output is obtained from the Purkinje cell by the inhibition sampling method (described in 7.1): that is, the level of inhibition is generally high, so that climbing fibre signals are only transmitted when the mossy fibre input is one that has been learned.

(ii) The olivary cell is driven by receptors whose stimulation is reduced (in any learned context) by the results of stimulating the corresponding Purkinje cell.

The learning of a context will arise if the combination of olivary cell firing and that particular context is a frequent one, as it would be, for example, while the child (under cerebral control) was 'learning' to stand. Once the context is learned, the reflex automatically becomes operative when it is required.

There is no reason why a particular olivary cell should not be driven by more than one kind of receptor, though receptors must be connected to Purkinje cell units whose activity reduces the stimulations they receive: the inhibitory nature of the Purkinje cell output may help to arrange this. The receptors connected to a given olivary cell have to be rather carefully chosen, but their number is limited only by the learning capacity of the corresponding Purkinje cell.

It is proposed that most cerebellar functions associated with maintaining balance and posture are carried out by forming the appropriate learned conditional reflexes in the sense of 7.2, while those motor skills which involve active movement rather than maintenance reflexes are generally learned in the manner described in 7.1.

### 7.3. *The cerebellar initiation of movements*

The two kinds of input–output relation give the cerebellum the power to learn any task whose execution is related in a rather rigid way to information sent through the mossy fibres, and at the same time to set up suitable reflexes to maintain balance and posture during execution of those tasks. The cerebrum is thus freed from at least the routine matters associated with motion and stance. There are, however, many instances in life when both the recognition that a job must be done, and its implementation, are simple operations. For example, information taken out of the visual system at a fairly low cortical level (say from areas 18 and 19) might be

useful as a source of cues during walking: and information about the mood one is in can sometimes influence in a simple (but learned) way the gestures one makes.

It is but a short step from believing that the cerebellum stores movements and gestures to proposing that visual cues and information about mood and so forth can form enough of a context actually to initiate an action; and it would be strange if something of this sort did not happen, though it doubtless occurs more frequently in the motor cortex. Where it is possible to translate the combined activity of many cerebral fibres rather simply into physical directives, doing so in the cerebellum would free the cerebrum from an essentially tedious task. In these circumstances, the cerebellum becomes rather more than a slave which copies things originally organized by the cerebrum: it becomes an organ in which the cerebrum can set up a sophisticated and interpretive buffer language between itself and muscle. This can be specially tailored to the precise needs of the animal, and during later life leaves the cerebrum free to handle movements and situations in a symbolic way without having continually to make the retranslation. The automatic cerebellar translation into movements or gestures will reflect in a concrete way what may in the cerebrum be diffuse and specifically unformulated, while the analysis leading to that diffuse and unformulated state can proceed in its appropriate language.

## §8. THE MAIN PREDICTIONS OF THE THEORY

### 8.1. *Modifiable synapses*

The main test of the theory is whether or not the synapses from parallel fibres to Purkinje cells are facilitated by the conjunction of presynaptic and climbing fibre (or post-synaptic) activity (5.1). If this is not true, the theory collapses.

It is likely that no other cerebellar synapses are modifiable. The mossy fibre-granule cell synapses are discussed in 3.3.3, and the Golgi cell afferent synapses in 4.5. The function of the stellate cells is fixed throughout the life of the cerebellum, and so they probably do not possess modifiable synapses. Though it is difficult to see how these predictions could be wrong, they might be: such a disproof would be embarrassing but not catastrophic, since something of the bones of the theory would remain.

### 8.2. *Cells*

The roles of the various cells in the cortex are roughly determined once the main prediction about modifiable synapses is verified. There are, however, three predictions which can be tested. The first concerns the Golgi cells. They have been discussed at some length, and arguments were

produced for the view that there should be little if any summation between the upper and lower dendritic trees (4.4). The cell should be driven by that tree which is currently the more powerfully stimulated. Refutation of this would again be awkward but not fatal.

Secondly, the interpretation of the stellate cells as a threshold setting mechanism for the Purkinje cells depends strongly upon the presumed distribution of the mossy fibre rosettes below the cortex. The theory requires that each mossy fibre extends a fair distance perpendicular to the line of the folium (6.1), and this can be investigated.

Thirdly, the number of granule cells active at any one time (say in any 50 msec period) is a small fraction (less than 1/20) of all granule cells.

### 8.3. *Input–output relations*

The two forms of input–output relation are experimentally distinguishable, and the same olivary cell–Purkinje cell pair may at different times be used both ways. For the learned movement form, the olivary cell should respond to a command for the same elemental movement as is initiated by the corresponding Purkinje cell. For learned conditional reflexes, the activity provoked by the Purkinje cell must tend to cause a reduction in the receptor activity which drives the olivary cell.

It is proposed that these two input–output relations are used for fairly different tasks. This division of labour is not logically necessary, since in principle each form can execute either task: but it would be surprising if the observed division differed substantially from the one suggested, since that particular arrangement is the most economical.

## §9. THE CODON REPRESENTATION

The notion central to the present theory is that the afferent input events communicated by the mossy fibres to cerebellar cortex are turned into a language of small subsets and then stored; and this has been called the codon representation of an input. This formulation is new, but the principle is closely related to the feature analysis ideas current in the machine intelligence literature (see e.g. Uhr & Vossler, 1961). 'Features' are merely rather specially chosen codons. This author in particular owes a debt to the paper by Brindley (1969) which contains what may be regarded as a degenerate case of codon representation, though from rather a special point of view.

The idea of codons arose in an unlikely way as the result of a search for a representation which the cerebrum might use for storing information. Its relevance to the cerebellum was noticed only when it was realized that any neural net built to implement the representation must contain something

like granule cells. An analysis of the properties of the codon representation and of its possible place in the theory of cerebral cortex will form the subject of a later paper.

I wish to thank Professor G. S. Brindley and Dr I. M. Glynn for their very helpful criticism; Professor Sir John Eccles and Springer-Verlag for permission to use Figs. 1 and 3, and C.S.I.C. Madrid for permission to use Fig. 2. Most of this work was carried out during the author's tenure of an M.R.C. research studentship, and formed part of a fellowship dissertation offered to Trinity College, Cambridge in August 1968. The ideas of §7.2 were formulated later to overcome criticism made by S. J. W. Blomfield and Professor G. S. Brindley. This work was supported by a grant from the Trinity College research fund.

## REFERENCES

ARMSTRONG, D. M., ECCLES, J. C., HARVEY, R. J., MATTHEWS, P. B. C. (1968). Responses in the dorsal accessory olive of the cat to stimulation of hind limb afferents. *J. Physiol.* **194**, 125–145.

BRINDLEY, G. S. (1964). The use made by the cerebellum of the information that it receives from sense organs. *Int. Brain. Res. Org. Bulletin* **3**, 80.

BRINDLEY, G. S. (1969). Nerve net models of plausible size that will perform many of very many simple learning task. *Proc. R. Soc.* B. (In the Press.)

BRODAL, A. (1954). Afferent cerebellar connections. In *Aspects of Cerebellar Anatomy*, ed. JANSEN, J. & BRODAL, A. ch. II, pp. 82–188. Oslo: Johan Grundt Tanum Forlag.

CAJAL, R. Y. (1911). *Histologie du Système Nerveux*, Tome II, 1955 edn., p. 57, C.S.I.C.: Madrid.

ECCLES, J. C., ITO, M. & SZENTAGOTHAI, J. (1967). *The Cerebellum as a Neuronal Machine.* Berlin: Springer-Verlag.

ESCOBAR, A., SAMPEDRO, E. D. & DOW, R. S. (1968). Quantitative data on the inferior olivary nucleus in man, cat and vampire bat. *J. comp. Neurol.* **132**, 397–403.

HEBB, D. O. (1949). *The Organization of Behaviour*, p. 62. New York: Wiley.

UHR, L. & VOSSLER, C. (1961). A pattern recognition program that generates, evaluates and adjusts its own operators. *Proc. west. jt. Computer Conf.* **19**, 555–569.

WALBERG, F. (1954). Descending connections to the inferior olive. In *Aspects of Cerebellar Anatomy*, ed. JANSEN, J. & BRODAL, A., ch. IV, pp. 249–263. Oslo: Johan Grundt Tanum Forlag.

## ADDENDUM

Escobar, Sampedro & Dow (1968) have shown that in man, and probably also in cat, there are fewer cells in the inferior olive than there are cerebellar Purkinje cells. There may therefore exist other sources of climbing fibres. Statements in the present work about the inferior olive should be understood to refer to all sources of climbing fibres, including those as yet undiscovered. If olivo-cerebellar fibres are found to branch, the theory will require slight modification.

## EXPLANATION OF PLATE

Dendritic spines on a cat Purkinje cell. (From Eccles *et al.* 1967, Fig. 27*A*.)


DAVID MARR

(*Facing p.* 470)

**W. Thomas Thach**

Commentary on

# *A Theory of the Cerebellar Cortex*

Marr may not have been the first to suggest that the cerebellum dealt with learning and memory (Luciani, 1915; Rawson, 1932; Brindley, 1964), but he clearly was the first to propose a theory. The theory began with the intuitive observations of Brindley on the nature of the acquisition of skilled sequential acts. The performance of such acts passed from the attentive slow, picking out of one movement after another to the unconscious rapid uninterrupted flow of them: with time, they became essentially automatic. The elements of the machinery consisted of the cell types, the connectivities and the synaptic actions of the cerebellar cortex, as newly illuminated by Eccles, Llinas and Sasaki and others (Eccles, et al., 1967; Llinas, 1981). The process was one of context recognition and learning: the context recognition at the level of the mossy fiber-granule cell-golgi cell circuitry, and the learning at the level of the parallel fiber-Purkinje cell synapse, heterosynaptically reinforced by the inferior olive climbing fiber. Linked through learning to the context of the prior movement in the sequence, the Purkinje cell automatically recognized it and it fires to trigger the next movement in the sequence. The model was described in lucid, even vivacious language; critical predictions were itemized and rank ordered with one to four stars, as for generals and gourmet guidebooks. The model stimulated thought, experiment and controversy, all of which continue to this day.

Other theoretical papers soon appeared. Albus (1970) offered a similar model, explicitly likened to a Perceptron. Two differences between the Albus and the Marr models were 1) that Albus had the learning and the climbing fiber discharge driven by error (rather than intent), and 2) that the learning consisted of decreasing (rather than increasing) the parallel fiber-Purkinje cell synaptic strength, thus decreasing (rather than increasing) the output of the Purkinje cell that was associated with erroneous performance. Gonshor and Melvill-Jones (cf. 1976) showed in humans that the vestibulo ocular reflex (VOR) is indeed exquisitely adjustable. Ito (1972) adapted the Marr-Albus Theory in an attempt to account for the adjustment of the vestibulo-ocular reflex; Gilbert (1975) adapted the Theory to portray Purkinje cell learning in the spike frequency domain.

Experimentalists tested the model by ablation (Does the learning go away?), single unit recording (Does neural discharge correlate with learning?), and electrical stimulation (Can stimulation reproduce learning?). Ito (1974) et al. in the rabbit and Robinson (1976) in the cat ablated the cerebellar cortex, altered

the gain of the VOR, and prevented further adjustment. The experiment has been repeated in a variety of animals since. Other movement adaptations have also been altered by cerebellar cortex ablation, including the ocular saccades (Optican and Robinson, 1980), the conditioned eye blink (McCormick and Thompson, 1984; Yeo, et al., 1984) and the habituation of the acoustic startle response (Leaton and Supple, 1986).

Single unit recording during motor adaptations added two critical observations: first, the complex spike of the Purkinje cell (caused by climbing fiber discharge) was reported to occur preferentially in situations where adaptation is occurring (Gilbert and Thach, 1977; Thach, 1980; Watanabe, 1984) or is likely to occur (Gellman et al., 1985; Armstrong and colleagues; Andersson and Armstrong, 1987; Armstrong et al., 1988; Simpson and Alley, 1974). Second, repeated occurrence of the complex spike was reported to be associated with a reduction in the occurrence of the simple spike caused by the parallel fiber input (Gilbert and Thach, 1977; Watanabe, 1984) as predicted by the Albus model.

Electrical stimulation conjointly of climbing fiber and mossy fibers in the decerebrate cat (Ito et al., 1982) gave a result similar to that observed in the living animal (Gilbert and Thach, 1977). Coupled stimulation of climbing fibers and mossy fibers led to reduced efficacy of those mossy fibers in activating (through granule cell parallel fiber synapses) Purkinje cells. The frequency of stimulation, the required number of pairings, and the time course of the learning were similar in the electrical stimulation and in the natural behavioral adaptation experiments. The observation has been confirmed and extended in a variety of reduced preparations, including direct stimulation of parallel fibers coupled with climbing fibers in cerebellar slices (Rawson and Tiloskulchai, 1982; Eckerot, 1985; Ekerot and Kano, 1985; Kano and Kato, 1987). Attempts are being made to examine ionic/molecular membrane mechanisms whereby the climbing fiber could produce the reduction in parallel fiber efficacy. The climbing fiber discharge appears to release adenosine and aspartate (Cuenod et al., 1988), which causes an inward calcium current in the Purkinje cell dendrite (Llinas and Nicholson, 1971; Llinas and Hess, 1976; Llinas and Sugimori, 1980 a,b). The parallel fiber releases glutamate, and glutamate sensitivity of the Purkinje cell is reduced by climbing fiber action (Ito et al., 1982).

These results suggest that the Marr Theory, as amended by Albus, may be generally correct. Nevertheless, there are many investigators who disagree with this conclusion for a number of different reasons. One group objects to the adequacy of ablation in general and to some experiments in particular in establishing proof (Harvey and Welch, 1988). While it is now widely accepted that cerebellar cortical ablation abolishes some kinds of motor adaptation, some argue that the result is non-specific, and that the plastic synapses located elsewhere and possibly widely distributed (Llinas, 1981; Lisberger, 1988). As for the conditioned eye blink, it has been claimed that the ablation so impairs performance as to lead falsely to the interpretation that learning is

impaired, and that controlled studies are needed to dissociate the learning from the performance of movements (Harvey and Welch, 1988).

Another group objects that single unit recordings in the awake, VOR-adapted monkey have failed to show the gain change at the level of the Purkinje cell that should, according to the theory, account for the adaptation (Miles et al., 1980; Lisberger, 1988). Others raise the question of whether these recordings were done in the appropriate part of the cerebellum (Gerrits and Voogd, 1989). Lisberger has reported that patterns of neural discharge sufficient to explain the adaptation are seen only in the vestibular nuclei, and that the modified synapse cannot be in the cortex but rather must be the vestibular nerve synapse onto vestibular nuclear cells (Lisberger, 1988). Ebner and Bloedel (1981, 1983) have shown that climbing fiber activity may cause a short-term change in the gain of the Purkinje cell response to parallel fiber input. Nevertheless, they apparently prefer not to believe that this is a mechanism for "motor learning".

Finally, some object that the stimulation experiments of conjunction of climbing fiber and mossy or parallel fiber activities have not (in their own hands) been repeatable, or are insufficient to account for learning (Llinas, 1970, 1981; Llinas and Volkino, 1973; Llinas, et al., 1975).

Only time and work can answer these objections. Certainly some elements of Marr's Theory may require further modification. Yet, a growing number of network theoreticians and experimental neuroscientists appear to like the ideas, and to anticipate their being proven to be essentially and substantially correct. But whether the Theory is right or wrong, it has been useful, and is a fitting monument to the genius of David Marr.

## REFERENCES

Albus JS (1971): A theory of cerebellar function. *Math Biosci* 10:25-61

Andersson G, Armstrong DM (1987): Complex spikes in Purkinje cells in the lateral vermis (b zone) of the cat cerebellum during locomotion. *J Physiol* 385:107-134

Armstrong DM, Edgley SA, Lidierth M (1988): Complex spikes in Purkinje cells of the paravermal part of the anterior lobe of the cat cerebellum during locomotion. *J Physiol* 400:405-414

Brindley GS (1964): The use made by the cerebellum of the information that it receives from the sense organs. *IBRO Bull* 3(3):80

Cuenod M, Do KQ, Vollenweider F, Streit P (1988): Cerebellar climbing fibers: excitatory amino acid and adensoine release. *Neurobiology of the Cerebellar Systems: A Centenary of Ramón y Cajal's Description of the Cerebellar Circuits.* p. 26 (Abstr.)

Ebner TJ, Bloedel JR (1981): Role of climbing fiber afferent input in determining responsiveness of Purkinje cells to mossy fiber inputs. *J Neurophysiol* 45:962-971

Ebner TJ, Yu QX, Bloedel JR (1983): Increase in Purkinje cell gain associated with naturally activated climbing fiber input. *J Neurophysiol* 50:205-219

Eccles JC, Ito M, Szentagothai J (1967): *The cerebellum as a neuronal machine.* New York: Springer-Verlag, Inc

Ekerot CF (1985): Climbing fiber actions of Purkinje cells—plateau potentials and long-lasting depression of parallel fiber responses. In: *Cerebellar Functions* Bloedel JR, Dichgans J, Precht W, ed. New York: Springer-Verlag

Ekerot C-F, Kano M (1985): Long-term depression of parallel fibre synapses following stimulation of climbing fibres. *Brain Res* 342:357-360

Gerrits NM, Voogd J (1989): The topograhical organization of climbing and mossy fiber afferents in the flocculus and the ventral paraflocculus in rabbit, cat and monkey. *Exp Brain Res Series* 17:26-29

Gellman R, Gibson AR, Houk JC (1985): Inferior olivary neurons in the awake cat: detection of contact and passive body displacement. *J Neurophysiol* 54:40-60

Gilbert PFC (1974): A theory of memory that explains the function and structure of the cerebellum. *Brain Res* 70:1-18

Gilbert PFC, Thach WT (1977): Purkinje cell activity during motor learning. *Brain Res* 128:309-328

Gonshor A, Melvill-Jones G (1976): Extreme vestibulo-ocular adaptation induced by prolonged optical reversal of vision. *J Physiol Lond* 256:381-414

Harvey JA, Welch JP (1988): Cerebellar regulation of the conditioned and unconditioned nicitating membrane reflex: analysis of sensory, associative and motor functions after reversible and irreversible cerebellar lesions. *Neurobiology of the Cerebellar Systems: A Centenary of Ramón y Cajal's Description of the Cerebellar Circuits.* p. 36 (abstr.)

Ito M (1972): Neural design of the cerebellar control system. *Brain Res* 40:81-84

Ito M, Sakurai M, Tongroach P (1982): Climbing fibre induced depression of both mossy fiber responsiveness and glutamate sensitivity of cerebellar Purkinje cells. *J Physiol Lond* 324:113-134

Ito M, Shiida TN, Yamamoto M (1974): The cerebellar modification of rabbit's horizontal vestibulo-ocular reflexin induced by sustained head rotation combined with visual stimulation. *Proc Japan Acad* 50:85-89

Kano M, Kato M (1987): Quisqualate receptors are specifically involved in cerebellar synaptic plasticity. *Nature 325:276-279*

Leaton RN, Supple WF, Jr. (1986): Cerebellar vermis: essential for long-term habituation of the acoustic startle response. *Science* 232:513-515

Llinas R (1970): Neuronal operations in cerebellar transactions. In: *The Neurosciences: Second Study Program.* Schnitt FO, ed. New York: Rockefeller University Press, pp 409-426

Llinas R (1981): Electrophysiology of cerebellar networks. In: *Handbook of Physiology*, Section 1, Volume II, Part 2. Brooks, VB ed. pp 831-876 Llinas R, Hess R (1976): Tetrodotoxin-resistant spikes in avian Purkinje cells. Proc Nat Acad Sci 73:2520-2523

Llinas R, Nicholson C (1971): Electrophysiological properties of dendrites and somata in alligator Purkinje cells. *J Neurophysiol* 34:532-551

Llinas R, Sugimori M (1980a): Electrophysiological properties of *in vitro* Purkinje cell somata in mammalian cerebellar slices. *J Physiol* 305:171-195

Llinas R, Sugimori M (1980b): Electrophysiological properties of *in vitro* Purkinje cell somata in mammalian cerebellar slices. *J Physiol* 305:197-213

Llinas R, Volkind RA (1973): The olivocerebellar system: functional properties as revealed by harmaline-induced tremor. *Exp Brain Res* 18:69-87

Llinas R, Walton K, Hillman DE, Sotelo C (1975): Inferior olive: its role in motor learning. *Science* 190:1230-1231

Lisberger SG (1988): The neural basis for learning of simple motor skills. *Science* 242:728-735

Luciani, L (1911-1924): *Human Physiology*. Welby FA, trans. London: MacMillan and Co., Ltd

Marr D (1969): A theory of cerebellar cortex. *J Physiol* 202:437-470

McCormick DA, Thompson RF (1984): Cerebellum: essential involvement in the classically conditioned eyelid response. *Science* 223:296-299

Miles FA, Fuller JRH, Braitman DJ, Dow BM (1980): Long-term adaptive changes in primate vestibulo-ocular reflexes III. Electro-physiological observations in flocculus of adapted monkeys. *J Neurophysiol* 43:1437-1476

Optican LM, Robinson DA (1980): Cerebellar-dependent adaptive control of primate saccadic system. *J Neurophysiol* 44:1058-1080

Rawson NR (1932): The story of the cerebellum. *Canad MAJ.* 26:220-225

Rawson JA, Tilokskulchai K (1982): Climbing modification of cerebellar Purkinje cell responses to parallel fiber inputs. *Brain Res* 237:492-497

Robinson DA (1976): Adaptive gain control of the vestibulo-ocular reflex by the cerebellum. *J Neurophysiol* 39:954-969

Simpson JI, Alley KE (1974): Visual climbing fiber input to rabbit vestibulo-cerebellum: a source of direction-specific information. *Brain Res* 82:302-308

Thach WT (1980): Complex spikes, the inferior olive, and natural behavior. In: *The Inferior Olivary Nucleus*. Courville J, ed. New York: Raven, pp 349-360

Watanabe E (1984): Neuronal events correlated with long-term adaptation of the horizontal vestibulo-ocular reflex in the primate flocculus. *Brain Res* 297:169-174

Yeo CH, Hardiman MJ, Glickstein M (1984): Discrete lesions of the cerebellar cortex abolish classically conditioned nictitating membrane response of the rabbit. *Behav Brain Res* 13:261-266

*Professor*
*Washington University*
*School of Medicine*
*St. Louis, Missouri*

# How the Cerebellum may be Used

by
STEPHEN BLOMFIELD
DAVID MARR

Department of Physiology,
Institute of Psychiatry,
De Crespigny Park,
London SE5

Recent anatomical information suggests new input–output relations for the cerebellum. These have interesting implications about the role of motor cortex in the learning and controlling of voluntary movements.

THE vertebrate cerebellar cortex has a very uniform structure, and may, for the purpose of this article, be regarded as being composed of many units like that appearing in Fig. 1. Its only output is the projection of large inhibitory cells, the Purkinje cells (*Pu*), to the intracerebellar nuclei, and to some of the vestibular nuclei[1,2]. In man, a major projection from the intracerebellar nuclei is to the ventro-lateral nucleus of the thalamus (*VL*)[1,2]. *VL* cells project to the motor cortex.

There are two kinds of input to the cerebellar cortex: the mossy fibres, which synapse with the numerous granule cells; and the climbing fibres, which project directly to the Purkinje cells and wrap themselves around their dendrites. Each Purkinje cell receives one climbing fibre[1], and can be powerfully excited by it. The climbing fibres arise from a group of cells in the contralateral brain stem[1]; the curious shape of this group has led to its being named the olive. The inferior olive (*IO*) receives connexions from a wide variety of sources, in particular from the cerebral cortex[2]. The mossy fibres have several different sites of origin[3]; particularly important are the pontine nuclei (*PN*) of the brain stem. The cerebellar granule cells, with which the mossy fibres synapse, send axons (the parallel fibres) to the Purkinje cells, and to the inhibitory interneurones of the cortex.

In a recent article[3], it was shown that the known anatomy and physiology of the cerebellar cortex are consistent with its interpretation as a simple memorizing device. It was predicted that the synapses between parallel fibres and Purkinje cells are modifiable, being facilitated by the conjunction of pre-synaptic and climbing fibre activity. It was shown how this would allow any single Purkinje cell to learn to recognize, without appreciable confusion, more than 200 different mossy fibre input patterns. Two methods were outlined by which such a memorizing device might learn to perform motor actions and maintain voluntary postures initially organized elsewhere. Since then, three relevant facts have come to our attention: (i), anatomical information concerning the origin of the cortico-olivary and cortico-pontine projections[4]; (ii), the discovery that the olivo-cerebellar (that is, climbing) fibres branch[5,6]; and (iii), the prediction that climbing fibres can organize more than simple memorizing phenomena[7]. These facts have implications about the way the cerebellum may be used by the rest of the nervous system that will be of interest to experimenters, and we therefore give here an outline of their principal consequences.

## New Information

(i) The origin of the descending projection to the olive has long been known to include cortical cells, of which the majority lie in the motor and pre-motor areas. But it has recently been shown that these fibres arise almost

Reprinted by permission from *Nature* Volume 227, pp 1224-1228. 

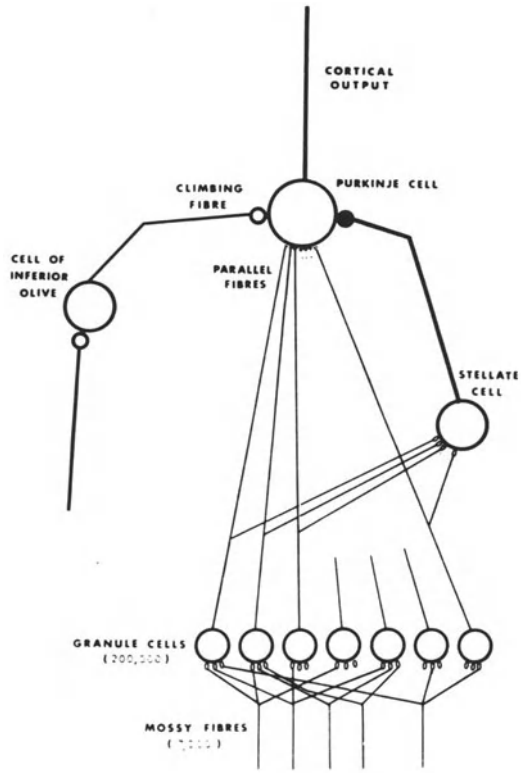



Fig. 1. The diagram selects the principal elements of the cerebellar cortex. There is one output system, the Purkinje cell axons, and two input systems, the climbing and the mossy fibres. The climbing fibres originate in the inferior olivary nucleus, and each Purkinje cell usually receives exactly one. The mossy fibres, which come from many parts of the body and brain, are imagined to convey information about the state of the animal—information referred to as the "context" at that time. The mossy fibre input is translated by the granule cells into a language of subsets, and the granule cell axons become the parallel fibres. It is predicted that the parallel fibre synapses with a Purkinje cell that are coactive with its climbing fibre are facilitated. The inhibitory cell prevents the Purkinje cell from firing unless almost all its active afferent synapses have been facilitated. The numbers of the various kinds of fibre projecting to one Purkinje cell in cat are as shown: this enables a single cell to learn at least 200 different mossy fibre inputs, without confusion between learned and unlearned events.

entirely from small pyramidal cells[4]. In contrast, the pontine nuclei receive collaterals from both large and small pyramidal cells[4]. The distinction may be that superficial pyramidal cells project to the inferior olive, while deep pyramidal cells give off collaterals to the pontine nuclei on their way to the spinal cord[4,6]. Further, the projection from the ventro-lateral thalamic nucleus to the motor cortex is direct to the deep pyramidal cells, and perhaps by way of an excitatory interneurone to both the superficial and the deep pyramidals (see Fig. 1[2,6,8]).

(ii) The inferior olive contains fewer cells than there are cerebellar Purkinje cells[10]. This means that either there are other sources of climbing fibres or the olivo-cerebellar fibres branch. It seems that the latter explanation is correct[5,6]. The distribution of the branches of one climbing fibre also seems to be restricted to a parasagittal plane[6].

(iii) The hypothesis[3] that the parallel fibre–Purkinje cell synapses are facilitated by simultaneous pre-synaptic and climbing fibre activity has implications deeper than merely allowing each Purkinje cell to memorize 200 or so different mossy fibre inputs. If a number of similar mossy fibre inputs have been learned and later an unlearned input is presented which is near enough to those which have been learned, then the Purkinje cell may treat the new input as if it had been learned. This is probably not the disadvantage it was once thought[3]. It means that a Purkinje cell will generalize its response to all events in those regions where learned events are sufficiently clustered together. The implications of this generalization are set out elsewhere[7].

## Consequences of this New Information

*Input–output relations.* In Fig. 2, the new information (i) is combined with the previous knowledge of cerebellar anatomy. All the synapses in the diagram are excitatory, except those from the Purkinje cells to the cells of the cerebellar nuclei. One very striking feature of this circuit diagram is the loop formed from the deep pyramidals through the pontine nuclei, cerebellar nuclei and *VL* nucleus of thalamus back to the deep (and also superficial) pyramidals. This arrangement has been commented on before[1,11]. A necessary assumption of the present theory is that this loop, which will provide a positive feedback from the deep pyramidals to themselves, is so arranged as to give rise to temporally extended pyramidal cell outputs. One possibility would be that the feedback is chiefly to the original area, so that a movement—once initiated—will tend to continue indefinitely (at least well beyond the normal firing period of pyramidal cells in response to an excitatory input): and this will only be terminated either by applying direct inhibition to the deep pyramidal cells or by breaking the feedback loop. In the original cerebellar theory[3], two possible forms of input–output relation were described, both of which required that each individual Purkinje cell could initiate one of the elemental movements into which it was postulated all actions were broken down. For executing actions it was thought necessary only to copy the correct pattern of elemental movements. It was shown how the cerebellar cortex could arrange this by having every elemental movement driven by the context in which it is required.

The anatomy of Fig. 2 is not wholly compatible with this simple programme for copying patterned sequences of elemental movements. In general, if a machine has to execute a sequence of movements, it can operate either by turning on the correct elemental movements at any instant, or by turning off all the incorrect ones. We believe that the design of the cerebellum suggests that the second scheme, the converse of the original input–output relations, is in fact used for learning motor actions. The second scheme is at first sight absurd, because the number of elemental movements required at any instant is far smaller than the number of possible elemental movements. It only becomes more economical than the first scheme if the number required exceeds the number which need to be turned off. In practice, this means that some agency must, at any instant, select from the vocabulary of elemental movements a particular set of "possibles", which includes all those actually required. If this can be done so that the number of "actuals" is greater than the number of "possibles" minus "actuals", it becomes cheaper to operate by deleting unwanted elemental movements from the set of "possibles".

Such an agency would have to satisfy the following properties: (*a*), it must consist of cells capable of driving elemental movements; (*b*), these cells must be capable of being context-driven; (*c*), the set of situations to which each cell responds must include those in which it is needed; and (*d*), cerebellar action upon it must be such that Purkinje cell activity turns off instructions for one (or more) elemental movements. We propose that the set of deep pyramidal cells in the motor cortex is such an agency, and that the conditions (*a*) to (*d*) are satisfied by them.

One can now assign a definite role to the small superficial[8] pyramidal cells which project to the inferior olive.

These, we assume, project to regions of the inferior olive which drive climbing fibres in the same general region of the cerebellum as that which projects back to the deep pyramidal cell beneath them. If the ideas described earlier are correct, these cells must fire when the large pyramidal cells related to them are firing but should not. That is, the small superficial pyramidal cells should detect the need to correct the current motor activity by deleting the messages from their corresponding deep pyramidal cells. In this respect it is of interest that the *VL* nucleus of the thalamus sends excitatory connexions to both deep and superficial pyramidals; this will inform the superficial pyramidals of the feedback excitatory input to the deep pyramidals; clearly there is no point in the superficial pyramidals making deletions when the deep pyramidals are, in fact, not going to be fired. Learning may be necessary for organizing the details of the projection from the *VL* nucleus to the cortex.

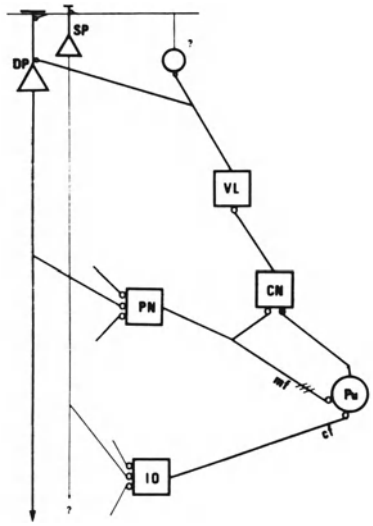



Fig. 2. There are two relevant kinds of cell in the motor cortex: small, superficial pyramids (*SP*) and large, deep pyramids (*DP*). The *DP* cells send collaterals to the pontine nuclei (*PN*), and the *SP* cells to the inferior olive (*IO*). The axons from the inferior olive terminate as climbing fibres (*cf*) on the Purkinje cells (*Pu*) in the cerebellar cortex; those from the pontine nuclei become mossy fibres (*mf*). Purkinje cells are inhibitory, and send synapses to the various cerebellar nuclei (*CN*): these nuclei also receive excitatory synapses from mossy fibre and climbing fibre collaterals. The cerebellar nuclei send excitatory synapses to the ventro-lateral nucleus of the thalamus (*VL*). *VL* projects back to the motor cortex by way of a fast and a slow path: the fast path goes only to the large, deep pyramids; the slow path goes to both deep and superficial pyramidal cells, perhaps by way of an interneurone.

Superficial pyramidal cells therefore recognize the classes of events which are incompatible with the current firing of the corresponding deep pyramidals. The analysis behind the recognition of the need for such corrections may be complicated, because it involves ideas about what the animal is trying to do. Its results can, however, be tied to specific contexts, using the kind of learning of which the cerebellar cortex may be capable[3]. There is thus a clear advantage to be gained by storing the corrections in the cerebellum.

It may fairly be objected that nothing has been said about the way in which the small pyramidal cells detect the need for the corrections which they can implement. This problem is in principle no greater, however, than the analogous assumption concerning the deep pyramidal cells: how do they recognize the need for their elemental movements? On a superficial level it is clear that all pyramidal cells could be capable of learning contexts[7] using the same mechanisms that have been described for the cerebellar Purkinje cells[3]. The deeper aspects of these problems have also begun to yield[7], and a full account of them will appear elsewhere.

The following summary states the conditions under which the inverted input–output relations could work, and hence the experimental findings needed to prove or disprove the hypothesis:

(*A*) Elemental movements are coded by deep pyramidal cells in the motor cortex.

(*B*) The set of situations to which such cells respond includes those in which they are needed.

(*C*) Their axon collaterals to the pontine nuclei provide a positive feedback loop (via the cerebellar nuclei) which is necessary, during normal operation of the system, for the proper initiation and continuation of their elemental movements.

(*D*) Small, superficial pyramidal cells recognize the need for correction of current motor cortex output. These corrections involve the prevention of firing of certain deep pyramidal cells.

(*E*) These corrections, whose initial computation is not necessarily easy, can eventually be run by the cerebellar Purkinje cells—in the same way as Purkinje cells were originally thought to drive the elemental movements themselves[3].

*Branching climbing fibres.* Purkinje cells in different regions of the cerebellar cortex are exposed to information, through the mossy fibres, that originates in different parts of the body and brain. A full description of the state of the body and brain as transmitted through mossy fibres will be called a full context, and a similar description of part of the body or brain a partial context. Then each Purkinje cell has access to a partial context; and the kind of contextual information which may reach each cell is probably fixed.

Each Purkinje cell usually receives exactly one climbing fibre. Hence if the axon from a single olivary cell gives rise to ten climbing fibres, the firing of that olivary cell effectively signals modification conditions to ten, presumably different, partial contexts. During the rehearsal necessary for the cerebellum to learn a given action, some of these partial contexts will recur and some, because they carry information which is essentially irrelevant, will not. Those Purkinje cells, the firing of whose climbing fibres is associated with a relatively unchanging partial context will learn that context—and this will be useful. Those which receive a different partial context each time will not learn (provided synaptic modification does not work first time); nor would it be of any use if they did. Indeed, it would be a disadvantage on two grounds. First, it would reduce the effective capacity of the Purkinje cell to learn useful contexts; second, it might cause incorrect deletions during an action in which an irrelevant partial context arises and the elemental movement is required.

There are many related questions concerning the number of corrections a Purkinje cell discharge can implement, the kind of convergence there is in the cerebellar nuclei, and so on, which cannot be properly studied until more information becomes available. The parasagittal distribution[6,12] of the climbing fibres may, however, shed some light on these problems. It is known that the cerebellar cortex tends to be organized into longitudinal strips, whose Purkinje cells project to restricted regions in the intracerebellar nuclei[2,12]; so the climbing fibres from a single olivary cell will tend to cause modification of Purkinje cells whose influences converge on a restricted zone of the cerebellar nuclei. One can even devise a

plausible embryological model which ensures that the Purkinje cells related to one olivary cell all converge on a single cerebellar nuclear cell—so that there is a one to one correspondence between olivary cells and cerebellar nuclear cells. But such a restriction is by no means necessary for the theory.

*Detection of clusters by Purkinje cells and climbing fibres.* It seems likely that two parts of the theory developed by one of us[7] for the cerebral pyramidal cells also apply to the cerebellar Purkinje cells. The first concerns the nature of the signals which the Purkinje cells actually transmit. It is possible that these cells do give a response which is strictly all-or-none, depending on whether the current input has been learned. We feel, however, that it is more likely that they signal a measure of how similar their current output is to the structure of the events that they have learned. It seems that the most suitable measure of this similarity is the fraction of the currently active afferent synapses to a cell which have been modified[7], provided that fraction is greater than some fixed lower bound p (say). A model has been proposed by which this quantity could be measured by a single cell[7], and we feel that this is likely to be more suitable for the theory of Purkinje cell dendrites than the simple one developed earlier[3].

This raises important questions concerning the need for convergence of Purkinje cell discharge onto cerebellar nuclear cells. Is it possible for a single maximally firing Purkinje cell to turn off a cerebellar nuclear cell completely, or does it need convergence from several Purkinje cells? And if several converging Purkinje cells are firing submaximally—in response to inputs rather dissimilar to their learned partial contexts—then is their summed effect sufficient to turn off the cerebellar nuclear cell?

The second application of the cerebral theory to the cerebellum concerns the discovery that a climbing fibre can organize a kind of cluster analysis[7]. Provided the information arriving at Purkinje cells is clustered and that the climbing fibre is coactive with enough events in a cluster, then the cell will respond to many more events, whether or not they have ever been associated with the climbing fibre activity. We think that this effect, certainly vital in the cerebral cortex[7], is probably important in the cerebellum also. It is a mechanism which can provide a kind of generalization to events which should "obviously" initiate the same responses as their neighbours without the necessity for a specific new learning trial.

The next topic we wish to raise concerns the Purkinje axon collaterals[1]. It has been pointed out[3] that the effect produced by them through their connexions with basket and stellate cells is simple, whereas their effect through the Golgi cells is not. One possible explanation of their existence is that, when active in the region of a particular Purkinje cell **P**, they cause **P** to relax the scale on which it measures the similarity of the current input to the events it has learned. This is suggested by two facts: first, the inhibition reaching **P** will be decreased by collateral stimulation; and second, the Purkinje axon collateral inhibition of the Golgi cells will cause a slight decrease in the local granule cell threshold. This is the correct step for interpreting the current mossy fibre input within the structure formed by the other mossy fibre inputs which it has learned (by the interpretation theorem[7]).

It is therefore possible not only that direct generalization, of the sort described above, can occur in the cerebellar cortex, but also that the extent to which this generalization is permitted (that is, lowering the value of p) can be varied by Purkinje axon collateral activity. If this is so, it has implications about the distribution of these collaterals that one would expect to find: because the cues to lower p for a particular cell **P** must arise from information suggesting that it would be appropriate to do so. This means that the Purkinje axon collaterals ending in one region of cortex should fire only when it is likely that the corrections controlled from these are wanted; and in general, the more likely they are to be wanted, the greater will be the permissible degree of generalization there (that is, the lower p can be), and so the more activity there should be in the Purkinje axon collaterals terminating there. This implies that the collaterals from each Purkinje cell $\mathbf{P}_1$ tend to be distributed to regions of cortex containing Purkinje cells which are needed after or at the same time as $\mathbf{P}_1$. The most obvious of such regions would be those containing the Purkinje cells which are fired by the other branches of the olivo–cerebellar axon which sends a branch to $\mathbf{P}_1$. (It is interesting to note that Purkinje axon collaterals are often closely related to climbing fibres.) Those regions of cortex receiving collaterals from many currently active Purkinje cells would then be more likely to be needed next than those regions receiving from only a few. The known distribution of Purkinje axon collaterals tends to support this notion. The Purkinje axons first contribute collaterals to the transversely running infraganglionic plexus, whose fibres often bridge across several folia; branches are given off from this plexus to the longitudinally running supraganglionic plexus, whose distribution is much more limited. Hence Purkinje axon collateral effects will tend to be restricted to the parasagittal plane. We have already shown that there is reason to suppose that the Purkinje cells have closer relations to other Purkinje cells within such a plane than without.

There is one other piece of evidence in favour of this rather complex view of the Purkinje axon collaterals. It is that it also accounts for the climbing fibre collateral effects[1,3]. For, during learning, any instruction to generalize must be annulled, in order that a true record of the mossy fibre input may be stored. According to the theory[3], learning occurs at **P** when the relevant climbing fibre is also active; and when it is, the effect of its collaterals could roughly balance the effect of the Purkinje axon collateral near **P**. According to the available evidence[1], both types of collateral are weak and their effects are opposite.

## Timing Relationships

We have argued that the small, superficial pyramidal cells of the cerebral cortex detect incompatibilities in the current deep pyramidal cell activity, and that they modify the behaviour of the cerebro–cerebellar–cerebral loop to cope with this. We now consider the timing relationships involved.

The speed of the main "feedback" loop is astounding. It incorporates some of the fastest pathways in the nervous system, and its major links all include monosynaptic connexions[1,11]. In the cat, discharges in the pontine nuclei follow stimulation of the cerebral white matter by as little as 2 ms[4]. The corresponding times for the other stages are: pontine nuclei to cerebellar nuclei, 1 ms[11]; cerebellar nuclei to *VL* nucleus of thalamus, 2 ms[1]; *VL* nucleus of thalamus to cerebral pyramidal cells, an estimated 1 ms. The whole loop may therefore be traversed in as little as 6 ms, and certainly within 10 ms. Such a fast mechanism is clearly required in voluntary movements, especially those of a more complex kind when muscular groups have to be set into action in rapid sequence and at closely defined times.

Contextual information reaching the cerebellar cortex through the mossy fibres is also rapidly transmitted; indeed, it involves almost the same pathways. The time taken for stimulation of the subcortical white matter to evoke a mossy fibre response is 2·7 ms[4]. Mossy fibre responses to stimulation of forelimb and hindlimb peripheral nerves have delays as short as 5 ms and 7 ms respectively[4].

On the other hand, the cortico–olivary–climbing fibre pathway is quite slow. The climbing fibre discharge evoked by stimulation of the cerebral subcortical white matter has a delay of 15 ms[4]. At first sight, it would therefore seem impossible that the superficial pyramidals

could signal that the currently active deep pyramidals should be deleted: their commands would arrive too late to be effective.

It is, however, necessary to consider the time scale of the context in which these instructions are being made. The overall context of the movement changes much more slowly than the individual components of that movement. That a given group of deep pyramidal cells should not fire is not merely a decision whose effects last for a few milliseconds: the group will be required to be off for an extended period of time. The decision may have to be made and implemented quickly, but it will remain in force for much longer. This means that the modification conditions refer to extended contexts, of perhaps as long as 100 ms, rather than to instantaneous contexts.

It is therefore proposed that the inferior olive cells should fire in prolonged bursts, of up to 100 ms. During this time, the currently active synapses to the related Purkinje cells should be strengthened in proportion to their degree of activity. This allows the Golgi cell threshold system to be reset by the climbing fibre collaterals, so as to give the "correct" parallel fibre pattern during modification. More important, this ensures that the Purkinje cells can respond in good time to inhibit the cerebellar nuclei cells—because the mossy fibre context just before the climbing fibre activity (that is, when the input reaches the pontine nuclei) will differ only slightly from that during it. The ability of Purkinje cells to generalize will also help in this effect.

It may be found that the small, superficial pyramidal cells anticipate the large, deep pyramidal cells, and signal in advance that certain cerebellar nuclei cells must be inhibited within the context of the present developing movement.

## Cerebellar Disorders

The present theory can provide a tentative explanation for many of the disorders arising from damage to the cerebellum. One of the most striking effects of acute cerebellar lesions is the delay in the initiation and termination of movements[13]. The delay in initiation is probably caused by malfunction of the cerebellar nuclei. In the acute stage of such lesions, there is considerable oedema and consequently raised pressure in the cerebellum; this could account for such malfunction. The result is that, when the cerebral cortex tries to initiate the movement, there is little or no excitatory feedback to the motor cortex. The movement can only be got going by a considerably greater voluntary effort, and this involves both delay and slow pick-up. With recovery of functioning of the cerebellar nuclei (that is, in those lesions which are more superficial), such delays will tend to disappear[12].

Delay in termination[13] probably results from a combination of two factors. First, there is an inability to initiate the muscular contractions which are required to stop the movement: this again involves the cerebellar nuclei. Second, there is delay in switching off the current movement: this results from the malfunction of the cerebellar cortex. This latter effect should become more apparent as recovery proceeds, for the cerebellar nuclei will be functioning normally while the damage to the cerebellar cortex persists. In other words, the context which signifies that the movement should stop is no longer able to implement this operation, because the relevant Purkinje cells are lacking. This argument receives support from the observations of Gordon Holmes[13] that the start of relaxation in a movement is usually more markedly affected than the start of contraction.

The inability of patients with unilateral cerebellar lesions to maintain voluntary postures on the affected side, and the greater sense of effort involved in making any voluntary movements, are both common features in the early acute stages. Both are consequences of inadequate excitatory feedback from the cerebellar nuclei.

The phenomena of dysmetria[13], in cases of acute cerebellar lesions, and of hypermetria, which occurs in more persistent cases, are probably related to these disorders. Dysmetria will result from the malfunctioning of both cerebellar nuclei and cerebellar cortex. Movements, once initiated, are ill-gauged and tend to undershoot or overshoot the mark. Undershoot will be caused by an inability to maintain a voluntary movement (a symptom of cerebellar nuclei malfunction); overshoot will be caused by inability to stop voluntary movements (already considered). It is particularly interesting that hypermetria should ensue—this is exactly what the theory would predict. It results from the lack of inhibitory control from the cerebellar cortex; as a result the movements consistently overshoot and are excessively forceful.

The decomposition of complex movements[13] is a natural consequence of any cerebellar malfunction. The errors arising in the initiation, continuation and termination of successive and concurrently running elemental movements should lead to hopeless confusion. The only hope for success would be to deal with one elemental movement at a time, so that errors may be consciously and deliberately dealt with as they arise.

An interesting disability which arises in cerebellar patients is that on trying to flex just one finger (in order to bring it into apposition with the thumb), they frequently flex all four fingers at the same time[13]. In this case, it may be that normally the cerebral command is to flex all four fingers but suppress flexion on the unwanted three. Certainly in early hand movements, flexion of all four fingers appears before flexion of individual fingers—though there is a cortical representation for each individual finger flexion. The suppression of the unwanted flexions is learned by the cerebellum during the early development of the child. Damage to the cerebellar cortex will interfere with the suppression, and a command to move one finger will initiate movement in all four.

We shall make just one reference to observations made on animals with lesions placed in the cerebellum. This concerns the effects of such lesions on the placing reaction[14]. Lesions which involve the dentate nucleus are found to abolish the placing reaction. In contrast, lesions confined to the cerebellar cortex may actually enhance it. Ablation of parts of the cerebral cortex which include the motor area is known to abolish the placing reaction. This is compatible with a learned reflex which passes through the cerebral motor cortex and whose output depends on positive feedback through the cerebro–cerebellar–cerebral loop. Clearly such a reflex is of use to the animal in standing and walking. Inhibitory control of this reflex is then exerted by the cerebellar cortex.

The functions of the ascending spino–cerebellar and spino–olivo–cerebellar tracts, and their utilization in the control of movements and postures, will be dealt with elsewhere.

We thank the Medical Research Council and Trinity College, Cambridge, for supporting this work.

Received February 9; revised June 29, 1970.

[1] Eccles, J. D., Ito, M., and Szentagothai, J., *The Cerebellum as a Neuronal Machine* (Springer-Verlag, Berlin, 1967).
[2] Jansen, J., and Brodal, A., *Aspects of Cerebellar Anatomy* (Johan Grundt Tanum Forlag, Oslo, 1954).
[3] Marr, David, *J. Physiol.*, **202**, 437 (1969).
[4] Kitai, S. T., Oshima, T., Provini, L., and Tsukahara, N., *Brain Res.*, **15**, 267 (1969).
[5] Armstrong, D. M., Harvey, R. J., and Schild, R. F., *J. Physiol.*, **202**, 106P (1969).
[6] Faber, D. S., and Murphy, J. T., *Brain Res.*, **15**, 267 (1969).
[7] Marr, David, *Proc. Roy. Soc.*, B (in the press).
[8] Towe, A. L., Patton, H. D., and Kennedy, T. T., *Exp. Neurol.*, **8**, 220 (1963).
[9] Branch, C. L., and Martin, A. R., *J. Neurophysiol.*, **21**, 380 (1958).
[10] Escobar, A., Sampedro, E. D., and Dow, R. S., *J. Comp. Neurol.*, **132**, 397 (1968).
[11] Tsukahara, N., Korn, H., and Stone, J., *Brain Res.*, **10**, 448.
[12] Voogd, J., *Prog. Brain Res.*, **25**, 94 (1967).
[13] Holmes, Gordon, *Brain*, **40**, 461 (1917).
[14] Dow, R. S., and Moruzzi, G., *The Physiology and Pathology of the Cerebellum* (University of Minnesota Press, Minneapolis, 1958).

# Jack D. Cowan

## Commentary on

## *How The Cerebellum May Be Used*

Many of the ideas concerning the neocortex (see Cowan's commentary, Chapter 4) are incorporated into a revised theory of cerebellar action. Figure 1 shows the overall architecture of cerebellar interactions:

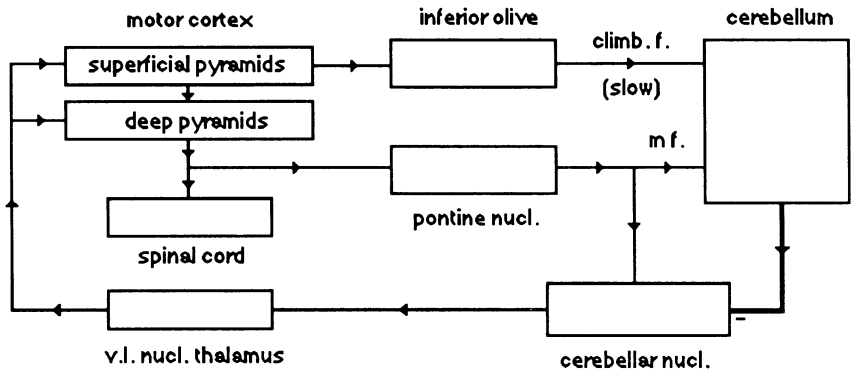



Fig. 1. Architecture of cerebro-cerebellar pathways. All interactions except the cerebellum-cerebellar nucleus are excitatory. All pathways except the inf. olive-cerebellum are fast, with conduction delays of no more than 2 msec.

Marr and Blomfield make the point that the execution of elemental movements is learned by turning *off* all the correct ones. By analogy with Marr's neocortex theory the deep pyramids of the motor cortex are presumed to act as classifiers capable of detecting the "context" of a sequence of elemental movements, and the smaller superficial pyramids are presumed to measure the need to correct current motor activity and to delete the messages from the corresponding deep pyramids. Their axon collaterals to the pontine nucleus provide a fast positive feedback loop via the cerebellar nucleus, which is required for the initiation and continuation of elemental movements. The corrections they initiate are presumed to be learned eventually and run by cerebellar Purkinje

cells. The point is made that Purkinje cells should function much like neocortical pyramidal cells and detect and generalize over clusters of elemental movements, using various collateral interactions not covered in Marr's original paper on the cerebellar cortex.

*Professor*
*Department of Mathematics*
*University of Chicago*
*Chicago, Illinois*

# SIMPLE MEMORY: A THEORY FOR ARCHICORTEX

By D. MARR
*Trinity College, Cambridge*

(*Communicated by G. S. Brindley, F.R.S.—Received* 27 *July* 1970—*Revised* 12 *November* 1970)

## CONTENTS



Vol. 262. B. 841. 3 [Published 1 July 1971

It is proposed that the most important characteristic of archicortex is its ability to perform a simple kind of memorizing task. It is shown that rather general numerical constraints roughly determine the dimensions of memorizing models for the mammalian brain, and from these is derived a general model for archicortex.

The addition of further constraints leads to the notion of a simple representation, which is a way of translating a great deal of information into the firing of about 200 out of a population of $10^5$ cells. It is shown that if about $10^5$ simple representations are stored in such a population of cells, very little information about a single learnt event is necessary to provoke its recall. A detailed numerical examination is made of a particular example of this kind of memory, and various general conclusions are drawn from the analysis.

The insight gained from these models is used to derive theories for various archicortical areas. A functional interpretation is given of the cells and synapses of the area entorhinalis, the presubiculum, the prosubiculum, the cornu ammonis and the fascia dentata. Many predictions are made, a substantial number of which must be true if the theory is correct. A general functional classification of typical archicortical cells is proposed.

## 0. Introduction

The cortex of the mammalian cerebrum admits a crude division into two classes: the archicortex, which is relatively simple and primitive; and the neocortex, which has developed more recently and is very elaborate, especially in man. In a recent paper (Marr 1970), a general theory for neocortex was set out. The present paper provides its counterpart for archicortex.

The comparatively simple structure of archicortex is probably reflected in its performance of a comparatively simple function. The central point of the neocortical theory was that a particular method of organizing information is likely to be useful in many different circumstances: it was shown how neocortex might take advantage of this to change the language in which incoming information is expressed by reclassifying it, as well as carrying out routine storage of associations between existing classes. It will be argued in the present paper that archicortex cannot reclassify information in this way. It will be shown that its histology is consistent with the proposition that it performs only a simple memorizing function—storing information in the language in which it is presented—rather than with organizing information in any more complicated sense. Recent work on the storage of information in nerve nets (Brindley 1969; Marr 1969, 1970) has reduced the construction of such a theory to little more than a technical exercise: it is an unavoidable one none the less, and various interesting factors emerge from this study.

The paper consists of three main divisions. In the first, §§ 1 and 2, the main ideas behind simple memory theory are explained. These ideas lead to a particular neural model which, it is proposed, captures the essence of much of the archipallial cortex. It is shown that under certain circumstances, the performance of such a model can be greatly improved by use of collateral synapses between its cells (the *collateral effect*, § 2.4).

The second part of the paper, § 3, takes an explicit model constructed along the lines suggested by the first part, and derives the equations which describe its expected performance. The model's storage capacity and recall abilities for a selection of values of the important parameters are displayed in a number of tables. The computations (§ 3.1) are followed in the rest of § 3 by a rough justification of the values of the parameters chosen.

The third part of the paper (§ 4) uses the model of § 3.1 to arrive at a theory of the hippocampal cortex. This theory produces many testable predictions, which are summarized in § 5. The theory is restricted to operations within the cortex, and does not describe any input–output relations. The reason is that they are much more complex than, for example, those of the cerebellar cortex, and their inclusion in this paper would have made it prohibitively long. They will therefore be set out elsewhere, together with the necessary extra theory.

### 0.1. *Notation*

Many of the terms and symbols of Marr (1970) are used in this paper, and it is convenient to repeat their definitions here. A *fibre* (e.g. $a_i(t)$) is a function of discrete time $t\,(=0, 1, 2, \ldots)$ and has the value 0 or 1. An *event* on the set $A = \{a_1, \ldots, a_N\}$ of fibres assigns to each fibre a value 0 or 1. Letters like $E, F$ are used for events, and the value that $E$ assigns to the fibre $a_i$ is written $E(a_i)$. The phrase '$a_i$ in $E$' means '$a_i$ takes the value 1 in the event $E$'. A *subevent* on the set $A = \{a_1, \ldots, a_N\}$ of fibres is an event on a subset of $A$. Letters like $X, Y$ denote subevents; and the set of fibres to which $X$ assigns a value is called the *support* of $X$, and is written $S(X)$. Gothic letters like $\mathfrak{E}, \mathfrak{F}$, denote collections of events; and letters like $\mathfrak{X}, \mathfrak{Y}$ denote collections of subevents.

The event $E$ is said to be a *completion* of the subevent $X$, written $E \vdash X$, if $E$ and $X$ agree at all the fibres to which $X$ assigns a value.

Let $\mathfrak{E}$ be the space of all events over $\{a_1, \ldots, a_N\}$. An *r-codon* $c$ on $\mathfrak{E}$ is a function, taking the values 0 or 1, such that $c(E) = 1$ if and only if a particular subset of $r$ fibres $(a_{i_1}, \ldots, a_{i_r})$ all have the value 1 in $E$; $c$ may be regarded as a detector of the subset $(a_{i_1}, \ldots, a_{i_r})$. An *$(R, \theta)$-codon* is a similar function $c$ such that $c(E) = 1$ if and only if at least $\theta$ of a particular collection $(a_{i_1}, \ldots, a_{i_r})$ of fibres have the value 1 in $E$.

## 1. General constraints

### 1.0. *Introduction*

It has recently been argued that neocortex may be regarded as a structure which classifies the information presented to it (Marr 1970). The detectors of the classes it forms are the pyramidal cells of layers V, III and possibly also of layer II. An incoming signal will probably pass through many such classifications during the course of its analysis. The number through which it passes will depend upon the animal, and upon its interest in that kind of information at that moment: it is clear that information is often abandoned as uninteresting before it has been examined to the maximum depth of which the animal is capable.

It is probably reasonable to suppose that at a given moment, there will exist in an animal's brain information whose expression is now as sophisticated as the animal either requires, or can provide. Further classification of the information may be carried out later but, at that moment, the animal needs simply to be able to store it in its present form. Such an expression of the input is called the animal's *current internal description* of the environment, and it is the storage of the current internal description which constitutes the animal's memory of the information. From these memories, he will form new classificatory units, organize temporally extended actions, and arrange to respond in the appropriate way to pieces of subsequent current internal descriptions.

The problems that are studied in this paper are those which arise in the storage and the free association of such current internal descriptions. The central problem may, by the neocortical theory (Marr 1970), be translated into the following form. $\mathscr{P}$ is a large population of neocortical pyramidal cells, of which some are firing. It is required that this should be recorded in some way, so that firing in a few of the cells which are active together in some event $E$ can later elicit the firing of all cells active in $E$. This scheme is probably only remotely analogous to hippocampal input-output relations in most mammalian brains, but it is a convenient model with which to introduce the cortical theory.

Three considerations necessitate the construction of a special theory for this problem. First, although it has been shown that the neocortex can store associations between classificatory units

(Marr 1970, §4)—for example through the pyramidal cells' basilar dendrites—this kind of storage requires a rather special kind of pre-existing structure: the relevant fibres have already to be distributed to roughly the correct places. Direct storage of associations in this way makes heavy demands on the abundance of interconnexions.

The second consideration concerns the way this kind of associational storage works. It essentially involves recording at each active pyramidal cell $\Omega_i$ in $\mathscr{P}$ a list of many of the cells $\Omega_j$ co-active with $\Omega_i$. This can become very expensive, and there are ways of improving upon it. Furthermore, it is only worth recording information in a permanent memory when it is known fairly certainly how that information should be expressed. It may, for example, turn out that part of a current internal description should be recoded to form a new classificatory unit. If this were done, a direct associational storage of that current internal description would soon be obsolete: it is better to store it temporarily in a special associative memory, until it becomes quite clear how it should be permanently set down.

Thirdly, there are many instances in which the control of behaviour would be made rather easy if an associative memory were available as a temporary storage place for instructions. This facility would, for example, allow an instruction of the form 'see post-box–post letter' to be set up before one started out on a walk.

### 1.1. *Simple memory*

Let $\mathfrak{E}$ be the set of all events and all subevents on the fibres $\{e_1, e_2, \ldots, e_m\}$, and let $\mathfrak{F}$ be the set of all events on the fibres $\{f_1, f_2, \ldots, f_n\}$ (see §0.1 for definitions of these terms). As time $t$ progresses ($t = 0, 1, 2, \ldots$), denote the event at time $t$ in $\mathfrak{E}$ by $E_t$, and that at time $t$ in $\mathfrak{F}$ by $F_t$. A *simple memory* is a device which connects $E_t$ and $F_t$, for each $t$, in the following sense. Let $X$ be a subevent or an event in $\mathfrak{E}$. Let $X_1, \ldots, X_J$ be all the completions of $X$ in $\mathfrak{E}$; that is $X_i \vdash X$ for $1 \leqslant i \leqslant J$, and there are no others. (If $X$ is an event, its completion is unique and is itself.) Suppose that exactly one of the events $X_1, X_2, \ldots, X_J$ has occurred. That is, the equation $X_j = E_t$ has exactly one solution, for all values of $j$, and of $t$ up to the present time. Then $\mathfrak{E}$ and $\mathfrak{F}$ are joined by a *simple memory* if presentation of $X$ subsequently causes the event $F_t$ in $\mathfrak{F}$.

Two special cases deserve separate names. In the case where $\{e_1, \ldots, e_m\} = \{f_1, \ldots, f_n\}$, the memory described above is called a *free simple memory*: if the memory is not free, it is called a *directed simple memory*. The reason for these names is that in a free simple memory, there are no constraints upon the way the associations may flow. Any collection of fibres from the set $\{f_1, \ldots, f_n\}$ may be used to recall the activity of the rest of these fibres at a particular time. In directed simple memory, this is not so. For example, $f_1$ may not be included in $\{e_1, \ldots, e_m\}$, in which case information about $f_1$ can never be used to recover information about the rest of the $f_i (2 \leqslant i \leqslant n)$.

In the models that are studied in this paper, rather little is said about whether

$$\{e_1, \ldots, e_m\} = \{f_1, \ldots, f_n\}.$$

The question is unimportant until the problem of input–output relations is studied. It is enough to note here that the same basic memory mechanism can be used for both free and directed simple memories.

### 1.2. *Numerical constraints*

There are various arguments which roughly determine the shape of simple memory theory; they are best presented in the form of order-of-magnitude calculations. This section contains four such arguments: the first is concerned with the proportion of learned to possible input events; the second with the likely size of input vocabulary—i.e. the number of input fibres; the third with the number of events which have to be held in the memory; and the fourth with the proportion of cells of the population concerned with the storage that is used for each event.

#### 1.2.1. *The constraint of a limited history*

The number of fibres that may be involved in a current internal description must be expected to be quite huge; but even if it were only 1000, and a mere 10 were involved at each unit of time (say 1 ms), there are enough possible events for the system to run for more than $10^{12}$ years without repetition. The world is, of course, not random; but the figures 10 and 1000 are certainly underestimates. From this observation follow two conclusions. First, information about the current internal description concerns *whether* a particular event has occurred, rather than *how often* it has done so, since the answer to the latter question is almost certainly never or once. Secondly, very few of the possible events will ever actually occur. Recovery of an event will therefore be theoretically possible from an extremely small amount of information, and the design of neural models must be such as to allow this.

#### 1.2.2. *Cortical indicator cells*

It is supposed that neocortical pyramidal cells of layers III and V are output cells for classificatory units, and that some, though not necessarily all, of such cells can take part in a current internal description. The human cerebrum contains about $7 \times 10^9$ cells (Shariff 1953) of which at least say $10^8$ could be classed as cortical pyramids. This is a huge number, and any attempt to allow all the cells in a population of this size to have access to a simple memory would lead to an unacceptably large neural structure for that memory. If, however, the memory is used for a relatively small number of events (of the order of $10^5$), information then being removed to the neocortex, an important simplification can be made.

Suppose that scattered more or less uniformly over the cerebral neocortex were cells which responded simply to activity in their neighbourhood of the cortex. If such a cell were driven by a very small region of the cortex—an area of perhaps 0.03 $mm^2$—it would serve as a marker of activity in the cortical pyramids within that region. Each cortical pyramid represents a separate classificatory unit, and it can probably be assumed that within such a region not all the pyramids will be active simultaneously. The non-specific cell which marks activity in that region is called an *indicator cell*: the best design for such a cell would probably assign to it a thin, unbranched ascending dendritic stem which passes through all layers of the neocortex, and which is sensitive to excitatory influences throughout its length.

The great advantage of indicator cells is that they can be used as entry fibres to a simple memory, provided that the return fibres synapse with the true cortical pyramids and not with the indicators. In this way, whenever a pyramidal cell is used, its nearby indicator(s) cause an entry to be made to the memory, while the return synapses to the pyramid itself are modified. The memory can later use these synapses to drive the original pyramidal cell. The only disadvantage arises when two nearby pyramidal cells are used in two different but very similar situations, but this problem is not a severe one.

A density of 30 indicator cells/mm$^2$ allows a quite sensitive specification of location; and although this figure is only a guess, we shall see in § 3.1 that it can be changed by a factor of 10 without much disruption of the models analysed there. In general, the density of such cells should reflect the frequency with which the various regions of neocortex use the simple memory facility, the density being high in regions expressing information which often needs temporary storage, and low elsewhere. If indicator cells are used, one would expect their dendritic design to vary as well, being very compact in areas where their cell density is high, and perhaps arborizing where they are rare.

The total area of one hemisphere of the human cerebral cortex is estimated to lie between 800 and 1300 cm$^2$. If it is supposed that about 400 cm$^2$ need to have access to the simple memory (this figure may be too large), the memory will possess about $10^6$ afferent fibres. This is the approximate number of fibres needing free simple memory, and does not include the various kinds of directed simple memory which may, for example, be involved in the planning of temporally extended actions.

### 1.2.3. *Capacity requirements*

The design of a memory requires some idea of the number of events to be stored, and of the amount of information from which recovery of a whole event should be possible. These two factors are linked, since if a memory has to be capable of recovering events from a very small amount of information, its capacity is much smaller than if most of the original event can be used to initiate recall. It is necessary to make a rough estimate of both requirements.

Simple memory has many uses, and the brain probably employs different structures for each use, though the structures are likely to conform to the same basic plan. For directed simple memory, it is very difficult to provide even a rough guess at the storage requirements. For free simple memory (an explicit model for which is developed in § 3.1), some idea of the necessary capacity can be obtained. The figure will not be very high, since it is part of the general theory that information is moved out of the simple memory when it is known how best to do this. The two possibilities for the re-storing of the information currently in simple memory are (i) that it is moved to neocortex in the form of new classificatory units (see Marr 1970, §§ 4, 5); (ii) that it is moved to neocortex in the form of associations between existing classificatory units (through, for example, the basilar dendrites of neocortical pyramidal cells).

It has been suggested that at least a part of the transfer between simple memory and the neocortex takes place during sleep (Marr 1970, § 5). This implies that simple memory must have adequate capacity for holding the events of at least one day. There are 86 400 s in 24 h, and although many events will not be moved out for some time, one probably does not store a new event every second. The figure of $10^5$ is therefore taken as the kind of capacity required of the free part of the simple memory.

The amount of information which can recall an event is even harder to estimate, but it should probably be very small, less than a tenth of the information contained in the original event. The model of § 3.1 operates at a level considerably below this figure.

### 1.2.4. *The activity of a collection of cells*

Let $\mathscr{P}$ be a population of $M$ cells, $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_M$. Suppose that at time $t$, exactly $L$ of these cells are firing: then the *activity* of $\mathscr{P}$ at time $t$ is defined to be $L/M$, and is written $\alpha = \alpha(t)$.

If $\mathscr{P}$ is being used to store $n$ input events, and if its activity during each is $\alpha$, then each cell of

$\mathscr{P}$ may expect to be used in $\alpha n$ input events. If the storage is taking place in the cells of $\mathscr{P}$, each cell will have to learn part of about $\alpha n$ input events. The number of subevents a single cell can learn is determined by the number of modifiable afferent synapses it has, and by the number that are used in each subevent. For example, the number of fairly dissimilar events that a cerebellar Purkinje cell can learn is probably about 200 (Marr 1969). Purkinje cells have more afferent synapses than any cortical cells, and so it follows that most cortical cells will not be able to learn substantially more than 200 subevents. The number of input events that the population $\mathscr{P}$ described above may learn is therefore bounded by about $200\alpha^{-1}$. This is an important and rather general constraint.

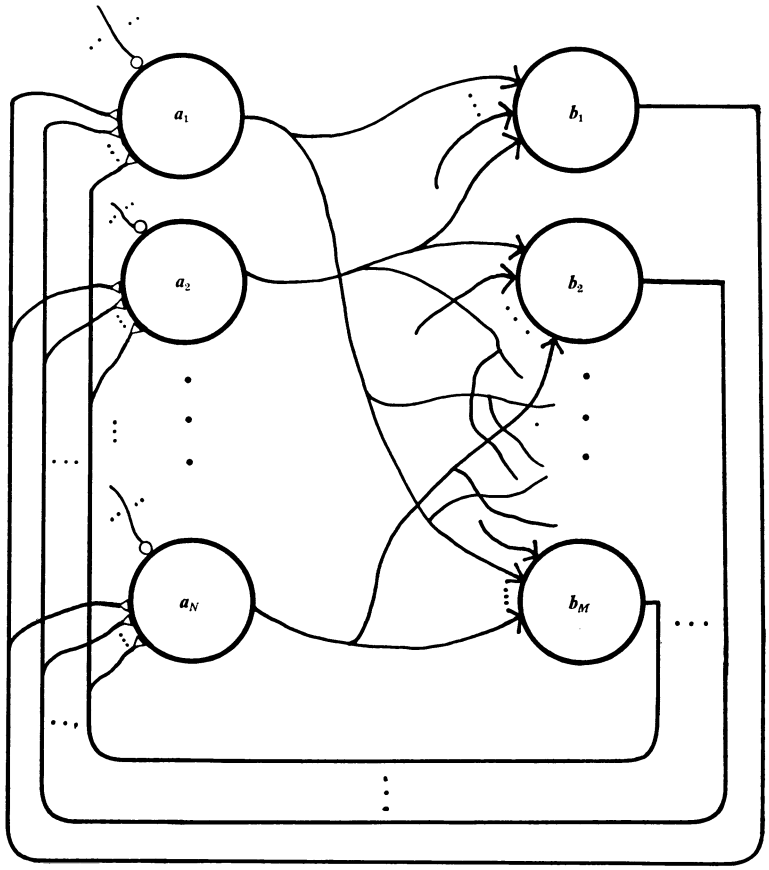



FIGURE 1. A primitive associative memory. The current internal description is an event on the cells $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N$: this is given a codon representation in the cells $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_M$ (which have Brindley afferent synapses), and the return to the $\boldsymbol{a}_i$-cells is through Hebb modifiable synapses. The various inhibitory interneurons necessary for the correct operation of the system have been omitted. This class of model provides an efficient associative memory for events on the $\boldsymbol{a}_i$ as long as their number and size are suitably restricted.

### 1.3. *The form of the analysis*

The model of figure 1 shows almost the simplest design for a free simple memory for events on the set of fibres $A = \{a_1, \ldots, a_N\}$. This model may be derived most quickly as follows. Let $X$ be a subevent on $A$. Then the problem of recovering the completion $E$ of $X$ (assuming that exactly one such $E$ has occurred) may be regarded as the problem of diagnosing those $a_i$ with $E(a_i) = 1$ from

the information contained in the subevent $X$ on the basis of the information stored in the memory. It is now possible to apply the interpretation theorem (Marr 1970, §§ 2, 4) to the problem, and figure 1 contains one arrangement for applying the corresponding neural analysis.

The inputs $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N$ are the cells which constitute the vocabulary of the current internal description, and the cells $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_M$ are suitable evidence cells. The technique of codon formation is used to construct suitable evidence cells (see Marr 1970, § 4), and for this reason, the $\boldsymbol{b}_j$ afferents end in Brindley synapses. (*Hebb* synapses will be taken to mean synapses that are initially ineffective, but are facilitated by simultaneous pre- and post-synaptic activity. *Brindley* synapses are Hebb synapses that also contain an unmodifiable excitatory component (Marr 1970, § 4.3.1; Brindley 1969).) The $\boldsymbol{b}_j$-cell population contains appropriate threshold-setting inhibitory interneurons, whose function is to keep the number of $\boldsymbol{b}$-cells that are active roughly constant during both storage and recall. These interneurons do not appear in the figure.

The return projection to the $\boldsymbol{a}$-cells ends in Hebb synapses. There are inhibitory interneurons in the $\boldsymbol{a}$-cell population which, during recall, allow firing in only those $\boldsymbol{a}$-cells the highest proportion of whose active afferent synapses from the $\boldsymbol{b}$-cells have been modified. This corresponds to implementing the interpretation theorem at the $\boldsymbol{a}$-cells, in response to the subevent $X$. The cell $\boldsymbol{a}_i$ measures $\mathcal{P}(a_i|X)$ when $X$ is applied to the set $\{a_1, \ldots, a_N\}$ (Marr 1970, § 2.5), and the $\boldsymbol{b}$-cell thresholds are lowered in such a way as to keep the number of $\boldsymbol{b}$-cells that are firing roughly constant (Marr 1970, § 4.4).

In principle, free simple memory is obtained by allowing the projections from the $\boldsymbol{a}$-cells to the $\boldsymbol{b}$-cells and back to be distributed freely over both populations (as in figure 1). A directed simple memory is obtained, for example, by arranging that only certain $\boldsymbol{a}$-cells project to the $\boldsymbol{b}$-cells, and that only certain $\boldsymbol{a}$-cells receive projections from the $\boldsymbol{b}$-cells.

### 1.4. *The consequences of the numerical constraints*

In this section are outlined the principal effects of the constraints described in § 1.2 when they are applied to the kind of model to which the methods of § 1.3 give rise. The development is informal, and is designed to give the reader an overall view of the theory developed in §§ 2 to 4. Its main purpose is to show roughly why it is that the basic model of figure 1 is inadequate for simple memory, and how this leads to the idea that a special working representation of each input has in fact to be created in the memory. This central representation is a kind of template for each event; it probably involves rather few cells—perhaps only 100 to 1000 even in man—and provides an economical central storage pattern from which the event in the output space $\mathfrak{F}$ at that particular instant can be recovered. This representation, called the *simple representation* of the current internal description, is a central feature of the present paper.

#### 1.4.1. *Synaptic modification*

Where codon formation occurs, the relevant synaptic modification has been regarded as an all-or-none process (Marr 1970, § 4). In contrast, the afferent synapses to output (cortical pyramidal) cells need to have variable strength in order to measure $P(\Omega|c_i)$, although it may be that this is in practice approximated by an all-or-none process (Marr 1970, §§ 4, 7). The numerical constraints of § 1.2 imply that in the theory of archicortex, synaptic modification should probably be regarded as an all-or-none process, although it is allowed that different classes of synapses may have different maximum strengths.

One reason for this is as follows. For evidence cells (i.e. in codon formation) the arguments are the same as for neocortex: these synapses are involved in representing a diagnostic space, not in measuring probabilities therein. For diagnostic processes in a simple memory, the argument rests on the peculiar way in which the memory is used—as a temporary store to which new information is continually being added. At a neocortical output cell, the notion of a conditional probability has a practical meaning, since the output cell and its supporting evidence cells are structures which form a permanent part of the brain's interpretive apparatus. This is not true of simple memory. Much of the information held therein is needed only temporarily, and that which is not will be removed to the neocortical store when it becomes clear how it should be represented there. The notion of conditional probability in such circumstances has at best only a changing meaning.

1.4.2. *Inadequacy of the simple model*

It is easy to show by using order-of-magnitude calculations that the simple model of figure 1 cannot be applied to the case where there are as many as $10^6$ input cells $\boldsymbol{a}_i$. Since neocortical pyramidal cells probably possess fewer than 100000 afferent synapses, most of which will be occupied with standard diagnostic evidence and with permanent neocortical associative information, it can probably be assumed that only about $10^4$ synapses are available for the simple memory function. In the simple model outlined in figure 1, this means that the number of $\boldsymbol{b}$-cells, $M$, may be taken as $10^4$, each one synapsing with every one of the $10^6$ $\boldsymbol{a}$-cells. The $\boldsymbol{b}$-cells must possess modifiable synapses since, otherwise, recall from subevents of learnt events would be impossibly bad. If the capacity of the memory is taken to be about $10^5$ events, and each $\boldsymbol{b}$-cell can learn $10^2$ (§ 1.2), the activity $\alpha$ of the $\boldsymbol{b}$-cell population must be as low as $10^{-3}$—that is, 10 cells active at any instant. This number is too small to allow a reliable representation of the whole input event by the $\boldsymbol{b}$-cells, and the model is therefore inadequate.

1.4.3. *The simple representation of the current internal description*

Arguments like that outlined in § 1.4.2 show two things: first, that there must be more than one layer of cells (like the $\boldsymbol{b}$-cells) between the input and the return of a simple memory, if it is bound by numerical constraints like those described in § 1.2. Secondly, the small number of synapses available at neocortical pyramids for the simple memory means in effect that there will be rather little spare capacity in the projection back from the simple memory. That is, most of the storage capacity at these synapses will be exhausted by the straightforward task of relating the pyramids to the activity in the projection from the memory during full events: there will be little left over to help in the task of completing a subevent of a learnt event. This means that during recall of a learnt event from a subevent, the recall must have been virtually achieved by the memory *before* the signals reach the projection back to the neocortex. Hence most of the diagnostic analysis involved in discovering the completion of a subevent takes place in the memory itself, not at the $\boldsymbol{a}$-cells. In the simple case of figure 1 (which can be used to store rather few events), this would mean that a subevent $X$ of $E$ could recall $E$ only if it caused activity in the same $\boldsymbol{b}$-cells as did $E$.

This is a rather stringent condition on the structure of the memory. It means that there exists a stage—a layer of cells—in (and by) which the completion process is achieved. Each input event $E$ has a representation as a firing pattern in this population of cells, and the problem of completing

a subevent $X$ of $E$ is equivalent to the problem of recovering its corresponding firing pattern. This pattern is called the *simple representation* of the input $E$.

### 1.4.4. *Advantages of the simple representation*

The notion of the simple representation of an event $E$ of the current internal description makes many of the problems of free and directed simple memory easy to express. A simple representation needs to be formed only of those parts of $E$ that contain the subevents through which $E$ will later be addressed: and the simple representation needs to be associated back (through the return from the memory) only to those parts of $E$ that will need to be recalled.

It will turn out that simple representations consist of collections of cells in a population whose activity $\alpha$(§ 1.2.4) is very low. The activity is in fact so low ($\alpha \approx 0.001$) that the cells of a simple representation can be directly associated to each other by collaterals terminating in Hebb synapses. The simple representation of $E$, written $[E]$, can thus be regarded as a firing pattern which can complete itself through its collateral synapses (called the *collateral effect*, § 2.4). Again, simple representations are somewhat limited in the maximum size they can attain, and this leads to the notion that more than one simple representation may be formed, each dealing with a different subevent of $E$. Within each simple representation, there is a full collateral effect, but between any two, it is less full (see §4.5.1).

## 2. The basic model for archicortex

### 2.0. *Introduction*

The arguments of § 1 show that simple memory may be divided into two operations: the creation of suitable diagnostic spaces for the input events as they occur; and the performance, during recall, of diagnostic operations within those spaces. The representation of these two basic functions requires a model consisting of two parts, closely analogous to codon formation and output cell selection in the neocortical theory. Many of the factors which determine the shape of each component have already arisen in the theory of the neocortex: they can therefore be derived rather quickly, and with this the first two parts of this section are concerned.

Within the outlines established by these two basic models, the actual shape of a simple memory is determined largely by numerical constraints. The rest of this section therefore shows how the capacities and characteristics of various models may be calculated, and derives the conditions imposed by the fact that the cells involved have to be physiologically plausible.

### 2.1. *Codon formation*

The first task to be discussed is the construction of evidence functions by input events. The obvious way to do this is to use the technique of codon formation, described in some detail by Marr (1970, § 4.3). (Compare also the *s*-cells of Brindley 1969.) The basic models for this appear in figure 2, and the arguments for each will be set out here only in so far as they differ from those put forward in the neocortical theory.

#### 2.1.1. *Preference for the model 2 using Brindley synapses*

The main differences between the arguments appropriate here and those for the neocortex arise because the function of simple memory is to record *all* its incoming information: the difficulties which arose in the neocortex, concerning the formation of evidence only over the appropriate

diagnostic space, do not arise here. Model 1 of figure 2 is excluded for the same reasons as in the neocortical theory: each cell can represent only one event, since after one modification, all the synapses not used in that event become ineffective. Model 3 is excluded for two reasons: (*a*) a climbing fibre system cannot both be simple and choose those cells most appropriate for each event (i.e. those at which the greatest number of active afferents have synapses); (*b*) a climbing fibre system in any case requires more cells than model (2).

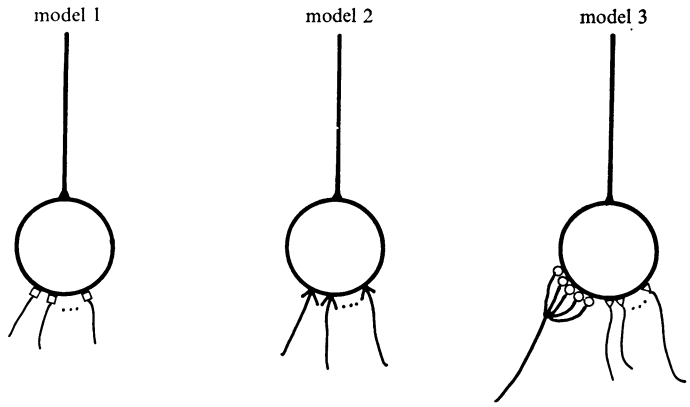



FIGURE 2. Three models for codon formation: model 1 uses synapses which are initially excitatory, but become ineffective as a result of post- without pre-synaptic activity; model 2 uses Brindley synapses; model 3 uses a climbing fibre and Hebb synapses.

2.1.2. *Threshold setting in model* 2

Brindley synapses contain an unmodifiable excitatory component, and are facilitated by a combination of pre- and post-synaptic depolarization. The post-synaptic threshold for the existence of modification conditions there will have to vary for two reasons: first the number of active afferents will not be constant; and secondly the overall proportion of synapses that have been modified will change, thus changing the amount of post-synaptic depolarization that an unlearned input of fixed size may be expected to cause. These problems do not arise in the special case considered by Brindley (1969), where the number of active afferents is always two, and the ratio of modifiable to unmodifiable components in the synapses is 1:2.

Synaptic modification probably depends on the local conditions prevailing in a piece of dendrite, and hence inhibition intended to prevent these conditions from arising must be applied directly to the dendrite. The use of Brindley synapses in codon formation therefore requires that inhibition of the appropriate strength should be applied to the dendrites containing those synapses.

There are broadly speaking two methods of providing such inhibition: either it is done by inhibitory cells which are otherwise identical to the codon cells; they learn inputs at the same rate, and are therefore excited at a rate which increases with the number of learnt events: or a negative feedback system is used, built to keep the number of codon cells that are active roughly constant. The first scheme is probably unsatisfactory, and the second is embodied in the model of figure 3. This model contains two kinds of inhibitory influence on the codon cell dendrites (often through different dendrites of the same inhibitory cell—e.g. the *G*-cells). One influence,

the inhibition driven directly by the afferent fibres, sets the cell thresholds on the assumption that no synapses have been modified. The other, a negative feed-back driven by codon cell axon collaterals via the *G*-cells, provides the component required to counteract the extra excitation which arises because a fraction of the population of synapses will have been modified by previous events. The system is imagined to be constructed so as to maintain a constant activity $\alpha$ in the set $\mathscr{P}$ of codon cells. The effect of all inhibition described here is subtractive, and dendritic branches which are not close are imagined to be independent.

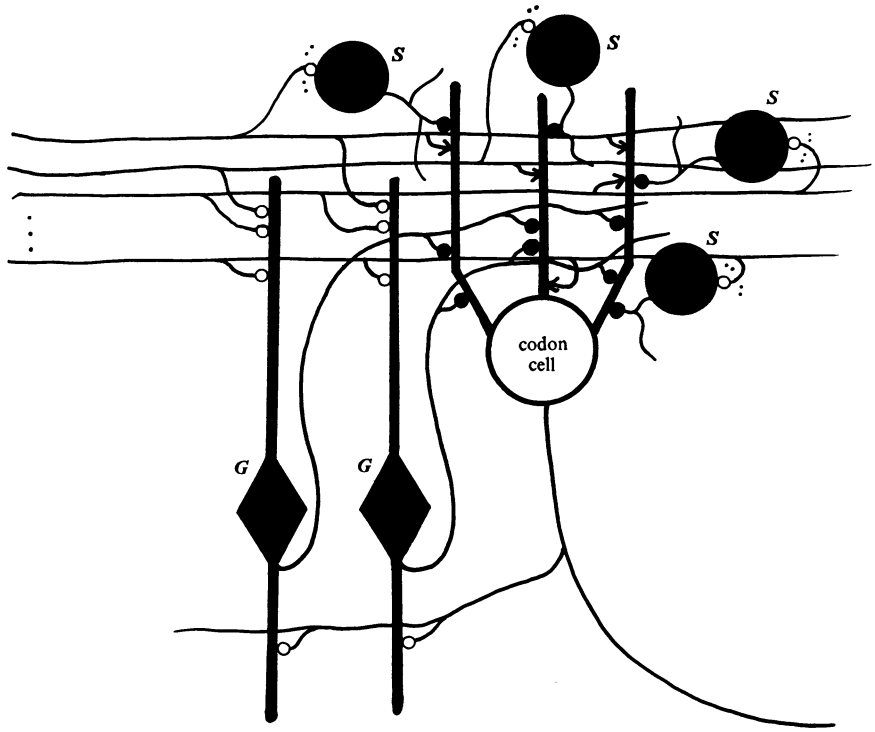



FIGURE 3. The full model for codon formation using Brindley synapses. Modification conditions are decided locally in the codon cell dendrites, and hence inhibition which controls these conditions is itself applied to the dendrites. The *S*-cells, driven by codon cell afferents, subtract roughly the expected excitation due to the unmodifiable component of the Brindley synapses. The *G*-cells, driven in part by codon cell axon collaterals, use negative feedback to compensate for changes in the size of the input event, and in the number of synapses which will already have been modified.

G. S. Brindley (personal communication) has pointed out that the need for *G*-cells in codon formation evaporates if information decays in the memory at about the same rate as it is acquired.

### 2.1.3. *Recalling an event*

The recall of an event is initiated by addressing the memory with a subevent. In order to avoid the problem of how the memory knows whether to store a given input, or to use it to recall the event most like it, it will be assumed that events which are to be stored are much larger than the

subevents which initiate recall. The reason for making this assumption is that the effect of a small subevent on the dendrites of the codon cells may then be regarded as being too mild to provoke synaptic modification there, since synaptic modification presumably requires a rather severe dendritic depolarization. The more general problem of controlling when a memory does and does not store its inputs will be dealt with in the paper on input–output relations.

One other point is needed to complete the discussion of codon cells. If the subevent that is being used for recall is wholly contained in the event to be recalled, then the best strategy is to lower the codon cell threshold until about the usual number of cells becomes active. This step is part of the usual procedure for implementing the interpretation theorem (Marr 1970, §§ 2.5 and 4.4). If, however, the subevent is only partially contained in the event to be recalled, then it will be shown in § 3.1 that better results are obtained if codon cells are treated like output cells (see § 2.2). This is essentially because output cells (with afferent basket synapses) are regarded as being capable of performing a division (Marr 1970, § 4.1.6); and, in the second situation, it turns out that the fraction of active afferent synapses which have been modified is a more suitable measure than the absolute number of such synapses.

### 2.2. *Diagnosis in simple memory*

It has been argued informally (§ 1.4.3) that the recall process in a simple memory has to be virtually complete by the time information is returned to the neocortical pyramidal cells. This means that the memory must contain internal diagnostic structure capable of recovering the pattern of firing appropriate to the learnt event of which the current input subevent formed a part. In this section, the cells at which the recovery is performed are described.

#### 2.2.1. *The simple representation*

In the neocortical theory, it was imagined that information was represented by a family of classes, each of which was formed because of a clustering of input subevents. The function of simple memory is to record information as it occurs, without trying to produce the best possible classification of the input on the spot. It is proposed that information in a simple memory is also represented by a family of classes, but that in this case, the classes are chosen randomly. An incoming event is assigned to a family of cells, analogous to neocortical output cells, chosen because they happen to have more relevant synapses than any others. These cells may be regarded as 'random' variables taking the value 0 or 1: the probability that they have the value 1 is assessed at each moment by consulting the relevant evidence, in the usual way.

When viewed as random classes in this way, it is seen that the diagnosis and interpretation theorems may be applied to the assessment of the incoming evidence: indeed, these results, strictly speaking, are more accurately applied to the problem of the diagnosis of random classes than of the more organized objects for which they were developed (Marr 1970, § 2). Since it is assumed that modifiable synapses for simple memory have all-or-none modification characteristics, it follows that they should transmit a measure of the fraction $f$ of their active afferent synapses which have been modified, provided that $f$ exceeds some (variable) lower bound $p$ (say).

It is thus proposed that the simple memory sets up, by a more or less random process, a set of classes which is unique (with very high probability) to each input. Each class is represented by a separate cell, although a given cell may represent more than one class. The set of cells which represent a given input in this way is called the *simple representation* of that input. The recall of an

event from a subevent is performed by recovering those classes by which the subevent is best interpreted, in the sense of the interpretation theorem (Marr 1970, § 2.5). In order to do this, the cells involved in a simple representation need to be able to measure the fraction $f$ defined above.

2.2.2 *Output cells for a simple representation*

The theory of output cells for the random classes described in § 2.2.1 falls into two parts: the first describes the formation of the classes, and the second deals with the subsequent interpretation of inputs. The idea that these cells do two things—i.e. store and interpret—and that they do both things all the time, leads naturally to the question of how they know what to do to a given input. For now it is enough to assume that if an input is a subevent of a previously learnt event, it will automatically cause recall of that event. If not, it is simply stored.

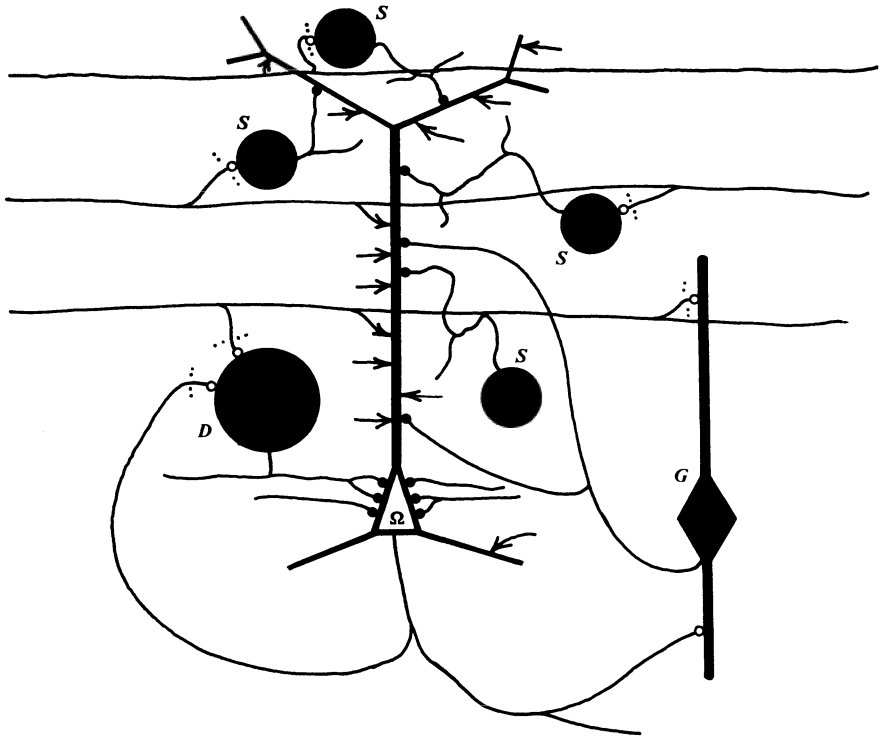



FIGURE 4. The output cell Ω has three kinds of afferent synapse: Brindley synapses (arrows) from codon cells, and two kinds of inhibitory synapses. Those from *S*- and *G*-cells are spread over the dendritic tree (cf. figure 3), and their effect is subtractive: those from the *D*-cells, concentrated at the soma, perform a division.

The problem of the formation of classes for the simple representation of an input has much in common with the problems surrounding codon formation. The central requirement is to choose, from the given population of cells, those which are best suited to representing the current input. This is exactly the problem that was discussed in § 2.1.1, and the possible mechanisms are again those of figure 2. For the same reasons as were given there, Brindley synapses provide the most suitable method of selecting such cells, and may therefore be expected at the cells involved in a simple representation.

It is interesting to note that output cells for the random classes involved in a simple representation require Brindley synapses, whereas output cells for classificatory units proper are best served by having climbing fibres. The freedom allowed by Brindley synapses—independence of different dendrites, and the ability to choose the most appropriate cells—which is such an advantage in simple memory, is only a disadvantage in the neocortical representation of classificatory units. The reason is that in the neocortex, it is crucial that all the relevant evidence for deduction of a property be held at the synapses of a single cell. Modification conditions have to occur everywhere on its dendrites simultaneously, and for all (or enough) of the relevant subevents. Without a climbing fibre, this cannot easily be arranged: a cell which is optimal for one subevent is not especially likely to be optimal for its neighbours as well.

The second part of output cell theory for a simple representation concerns the diagnosis of incoming events. Most of the problems that arise have been considered in output cell theory for the neocortex (Marr 1970, §4.1). These arguments show that two kinds of inhibition are needed: one to perform a subtraction (the *S*-cells of figure 4), and one to perform a division (the *D*-cells or basket cells of figure 4). Such cells would cause the output cells' firing rates to be proportional to $f-p$. In the present case, however, some further information is available: the output cells for a particular event were originally selected (through Brindley synapses) because they had the greatest number of active afferent synapses. Such cells will therefore tend to have more modified active afferent synapses during recall than other cells, and preliminary selection can usefully be made by subjecting the population of output cells to a suitable absolute threshold $T$ (say). In figure 4, it is imagined that inhibition to produce this is provided by the *G*-cells (driven in part by output cell axon collaterals). *G*-cells thus have two functions: to arrange suitable modification conditions during the storage of an event, and to provide a (variable) absolute threshold $T$ during recall. It will be shown in §3.3 that the introduction of two kinds of threshold into output cell theory—i.e. specifying both $T$ and a lower bound on $f$—greatly improves the performance of a memory.

In figure 5, the apparatus of figure 3 is added to that of figure 4 to produce the basic unit of simple memory. This type of model is examined in detail in §3.

### 2.2.3. *Structural differences between archicortex and neocortex*

There are various differences in the fine structure of the models devised for archi- and neocortex, of which perhaps the most striking concerns the absence of climbing fibres in archicortex. It is also possible to deduce differences that are predicted by the theory and which concern the large-scale arrangements of the two structures. If all of a large population of output cells tend to receive afferents from the same collection of evidence cells, the disposition of cells and fibres will contrast strongly with their arrangement in neocortex, where one expects that evidence cells are relatively private constructions. There is no reason in archicortex to have evidence and output cells particularly near one another: one can therefore expect to find cells involved in different stages placed rather far apart, and joined by powerful projections. (The so-called perforant path in the hippocampal formation may be an example of such a projection.)

For this reason, the numerical analysis which follows (§3.1) deals with layers of cells $\mathscr{P}_i$, which project to one another with various contact probabilities. Some layers will contain evidence cells, and some, output cells. The difference is however unimportant except in calculations about the recalling abilities of the system.

### 2.3. *The basic equation, and various constraints*

The calculation of the capacity and recalling ability of the simple memory described in § 2.2 rests on various assumptions and approximations. These are set out together in this section, and the relations derived here are used in § 3.

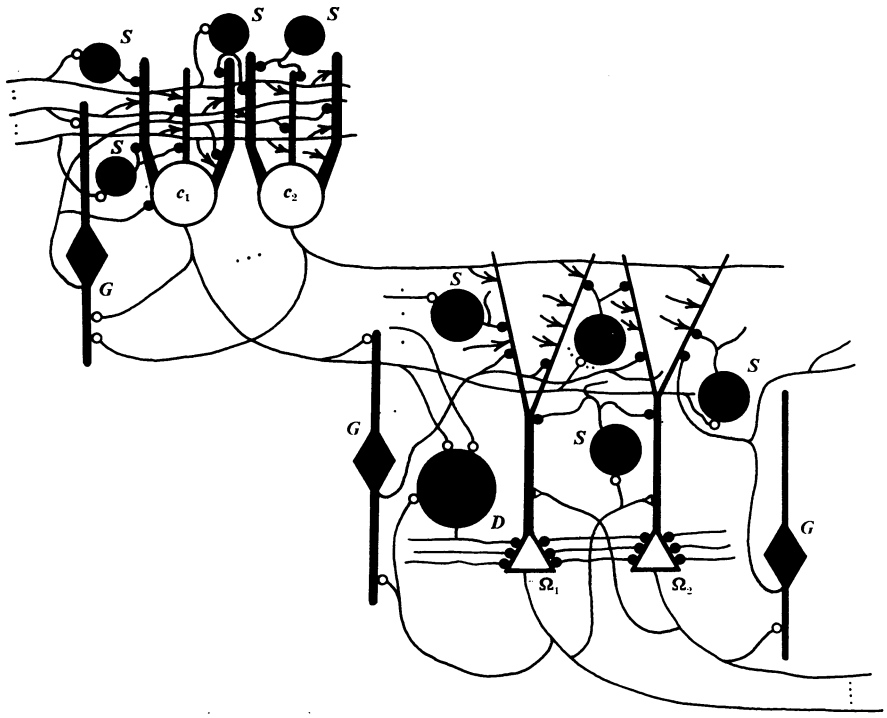



FIGURE 5. A model for simple memory, obtained by combining figures 3 and 4. The output cell axons return to the cells of the current internal description, after giving off collaterals which terminate in Hebb synapses at other output cells. This kind of model is analysed in § 3.1.

#### 2.3.0. *Notation*

$\mathscr{P}_i$ ($i = 0, 1, 2, \ldots$) is a population of $N_i$ cells with activity $\alpha_i$. The set of cells of $\mathscr{P}_i$ which fire in response to an input is called the $\mathscr{P}_i$-*representation* of the input. The terms *event*, *subevent*, and *codon* will have their usual meanings. In addition, the following notation will be standard:

$\boldsymbol{E}$ denotes an expectation;
$N_i$ the number of cells in $\mathscr{P}_i$;
$L_i$ the number of active cells of $\mathscr{P}_i$ $(L_i = \alpha_i N_i)$;
$R_i$ the threshold of the cells in $\mathscr{P}_i$ during the storage of information;
$S_i$ the number of afferent synapses possessed by each cell of $\mathscr{P}_i$ (assumed constant over $\mathscr{P}_i$);
$Z_i$ the contact probability for the projection of the afferent fibres to $\mathscr{P}_i$ (usually from $\mathscr{P}_{i-1}$). (Thus $Z_i$ = the probability that an arbitrary cell of $\mathscr{P}_i$ receives a synapse from an arbitrary cell of $\mathscr{P}_{i-1}$);
$\Pi_i$ the probability that an arbitrary afferent modifiable synapse in $\mathscr{P}_i$ has been modified.

2.3.1. *The response in $\mathscr{P}_i$ to an input event*

If it is assumed that the afferents to $\mathscr{P}_i$ distribute there randomly with contact probability $Z_i$, the variables defined in § 2.3.0 are related by the following equation:

$$\boldsymbol{E}(L_i) = N_i \sum_{r=R_i}^{L_{i-1}} \binom{L_{i-1}}{r} Z_i^r (1-Z_i)^{L_{i-1}-r} \quad \text{(Marr 1970, § 3).} \tag{2.1}$$

$L_i$ is the sum of expectations (corresponding to the individual terms of the expression), of which one (obtained by putting $r = R_i$) will usually be far larger than the rest. This is because $R_i$ will usually be chosen to keep $\alpha_i$ rather small, which implies that only the terms in the tail of the binomial distribution are in practice used.

2.3.2. *Modifiable synapses in $\mathscr{P}_i$*

It is helpful to have a rough guide as to when it is useful to have synaptic modification at the cells of $\mathscr{P}_i$. Fortunately, it is easy to obtain a simple approximate criterion for this. $\alpha_i = L_i/N_i$ is the activity in $\mathscr{P}_i$: let $\alpha_{i-1} = L_{i-1}/N_{i-1}$ be the activity of the input fibres. This is done because the input to $\mathscr{P}_i$ will be from the population of cells $\mathscr{P}_{i-1}$. It is roughly true that the proportion of synapses active at each active cell of $\mathscr{P}_i$ is $\alpha_{i-1}$: it is certainly at least this; the amount by which it exceeds it decreases as the value of $S_i\alpha_{i-1}$ increases. Therefore, the probability that after $n$ events, an arbitrary synapse of $\mathscr{P}_i$ has been facilitated is $(1-\alpha_{i-1})^{n\alpha_i}$, which is approximately $1-\exp(-n\alpha_{i-1}\alpha_i)$ if $\alpha_{i-1}$ is small. It is only worth having modifiable synapses in $\mathscr{P}_i$ if, when the inputs have all been learned, not all the synapses there have almost certainly been facilitated—that is, if $n\alpha_{i-1}\alpha_i$ is of the order of 1. Hence a rough, necessary condition that it be useful to have modifiable synapses in $\mathscr{P}_i$ is

$$n\alpha_{i-1}\alpha_i \lesssim 1. \tag{2.2}$$

2.3.3. *The condition for full representation*

The second constraint also embodies a necessary condition—that the activity in $\mathscr{P}_i$ provides an adequate representation of the input event. In the present context, a rather weak criterion of adequacy is sufficient, namely that a change in the firing of the input fibres should produce a change in the cells which are firing in $\mathscr{P}_i$.

The probability that an arbitrary but fixed active input fibre to $\mathscr{P}_i$ does not terminate at any active cell of $\mathscr{P}_i$ is approximately $(1-S_i\alpha_{i-1}/L_{i-1})^{L_i}$. This is approximately

$$\exp(-\alpha_{i-1}S_iL_i/L_{i-1}) = \exp(-S_i\alpha_i N_i/N_{i-1}).$$

Most of the active cells of $\mathscr{P}_i$ would cease to fire if one of their active afferents were removed (by the remarks of § 2.3.1 about the tail of a binomial distribution), and hence the condition for full representation of the input in $\mathscr{P}_i$ is that the probability $\exp(-S_i\alpha_i N_i/N_{i-1})$ should be kept very small—say less than $e^{-20}$. The condition then becomes

$$S_i\alpha_i N_i \geqslant 20N_{i-1}; \quad \text{i.e.} \quad S_iL_i \geqslant 20N_{i-1}. \tag{2.3}$$

If $\mathscr{P}_i$ is being used to capacity, i.e. $n\alpha_{i-1}\alpha_i \sim 1$, we find that

$$S_iN_i \gtrsim 20L_{i-1}n. \tag{2.4}$$

2.3.4. *Four practical constraints*

It must always be remembered that the cells and synapses of $\mathscr{P}_i$ are physiological objects, which cannot be asked to perform unrealistic feats. One tendency of the theory is to use the populations $\mathscr{P}_i$ of cells with very low activities, $\alpha_i$. The thresholds of the cells in $\mathscr{P}_i$ have, however, to be set by negative feedback devices, like the *G*-cells of figure 3, and these are to a certain extent limited as to what they can do.

The basic difficulty lies in specifying the proportion of active afferent synapses to which a cell may reasonably expect to be sensitive. Negative feedback devices like the *G*-cell will operate by measuring afferent synaptic activity, and inhibiting the cells with which they synapse in such a way as to keep $\alpha$ at the appropriate value. In what follows, $\alpha$ will be assumed to exceed 0.001 since this figure seems about as small a fraction of active synapses as would allow the activity to be reliably detected. The true bound may be lower, but it cannot be a great deal lower, and certainly not by an order of magnitude.

The same problem applies to the cells of $\mathscr{P}_i$ as applies to the *G*-cells which set their thresholds. In the case where the $\mathscr{P}_i$-cells have Brindley modifiable afferent synapses, the conditions on $\mathscr{P}_i$-cells are probably more stringent than on their associated threshold controllers, since it seems plausible that a considerable degree of post-synaptic depolarization is necessary in a region of dendrite before the conditions for modification are created there. It is difficult to give a numerical translation of the condition on the proportion of active synapses necessary for implementing modification conditions: in what follows, the relevant lower bound will be taken to be 0.005. In practice, it will be possible to alleviate this difficulty by arranging for related synapses to be placed near one another on a dendrite.

Finally, the second tendency of the theory is to require that the number of synapses on a cell be as large as is plausible. Cragg (1967) has shown that the average number of synapses per cell in monkey motor cortex is 60000, and in monkey striate cortex it is 5600. Large archicortical cells are comparable with large motor pyramidal cells, so it is wise to restrict the possible value of $S_i$ to not much more than 60000. An absolute bound of $S_i \leqslant 10^5$ will always be assumed.

There is no direct information about the numbers of synapses on archicortical cells, or the contact probabilities of the various projections, or the activities ($\alpha_i$) of the various groups of cells. It will not be possible to apply detailed quantitative tests to the present theory's predictions until numerical information of this kind becomes available.

## 2.4. *The collateral effect*

Let $\mathscr{P}$ be the population of cells in which the simple representation of an input is formed. If each cell has about 60000 afferent synapses, then each one can probably learn about 100 input events (cf. the cerebellar Purkinje cells, Marr 1969). Hence, if the population as a whole is to learn about $10^5$ events, the activity $\alpha$ of $\mathscr{P}$ must be about $10^{-3}$.

Equation (2.2) of 2.3.2 shows that for learning to be profitable in $\mathscr{P}_i$ driven by cells of $\mathscr{P}_{i-1}$, it is necessary that $n\alpha_{i-1}\alpha_i \lesssim 1$. Let $\mathscr{P}_{i-1} = \mathscr{P}_i = \mathscr{P}$: then the condition becomes $n\alpha^2 \lesssim 1$, and is satisfied by the values of $n$ ($\approx 10^5$) and $\alpha$ ($\approx 10^{-3}$) appropriate to the cells of a simple representation. In other words, it is possible to make good use of learning in synapses from the cells of $\mathscr{P}$ to the cells of $\mathscr{P}$—that is, in synapses at cells of $\mathscr{P}$ driven by collaterals of other cells of $\mathscr{P}$. The practical importance of this is that an input to $\mathscr{P}$ need not be sufficient on its own to re-stimulate all the cells of the particular simple representation which that input is designed to stimulate: collateral activity in $\mathscr{P}$ will help the recall process. Provided that the afferent information causes

more than a critical fraction of the active cells in $\mathscr{P}$ to be cells of the required representation, the collateral system will take over, suppress the cells which should not be active, and stimulate those which should. The completion of a partially specified simple representation by $\mathscr{P}$-cell collaterals is called the *collateral effect*. It will be shown that the collateral effect is probably capable of completing a simple representation when the fraction of currently active cells which are in that representation is as low as one third.

The details of the structure required for the collateral effect are as follows:

(i) collaterals distributing in $\mathscr{P}$ with the appropriate contact probability (see § 3);
(ii) Hebb (or Brindley) modifiable synapses where the collaterals meet other cells of $\mathscr{P}$;
(iii) the usual inhibitory threshold controlling cells.

## 3. Capacity calculations

### 3.0. *Introduction*

For practical application of the theory, it is essential to have a firm grasp of the kind of performance that may be expected from the basic simple memory of § 2. This section gives the reader direct experience of the available storage and recall capacity, for reasonable values of the important parameters.

Storage of an event will be said to have been achieved when its simple representation has been formed; and recall of that event, when its simple representation has been recovered.

### 3.1. *Establishing and recovering a simple representation*

There are various arguments which roughly decide the number of cells and synapses in the different portions of the memory that is analysed here. The conclusions are stated first, in the form of specifications of properties of a network which will form simple representations. These conclusions are followed by the arguments which lead to them, and these, by remarks about the memory's storage and recall performance.

#### 3.1.1. *The basic memory*

There are three populations of cells, $\mathscr{P}_1$, $\mathscr{P}_2$ and $\mathscr{P}_3$. The cells of $\mathscr{P}_1$ send axons to $\mathscr{P}_2$, and those of $\mathscr{P}_2$ send axons to $\mathscr{P}_3$. $\mathscr{P}_3$ possesses a collateral system, and it is in $\mathscr{P}_3$ that simple representations are formed. Table 1 shows the basic parameters for each of the $\mathscr{P}_i$, using the notation defined in § 2.3.0. It is imagined that the $10^6$ cells of $\mathscr{P}_1$ are split into 25 so-called *blocks* of cells, each of which projects exclusively to a corresponding block in $\mathscr{P}_2$ (see figure 6). The parameters for each block are given in table 2. The projection from $\mathscr{P}_2$ to $\mathscr{P}_3$ has no block structure, and table 3 describes the parameters for this projection. $\mathscr{P}_3$ also possesses a collateral system, which may be regarded as a projection from $\mathscr{P}_3$ to $\mathscr{P}_{3'}$. The parameters for the collaterals appear in table 3 in the column for $i = 3'$. These values have all been obtained using the equations of § 2.3.

The probability that an arbitrary synapse has been modified can easily be calculated if it is assumed that synapses are effectively chosen randomly each time an event is stored. The assumptions behind this have been set out already (Marr 1969, § 5) in the calculation of the capacity of a cerebellar Purkinje cell. Suppose $n$ events have been stored; then the probability $\Pi_i$ that an arbitrary modifiable synapse in $\mathscr{P}_1$ will have been facilitated is

$$\Pi_i = 1 - (1 - x_i/S_i)^{n\alpha_i},$$

where $\alpha_i$, $S_i$ are as in §2.3.0, and $x_i$ is the expected number of synapses used at an active cell for one event. $x_i$ is near to $R_i$, the threshold of such a cell: in fact

$$x_i = \sum_{R \geq R_i} P_i(R) . R,$$

where $P_i(R)$ is the probability that an active cell of $\mathscr{P}_i$ has exactly $R$ active afferent synapses. $P_i(R)$ is calculated from the terms of the equation in §2.3. Table 4 shows values of $\Pi_i$ for $n = 5 \times 10^4$, and $n = 10^5$ stored events.

TABLE 1. GROSS PARAMETERS FOR A SIMPLE MEMORY $\mathscr{P}_1 \rightarrow \mathscr{P}_2 \rightarrow \mathscr{P}_3$

Cells of $\mathscr{P}_2$ and $\mathscr{P}_3$ possess Brindley modifiable afferent synapses

| $i$ ... | 1 | 2 | 3 |
|---|---|---|---|
| $N_i$ | $1.25 \times 10^6$ | 500 000 | 100 000 |
| $L_i$ | 2500 | 3025 | 217 |
| $\alpha_i$ | 0.002 | 0.006 | 0.002 |

TABLE 2. $\mathscr{P}_1$ AND $\mathscr{P}_2$ OF TABLE 1 ARE SPLIT INTO 25 BLOCKS, EACH HAVING THE FOLLOWING SPECIFICATIONS:

| $i$ ... | 1 | 2 |
|---|---|---|
| $N_i$ | 50 000 | 20 000 |
| $L_i$ | 100 | 121 |
| $R_i$ | — | 31 |
| $S_i$ | — | 10 000 |
| $\alpha_i$ | 0.002 | 0.006 |
| $Z_i$ | — | 0.2 |

TABLE 3. THE PROJECTION $\mathscr{P}_2 \rightarrow \mathscr{P}_3$ HAS NO BLOCK STRUCTURE, AND HAS THE FOLLOWING PARAMETERS:

| $i$ ... | 2 | 3 | 3′ |
|---|---|---|---|
| $N_i$ | 500 000 | 100 000 | 100 000 |
| $L_i$ | 3025 | 217 | 200 |
| $R_i$ | — | 351 | — |
| $S_i$ | — | 50 000 | 10 000 |
| $\alpha_i$ | 0.006 | 0.002 | 0.002 |
| $Z_i$ | — | 0.1 | 0.1 |

The column $i = 3'$ gives the parameters for the collateral system in $\mathscr{P}_3$.

The expected number of active afferent collateral synapses at a cell of $\mathscr{P}_3$ is 21.7, but has been taken to be 20 for simplicity.

TABLE 4. MODIFICATION PROBABILITIES $\Pi_i$ FOR MODIFIABLE SYNAPSES IN EACH $\mathscr{P}_i (i = 2, 3, 3')$ AFTER $n$ EVENTS HAVE BEEN STORED

$i = 3'$ gives values for the collaterals in $\mathscr{P}_3$

| $n$ | $\Pi_2$ | $\Pi_3$ | $\Pi_{3'}$ |
|---|---|---|---|
| $5 \times 10^4$ | 0.621 | 0.538 | 0.181 |
| $10^5$ | 0.857 | 0.787 | 0.330 |

### 3.1.2. *The collateral effect in* $\mathscr{P}_3$

The collateral system in $\mathscr{P}_3$ can aid the recovery of a simple representation in the following way. Suppose that an input $X$ is presented at $\mathscr{P}_1$, and that $X$ is a subevent of a previously learnt event $E_0$. Let $\mathscr{P}_{30}$ denote the simple representation of $E_0$ in $\mathscr{P}_3$ and let $\mathscr{P}_{31}$ denote the rest of $\mathscr{P}_3$.

Suppose that $X$ causes firing of $C_0$ cells in $\mathscr{P}_{30}$, and $C_1$ cells in $\mathscr{P}_{31}$. Since $E_0$ has already been learnt, all collateral synapses between cells of its simple representation will have been facilitated. Hence collateral synapses between cells of $\mathscr{P}_{30}$ will all have been facilitated, whereas those between other cells will have no more than the usual probability of having been facilitated.

In order to analyse the effects of the $\mathscr{P}_3$ collaterals, it is assumed that once firing in the collection $\mathscr{P}_3$ has been established by the afferents from $\mathscr{P}_2$, these afferents become silent, and the cells in $\mathscr{P}_3$ are driven solely by the collaterals. The effects of the collaterals alone can be discovered by regarding $\mathscr{P}_3$ as projecting to an identical set of cells, called $\mathscr{P}_{3'}$, in the same way as the collaterals distribute among the cells of $\mathscr{P}_3$. The behaviour of $\mathscr{P}_{3'}$, which represents the new state of $\mathscr{P}_3$ after one 'application' of the transformation on the $\mathscr{P}_3$ firing pattern induced by the collaterals, can then be calculated using the equations of § 2.3.

In the present theory, the important question is whether or not the collateral effect can lead to the recovery of the simple representation of $E_0$. Whether this happens depends on the parameters associated with the collateral distribution, and on the relative sizes of $C_0$ and $C_1$. For fixed parameters there is a threshold for the ratio $C_0 : C_1$ above which the collaterals will tend to increase this ratio, and below which they will tend to decrease it. The threshold is of a statistical nature, because above it, the collaterals are more likely to increase the ratio, and below it, they are more likely to decrease it. One has to move a little way away from this threshold before the outcome either way is virtually certain.

The *statistical threshold* (for $C_0 + C_1 = L_3$) is defined as the value of the ratio $C_0 : C_1$ such that the expected effect of the collaterals is to maintain it. It may be calculated as follows.

Let $\boldsymbol{b}$ be an arbitrary cell of $\mathscr{P}_{3'}$, the copy of $\mathscr{P}_3$ to which the collaterals are imagined to project. The number of active afferent synapses at $\boldsymbol{b}$ comes from a binomial distribution $b(L_3; Z_{3'})$ with expectation $L_3 Z_{3'}$ from population $L_3$. $L_3$ is the number of active cells in $P_3$ and $Z_{3'}$ is the collateral contact probability. Hence the probability that $\boldsymbol{b}$ has exactly $x$ active afferent synapses is

$$P_{3'}(x) = \binom{L_3}{x} Z_{3'}^x (1 - Z_{3'})^{L_3 - x}. \tag{3.1}$$

If $\boldsymbol{b}$ is not in $\mathscr{P}_{30}$, the simple representation of $E_0$, the number of these active synapses that will have been facilitated is drawn from the binomial distribution $b(x : \Pi_{3'})$ with expectation $x\Pi_{3'}$ from population of size $x$ (from the definition (§2.3.0) of $\Pi$). Hence if $Q_{3'1}(r)$ denotes the probability that exactly $r$ of the $x$ active afferent synapses to $\boldsymbol{b}$ have been modified,

$$Q_{3'1}(r) = \binom{x}{r} \Pi_{3'}^r (1 - \Pi_{3'})^{x - r}. \tag{3.2}$$

If $\boldsymbol{b}$ is in $\mathscr{P}_{30}$, all afferent synapses from other cells in $\mathscr{P}_{30}$ will have been modified. Hence the number of active afferent modified synapses at a cell in $\mathscr{P}_{30}$ is composed of two contributions: one, with distribution $b(C_0; Z_{3'})$ from cells of $\mathscr{P}_{30}$ with probability $Z_{3'}$, all of which have been modified: and one with distribution $b(C_1; Z_{3'})$ from $\mathscr{P}_{31}$ which have only chances given by (3.2) of having been modified. For the purposes of calculation, this situation has been approximated by assuming that, for a cell in the simple representation of $E_0$ with $x$ active afferent synapses, the number of those synapses which have been facilitated has distribution

$$b(x; (C_0 + C_1 \Pi_{3'})/(C_0 + C_1)).$$

Hence if $Q_{3'0}(r)$ denotes the probability that exactly $r$ of the $x$ active afferent synapses to $\boldsymbol{b}$ have been modified,

$$Q_{3'0}(r) = \binom{x}{r} (C_0 + C_1)^x (C_0 + C_1 \Pi_{3'})^r (1 - C_0 - C_1 \Pi_{3'})^{x - r}. \tag{3.3}$$

Hence, if the cells in $\mathscr{P}_{3'}$ all have a threshold $R$, the expected number of active cells that are not in the simple representation of $E_0$ is

$$C_1' = (N_3 - L_3) \sum_{r \geq R} \sum_{x=r}^{L_3} P_{3'}(x)\, Q_{3'1}(r), \tag{3.4}$$

and the expected number of active cells in $\mathscr{P}_{30}$ is

$$C_0' = L_3 \sum_{r \geq R} \sum_{x=r}^{L_3} P_{3'}(x)\, Q_{3'0}(r). \tag{3.5}$$

Thus, when all cells of $\mathscr{P}_{3'}$ have threshold $R$, the effect of the collaterals is to transform $C_0$ and $C_1$ into new numbers with expectations $C_0'$ and $C_1'$. Hence the statistical threshold, as defined above, for recovery of the simple representation of $E_0$ is that ratio $C_0 : C_1$ for which

$$C_0 : C_1 = C_0' : C_1', \text{ subject to } C_0 + C_1 = C_0' + C_1' \approx L_3. \tag{3.6}$$

In practice, however, the cells will not have a uniform threshold, since the theory allows that division can take place as well as subtraction. The effect of division may be incorporated by assuming that a cell only fires if at least a fraction $f$ of its active afferent synapses have been facilitated: $f$ is called the *division threshold* of the cell. The combined effects of a subtractive threshold $T$ and a division threshold $f$ are to give a cell $\boldsymbol{b}$ of $\mathscr{P}_3$, with $x$ active afferent synapses, a threshold $R = R(\boldsymbol{b})$ where

$$R(\boldsymbol{b}) = \max\{T, fx\}.$$

This transforms $C_i'$ of (4) and (5) into $C_i^*$ where

$$C_1^* = (N_3 - L_3) \sum_{r \geq \max\{T, fx\}} \sum_{x=r}^{L_3} P_{3'}(x)\, Q_{3'1}(r), \tag{3.7}$$

$$C_0^* = L_3 \sum_{r \geq \max\{T, fx\}} \sum_{x=r}^{L_3} P_{3'}(x)\, Q_{3'0}(r). \tag{3.8}$$

The statistical threshold becomes that ratio $C_0 : C_1$ for which

$$C_0 : C_1 = C_0^* : C_1^*, \text{ subject to } C_0 + C_1 = C_0^* + C_1^* \approx L_3, \tag{3.9}$$

the threshold parameters $T, f$ being chosen to minimize $C_0^*/C_1^*$. The expectations $C_0^*$, $C_1^*$ have been computed for the relevant parameters, and selected values appear in the tables 5 to 7. Cases $C_0 + C_1 = L_3$ and $C_0 + C_1 = \frac{1}{2}L_3$ have both been calculated, since it is often better to use the smaller values during recall. The case $n = 10^5$ and $C_0 + C_1 = L_3$ resembles table 6 in the same way as table 7 resembles table 5. Various other tables have been computed, and the statistical thresholds obtained for selected values of $L_3$ and $Z_{3'}$ are given in table 8.

Three points are worth noting about these results. First, $Z_{3'} = 0.2$ gives a statistical threshold about twice as good as that for $Z_{3'} = 0.1$. Secondly, recovery of the whole of the simple representation depends upon suitable juggling of $T$ and $f$, and is complete after about 3 cycles. $f$ must start low, and increase as the representation is recovered: $T$ must decrease in such a way that the activity in $\mathscr{P}_3$ is kept roughly constant. And thirdly, the overall performance of the collateral effect is impressive (see table 8): recovery of the whole of the simple representation of $E_0$ is almost certain for values of about $0.1\,L_3$ greater than the statistical threshold value (assuming that $C_0 + C_1$ is constant).

The collateral effect is valuable in any population of cells where $n\alpha^2 \lesssim 1$. This condition may

often be satisfied in and between regions of neocortex, and the effect may be an important means of providing indirect 'associational' aid for the interpretation of sensory inputs (see Marr 1970, §2.4).

TABLE 5. THE COLLATERAL EFFECT IN $\mathscr{P}_3$

$N_3 = 100000$; $L_3 = 200$; $Z_{3'} = 0.1$. 50000 simple representations have been stored.

| $C_0$ | $C_1$ | $T$ | $f$ | $C_0^*$ | $C_1^*$ |
|---|---|---|---|---|---|
| 100 | 0 | 3 | 1.0 | 200 | 6 |
| | | 6 | 1.0 | 188 | 0 |
| 80 | 20 | 6 | 0.8 | 151 | 15 |
| | | 6 | 0.9 | 119 | 3 |
| 60 | 40 | 8 | 0.6 | 70 | 19 |
| | | 7 | 0.7 | 86 | 26 |
| 50 | 50 | 7 | 0.6 | 70 | 82 |
| | | 6 | 0.7 | 73 | 101 |
| 40 | 60 | 7 | 0.6 | 41 | 82 |
| | | 6 | 0.7 | 41 | 101 |

Statistical threshold ~ 50:50.

TABLE 6. THE COLLATERAL EFFECT IN $\mathscr{P}_3$

$N_3 = 100000$; $L_3 = 200$; $Z_{3'} = 0.1$. 100000 simple representations have been stored.

| $C_0$ | $C_1$ | $T$ | $f$ | $C_0^*$ | $C_1^*$ |
|---|---|---|---|---|---|
| 100 | 0 | 6 | 1.0 | 188 | 14 |
| | | 9 | 1.0 | 136 | 1 |
| 90 | 10 | 10 | 0.9 | 89 | 8 |
| | | 7 | 1.0 | 86 | 6 |
| 80 | 20 | 8 | 0.9 | 110 | 77 |
| | | 9 | 0.9 | 86 | 27 |
| 60 | 40 | 10 | 0.7 | 38 | 80 |
| | | 9 | 0.8 | 48 | 72 |

Statistical threshold ~ 85:15.

TABLE 7. THE COLLATERAL EFFECT IN $\mathscr{P}_3$

$N_3 = 100000$; $L_3 = 200$; $Z_{3'} = 0.1$. 50000 simple representations have been stored.

| $C_0$ | $C$ | $T$ | $f$ | $C_0^*$ | $C_1^*$ |
|---|---|---|---|---|---|
| 200 | 0 | 4 | 1.0 | 200 | 0 |
| | | 9 | 1.0 | 200 | 0 |
| 160 | 40 | 4 | 0.8 | 167 | 1 |
| | | 8 | 0.8 | 167 | 0 |
| 120 | 80 | 10 | 0.6 | 160 | 9 |
| | | 11 | 0.6 | 148 | 4 |
| 80 | 120 | 11 | 0.4 | 88 | 102 |
| | | 10 | 0.5 | 98 | 61 |
| 40 | 160 | 8 | 0.5 | 24 | 186 |
| | | 9 | 0.5 | 20 | 115 |

Statistical threshold ~ 60:140.

3.1.3. *Recall performance* $\mathscr{P}_2 \to \mathscr{P}_3$

The analysis of recall performance $\mathscr{P}_2 \to \mathscr{P}_3$ and $\mathscr{P}_1 \to \mathscr{P}_2$ follows the same general line as the arguments of §3.1.2, except that the equations apply only to individual blocks. Let $E_0'$ denote the restriction of the input event $E_0$ to one block $\beta$ of $\mathscr{P}_1$, and suppose, as in §3.1.2, that $E_0$ has already

been learnt. A new input event is presented to the block $\beta$, $A_0$ cells of which were active in $E_0'$ and $A_1$ of which were not. These in turn evoke (in the corresponding block of $\mathscr{P}_2$) $B_0$ cells which were also active in response to $E_0'$, and $B_1$ cells which were not. The firing in $\mathscr{P}_2$ causes the firing in $\mathscr{P}_3$ described by the numbers $C_0$, $C_1$ of § 3.1.2. The situation when more than one block of $\mathscr{P}_1$ is active can be solved by a simple extension of the methods used for exactly one block. Figure 6 illustrates the recall problem.

TABLE 8. ESTIMATED STATISTICAL THRESHOLDS (s.t.) FOR VARIOUS VALUES OF THE MAIN PARAMETERS

$N_3 = 100000$; $C = C_0 + C_1$; s.t. accurate $\pm 0.05 L_3$.

| $L_3$ | $Z_3$, | $C$ | $10^{-4}n$ | s.t. |
|---|---|---|---|---|
| 100 | 0.1 | 100 | 5 | 30:70 |
| \| | 0.1 | \| | 10 | 40:60 |
| \| | 0.2 | \| | 5 | 15:85 |
| \| | 0.2 | \| | 10 | 20:80 |
| 200 | 0.1 | \| | 5 | 50:50 |
| \| | 0.1 | \| | 10 | 85:15 |
| \| | 0.2 | \| | 5 | 30:70 |
| \| | 0.2 | \| | 10 | 50:50 |
| \| | 0.1 | 200 | 5 | 60:140 |
| \| | 0.2 | \| | 5 | 40:160 |

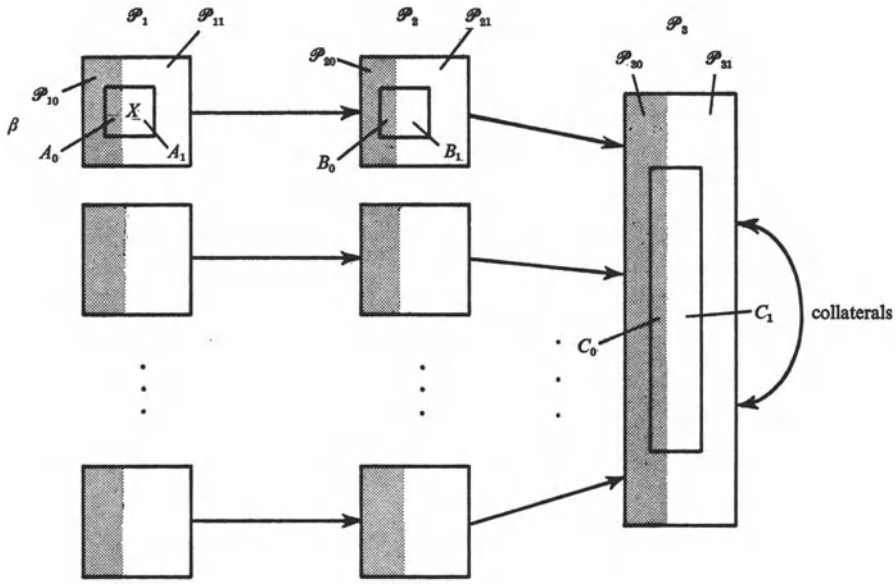



FIGURE 6. The recall problem. $\mathscr{P}_1$, $\mathscr{P}_2$ and $\mathscr{P}_3$ are the populations of cells defined in table 1. Shading represents the parts of these populations involved in the storage of an event $E_0$. A new subevent $X$ is presented to one block of $\mathscr{P}_1$, $A_0$ of whose cells were involved in $E_0$, and $A_1$ of which were not. This produces activity in one block of $\mathscr{P}_2$, and in $\mathscr{P}_3$. $B_0$ of the active cells in $\mathscr{P}_2$ were active in $E_0$, and $B_1$ were not: $C_0$ of the active cells in $\mathscr{P}$ were also active in $E_0$, and $C_1$ were not. The numbers $A_i$, $B_i$, $C_i$, $(i = 1, 2)$ are computed in the text.

The equations describing the relation between the $B_i$ and the $C_j$ $(i, j = 1, 2)$ are best derived through a series of steps. The notation of § 2.3 is assumed to hold for all processes concerned with the storage of the event $E_0$; for example, $L_3$ is the size of the simple representation of $E_0$ in $\mathscr{P}_3$. The relations between $L_i$, $N_i$, $R_i$, etc., are described by the equations of § 2.3.

S 1. *Additional notation*

The following symbols help to describe states occurring during recall. For $i = 1, 2, 3$:

$\mathscr{P}_{i0}$ = the set of cells of $\mathscr{P}_i$ which were in the $\mathscr{P}_i$-representation of $E_0$,

$\mathscr{P}_{i1}$ = the set of cells of $\mathscr{P}_i$ which were not in the $\mathscr{P}_i$-representation of $E_0$.

Thus there are

$C_0$ cells active in $\mathscr{P}_{30}$,

$C_1$ cells active in $\mathscr{P}_{31}$,

$B_0$ cells active in $\mathscr{P}_{20}$,

and $B_1$ cells active in $\mathscr{P}_{21}$.

Let $A_0$ be the number of active cells in $\mathscr{P}_{10}$,

and let $A_1$ be the number of active cells in $\mathscr{P}_{11}$.

S 2. *Calculation of contact probabilities*

The contact probability $\mathscr{P}_2 \to \mathscr{P}_3$ is $Z_3$, but the contact probability $\mathscr{P}_{20} \to \mathscr{P}_{30}$ is not $Z_3$, since the cells of $\mathscr{P}_{30}$ were selected (through Brindley synapses) because they had the most active afferent synapses from the $\mathscr{P}_2$-representation of $E_0$. Let $R_3$ be the threshold of the cells in $\mathscr{P}_3$ during the setting up of the simple representation of $E_0$: then the contact probability from the active cells of $\mathscr{P}_2$ to those of $\mathscr{P}_3$ at that time is

$$L_2^{-1} L_3^{-1} \sum_{r \geqslant R_3} N_3 \binom{L_2}{r} Z_3^r (1 - Z_3)^{L_2 - r} r = \xi_0 \quad \text{say:}$$

and the contact probability between active $\mathscr{P}_2$ cells and inactive $\mathscr{P}_3$ cells is depressed slightly: it is in fact $\xi_1$ where $\xi_1 = (N_3 Z_3 - L_3 \xi_0)/(N_3 - L_3)$. The contact probability between all other collections in $\mathscr{P}_2$ and $\mathscr{P}_3$ is $Z_3$. In the following calculations, it will be assumed that distributions between $\mathscr{P}_2$ and $\mathscr{P}_3$ are random, with the contact probabilities $\xi_0$, $\xi_1$, $Z_3$ between the special groups described above.

S 3. *Calculating the number of active synapses at a cell* $\boldsymbol{c}$ *of* $\mathscr{P}_3$

(i) If $\boldsymbol{c}$ is in $\mathscr{P}_{30}$ the number $s$ of synapses active at $\boldsymbol{c}$ is formed from two components: $s_0$ from the active cells in $\mathscr{P}_{20}$ and $s_1$ from the active cells in $\mathscr{P}_{21}$. $s_0$ comes from a binomial distribution $b(B_0; \xi_0)$, and $s_1$ from a binomial distribution $b(B_1; Z_3)$ (in the usual notation). Hence $P_{30}(s)$, the probability that exactly $s$ synapses are active at $\boldsymbol{c}$, is

$$P_{30}(s) = \sum_{s_0 + s_1 = s} \binom{B_0}{s_0} \xi_0^{s_0} (1 - \xi_0)^{B_0 - s_0} \binom{B_1}{s_1} Z_3^{s_1} (1 - Z_3)^{B_1 - s_1}.$$

(ii) If $\boldsymbol{c}$ is not in $\mathscr{P}_{30}$ the two components $s_0$ and $s_1$ have distributions $b(B_0; \xi_1)$ and $b(B_1; Z_3)$ respectively. Hence $P_{31}(s)$, the probability that exactly $s$ synapses are active at $\boldsymbol{c}$, is

$$P_{31}(s) = \sum_{s_0 + s_1 = s} \binom{B_0}{s_0} \xi_1^{s_0} (1 - \xi_1)^{B_0 - s_0} \binom{B_1}{s_1} Z_3^{s_1} (1 - Z_3)^{B_1 - s_1}.$$

S 4. *Calculating the number of active facilitated synapses at a cell* $\boldsymbol{c}$ *of* $\mathscr{P}_3$

(i) Let $\boldsymbol{c}$ be in $\mathscr{P}_{30}$ and have $s$ active afferent synapses, made up from the two components $s_0$ and $s_1$ of S 3(i). All the $s_0$ synapses will have been facilitated, and the number of the $s_1$ synapses

which will have been facilitated has distribution $b(s_1; \Pi_3)$ where $\Pi_3$ is the probability that an arbitrary $\mathscr{P}_3$ afferent synapse has been facilitated. So the probability that $\boldsymbol{c}$ has exactly $r$ active afferent facilitated synapses is $Q_{30}(r)$ where

$$Q_{30}(r) = \sum_{s_0=0}^{r} \left\{ \binom{B_0}{s_0} \xi_0^{s_0}(1-\xi_0)^{B_0-s_0} \sum_{s_1 \geqslant r-s_0} \binom{B_1}{s_1} Z_3^{s_1}(1-Z_3)^{B_1-s_1} \binom{s_1}{r-s_0} \Pi_3^{r-s_0}(1-\Pi_3)^{s_0+s_1-r} \right\}.$$

(ii) If $\boldsymbol{c}$ is in $\mathscr{P}_{31}$ the probability $Q_{31}(r)$ that $\boldsymbol{c}$ has exactly $r$ active afferent modified synapses is

$$Q_{31}(r) = \sum_{s \geqslant r} \binom{s}{r} \Pi_3^r(1-\Pi_3)^{s-r}\{P_{31}(s)\},$$

since all active afferent synapses have chance $\Pi_3$ of having been facilitated.

S 5. *Calculating the cells' thresholds*

All the cells in $\mathscr{P}_3$ are assumed to be subject to two kinds of threshold: an absolute threshold of $T_3$ (say), and a division threshold (defined in §3.1.2) of $f_3$. Thus if a cell has $s$ active afferent synapses, its threshold is set at

$$R_3 = \text{maximum}\,\{T_3, sf_3\}.$$

S 6. *Calculating expected numbers of active cells*

There are $L_3$ cells in $\mathscr{P}_{30}$ and $(N_3 - L_3)$ cells in $\mathscr{P}_{31}$. It is assumed that the cells of $\mathscr{P}_3$ are subject to thresholds $(T_3, f_3)$ of S 5. Then the expected numbers of cells active in $\mathscr{P}_{30}$ and $\mathscr{P}_{31}$ are respectively:

$$C_0 = L_3 \sum_{s_0+s_1 \geqslant T_3} \sum_{r \geqslant R_3} Q_{30}(r), \quad \text{where} \quad R_3 = \max\{T_3, (s_0+s_1)f_3\},$$

$$C_1 = (N_3 - L_3) \sum_{s \geqslant T_3} \sum_{r \geqslant R_3} Q_{31}(r), \quad \text{where} \quad R_3 \text{ is as defined in S 5 above.}$$

Close approximations to these distributions have been computed for various values of the important parameters, and some results appear in table 9. They are summarized in §3.1.5.

3.1.4. *Recall performance* $\mathscr{P}_1 \to \mathscr{P}_2$

The problem of describing the effect of presenting a learnt subevent to $\mathscr{P}_2$ can be solved by calculating the values of $B_0$, $B_1$ in terms of $A_0$ and $A_1$ (defined in S 1 of §3.1.3). These relations are very similar to those holding between the $B_i$ and the $C_j$ $(i,j = 0, 1)$. The following steps $S$ are analogous to those of §3.1.3, and can be derived by the same arguments. Write $\eta_0$ for the contact probability between the active cells of $\mathscr{P}_1$ and $\mathscr{P}_2$ during the original setting up, and write $\eta_1$ for the contact probability between active $\mathscr{P}_1$ cells and inactive $\mathscr{P}_2$ cells. $\eta_0$ corresponds to $\xi_0$ and $\eta_1$ to $\xi_1$.

S 2 (i) $\eta_0 = L_1^{-1}L_2^{-1} \sum_{r \geqslant R_2} N_2 \binom{L_1}{r} Z_2^r(1-Z_2)^{L_2-r}r,$

(ii) $\eta_1 = (N_2Z_2 - L_2\eta_0)/(N_2 - L_2).$

S 3 (i) $P_{20}(s) = \sum_{s_0+s_1=s} \binom{A_0}{s_0} \eta_0^{s_0}(1-\eta_0)^{A_0-s_0} \binom{A_1}{s_1} Z_2^{s_1}(1-Z_2)^{A_1-s_1},$

(ii) $P_{21}(s) = \sum_{s_0+s_1=s} \binom{A_0}{s_0} \eta_1^{s_0}(1-\eta_1)^{A_0-s_0} \binom{A_1}{s_1} Z_2^{s_1}(1-Z_2)^{A_1-s_1}.$

S 4 (i) $$Q_{20}(r) = \sum_{s_0=0}^{r} \left\{ \binom{A_0}{s_0} \eta_0^{s_0}(1-\eta_0)^{B_0-s_0} \times \sum_{s_1 \geqslant r-s_0} \binom{A_1}{s_1} Z_2^{s_1}(1-Z_2)^{B_1-s_1} \binom{s_1}{r-s_0} \Pi_2^{r-s_0}(1-\Pi_2)^{s_0+s_1-r} \right\},$$

(ii) $$Q_{21}(r) = \sum_{s \geqslant r} \binom{s}{r} \Pi_2^r(1-\Pi_2)^{s-r}\{P_{21}(s)\}.$$

S 5 $$R_2 = \text{maximum}\{T_2, sf_2\}.$$

S 6 (i) $$B_0 = L_2 \sum_{s_0+s_1 \geqslant T_2} \sum_{r \geqslant R_2} Q_{20}(r), \quad \text{where} \quad R_2 = \max\{T_2, (s_0+s_1)f_2\},$$

(ii) $$B_1 = (N_3 - L_3) \sum_{s \geqslant T_3} \sum_{r \geqslant R_3} Q_{21}(r), \quad \text{where } R_2 \text{ is as defined in S 5 of this section.}$$

Close approximations to these distributions have been computed for various values of the important parameters, and selected results are shown in table 10.

TABLE 9. ADDRESSING $\mathscr{P}_3$ WITH AN INPUT, FROM ONE BLOCK OF $\mathscr{P}_2$, WHICH CONTAINS A SUBEVENT OF A LEARNT EVENT $E_0$

The simple representation of $E_0$ occupied 217 cells of $\mathscr{P}_3$; $n$ such representations have been stored. Notation is from the text (see figure 6).

| $B_0$ | $B_1$ | $T_3$ | $f_3$ | $C_0$ | $C_1$ |
|---|---|---|---|---|---|
| $n = 50\,000$ | | | | | |
| 120 | 0 | 11 | 1.0 | 184 | 27 |
| | | 12 | 1.0 | 166 | 13 |
| | | 13 | 1.0 | 144 | 6 |
| | | 14 | 1.0 | 120 | 3 |
| 100 | 20 | 13 | 0.92 | 101 | 126 |
| | | 14 | 0.92 | 78 | 53 |
| | | 15 | 0.92 | 57 | 21 |
| | | 11 | 1.0 | 56 | 27 |
| 80 | 40 | 15 | 0.75 | 35 | 141 |
| | | 15 | 0.83 | 33 | 79 |
| | | 13 | 0.92 | 51 | 127 |
| | | 14 | 0.92 | 36 | 54 |
| 60 | 0 | 8 | 1.0 | 89 | 110 |
| | | 9 | 1.0 | 58 | 36 |
| 45 | 15 | 10 | 0.75 | 16 | 113 |
| | | 8 | 1.0 | 26 | 110 |
| $n = 100\,000$ | | | | | |
| 120 | 0 | 17 | 1.0 | 53 | 107 |
| | | 18 | 1.0 | 36 | 50 |
| 100 | 20 | 19 | 0.92 | 15 | 144 |
| | | 17 | 1.0 | 23 | 109 |
| 60 | 0 | 11 | 1.0 | 19 | 204 |

### 3.1.5. *General summary of recall performance*

Table 8 shows the statistical thresholds for recovery of a simple representation in $\mathscr{P}_3$ and tables 9 and 10 can be used to discover the minimal conditions on an input for it eventually to cause the recovery of such a representation. The memory consists of 1.25 million input fibres, divided into 25 blocks of 50 000 fibres. A single input event causes activity in 2500 fibres—100 in each block—and the simple representation of each event is formed. Suppose each $\mathscr{P}_3$-cell has 20 000 afferent collateral synapses. After 50 000 events have been learned, recovery of an event $E_0$ will have very

high probability of success from stimulation of 30 fibres, all of which were active in $E_0$, provided that those fibres belong to one block; or from stimulation of 100 fibres in one block, provided that about 70 of those fibres were active in $E_0$. After 100000 events have been learned, the corresponding figures are 60, and 90 out of 100, still from a single block.

TABLE 10. ADDRESSING ONE BLOCK OF $\mathscr{P}_2$ WITH AN INPUT, FROM ONE BLOCK OF $\mathscr{P}_1$, WHICH CONTAINS A SUBEVENT OF A LEARNT EVENT $E_0$

The $\mathscr{P}_2$-representation of the part of $E_0$ in this block occupied 121 cells of $\mathscr{P}_2$; $n$ such events have been stored. Notation is from the text (see figure 6).

| $A_0$ | $A_1$ | $T_2$ | $f_2$ | $B_0$ | $B_1$ |
|---|---|---|---|---|---|
| $n = 50000$ | | | | | |
| 20 | 0 | 7 | 1.0 | 57 | 50 |
| | 0 | 8 | 1.0 | 35 | 12 |
| 30 | 0 | 9 | 1.0 | 80 | 26 |
| | 0 | 10 | 1.0 | 61 | 8 |
| 40 | 0 | 11 | 1.0 | 94 | 12 |
| | 0 | 12 | 1.0 | 80 | 4 |
| 80 | 20 | 23 | 0.9 | 94 | 5 |
| | | 24 | 0.9 | 89 | 2 |
| 60 | 40 | 23 | 0.8 | 63 | 27 |
| | | 24 | 0.8 | 53 | 14 |
| $n = 100000$ | | | | | |
| 30 | 0 | 11 | 1.0 | 43 | 84 |
| | 0 | 12 | 1.0 | 27 | 27 |
| 40 | 0 | 13 | 1.0 | 64 | 101 |
| | 0 | 14 | 1.0 | 48 | 39 |
| 50 | 0 | 16 | 1.0 | 67 | 45 |
| | 0 | 17 | 1.0 | 52 | 18 |
| 80 | 20 | 28 | 0.9 | 73 | 79 |
| | | 29 | 0.9 | 62 | 39 |

## 3.2. *Justifying the model of* § 3.1

### 3.2.0. *Generalities*

There are three general constraints which are important in determining the general structure of the memory of § 3.1. They are

(i) that the memory should consist of a number of layers of cells, each receiving connexions from one layer and projecting to one other;

(ii) that the memory needs a capacity, $n$, of the order of $10^5$ events, with good recall capabilities and about $10^6$ input fibres;

(iii) that recall should be complete before the projection out of the memory.

The constraint (i) arises because the theory is devised for certain regions of the brain which, according to the available evidence, are connected in this way (see § 4). A theoretician has two general options when designing a memory: he can either specify an exact task, and prove that a particular model is the most economical for that task (cf. Brindley 1969); or he can describe an exact structure, and compute its performance (see, for example, Marr 1969). The present theory has the disadvantage of no exact information; its task is the relating of previously unrelated pieces of knowledge by deduction from plausible general assumptions, the whole being tested by the predictions to which it leads. Condition (i) represents the injection of existing anatomical information into the theory.

Constraint (ii) is important in so far as the design of the memory would have to be changed if

it were shown that the figure of $10^5$ was too low. If it were too high, the memory would need only to be shrunk; but a collateral effect is not possible where $n\alpha^2$ is much larger than 1.

It is a matter of common experience that few people can memorize more than 100 randomly chosen items in an hour, though the items may not correspond to the technical term 'event' since many are temporally extended. Even supposing each such item to correspond to 10 events, only 1000 events would need to be stored every hour. This would give 16000 in a 16 h day, which would allow a reasonable number of full days to be accommodated. This seems sufficient for a memory which, it is proposed, is only for temporary storage (information being transferred to the neocortex at least in part during sleep). There is therefore not much danger that $10^5$ is an underestimate for $n$.

The third constraint—that recall should be completed before the return projection—may be justified in two ways. If it is assumed that the return from the memory should occupy as few neocortical synapses as possible, then the return projection must be used only for addressing the neocortical pyramids. There will then be no spare capacity for noise elimination there, and so recall has to be complete before this stage. The second point is that the number of events that may be learned by a single cell is about 100 (§1.2.4). Hence if any neocortical pyramid is likely to be active in such a number of learnt events, all its afferent synapses from the memory will be occupied by the addressing problem. In this case also, there will be no spare redundancy for noise elimination.

These two arguments suggest, but do not compel, the view that the final efferent projection from the memory should perform little more than an addressing task. Constraint (iii) is therefore assumed; but it should be remembered that any spare capacity on the return projection would allow the memory to be correspondingly over-run in its earlier stages.

### 3.2.1. *The form of the simple representation*

It was shown in §1.4.2 that a model consisting of only one layer of cells (input $\mathscr{P}_1 \rightarrow \mathscr{P}_2 \rightarrow$ return) cannot be constructed to satisfy the general constraints set out in §1. In §3.1, it was shown that a memory with two intermediate layers ($\mathscr{P}_1 \rightarrow \mathscr{P}_2 \rightarrow \mathscr{P}_3 \rightarrow$ return) can. This section discusses how the specifications for $\mathscr{P}_3$ could differ from those of §3.1.

A collateral effect can only be operated usefully among the cells of $\mathscr{P}_3$ if $n\alpha_3^2 \lesssim 1$, i.e. $\alpha_3 \lesssim 0.003$. In order that $\alpha_3$ be this low, the number of cells in $\mathscr{P}_3$ must exceed 30000, since otherwise the number of active cells in $\mathscr{P}_3$ becomes unrealistically low. $N_3$ could be say 50000, but the chosen figure was 100000, since this allows a slightly lower $\alpha_3$ while remaining plausible.

Provided therefore that the need for a collateral effect in $\mathscr{P}_3$ is accepted, $N_3$ and $\alpha_3$ must be roughly as in §3.1. If there were no collateral effect in $\mathscr{P}_3$, the constraint that recall has to be complete by then implies that at least one of the projections into $\mathscr{P}_2$ and into $\mathscr{P}_3$ must have low values of $\Pi$; i.e. the probability, that an arbitrary modifiable afferent synapse to $\mathscr{P}_2$ or $\mathscr{P}_3$ has been modified, must be low. Hence, either $n\alpha_1\alpha_2 \ll 1$ or $n\alpha_2\alpha_3 \ll 1$. If recall is to be allowed from one block of $\mathscr{P}_2$, $\Pi_3$ must be low, and so $n\alpha_2\alpha_3 \ll 1$. Other things being equal, if $\Pi_3$ has to be so low that recall is achieved almost totally in $\mathscr{P}_3$ from one block in $\mathscr{P}_2$, $\alpha_3$ has to be less than it is in the model of §3.1 and thus a collateral effect is possible in $\mathscr{P}_3$.

The arguments are therefore strongly in favour of the form of simple representation shown in §3.1. The memory, if it is anything like that described there, must be rather similar to it. There may of course be other, very different solutions: but the available histological evidence suggests that, for example, the hippocampus is built to a plan along the lines of §3.1 (see §4).

3.2.2. *The specification of* $\mathscr{P}_2$

The block structure in $\mathscr{P}_1$ and $\mathscr{P}_2$ is simply a crude attempt to approximate to an ordering of some kind on the input fibres. The figures chosen have no particular justification: nor does it matter greatly if they are changed.

Once the values of $N_3$, $\alpha_3$ have been chosen by the need to create in $\mathscr{P}_3$ a favourable environment for the collateral effect, the shape of $\mathscr{P}_2$ is roughly determined by the number $S_3$ of synapses allowed for the projection $\mathscr{P}_2$ to $\mathscr{P}_3$. The best use of $\mathscr{P}_3$ requires that $\Pi_3$ lies between about 0.2 and 0.8; if $\alpha_3$ and $N_3$ are fixed, this roughly determines the number of active afferent fibres that each active cell of $\mathscr{P}_3$ should possess. This determines the relation between $L_2$ and $Z_3$, choice of one of these remaining. The final condition, which roughly decides $L_2$ (and hence $Z_3$) is the condition that each active afferent to $\mathscr{P}_3$ is received at an active cell of $\mathscr{P}_3$. This fixes an upper bound to $L_2$ near which (by economy arguments) $L_2$ should actually be found. The value of $L_2$ in the model of § 3.1 is 3000, but values up to about 6000 are acceptable, provided slight changes elsewhere are made.

3.2.3. *Input to* $\mathscr{P}_2$

Once $L_2$ has been roughly decided, the other parameters of $\mathscr{P}_2$ are determined by $n$ (the capacity), and by the input from $\mathscr{P}_1$. For modifiable synapses to be useful in $\mathscr{P}_2$, $\alpha_2$ must be less than 0.01, and recall performance is much impaired if $\mathscr{P}_2$ does not contain modifiable synapses. This constraint on $\alpha_2$, together with the rough estimate for $L_2$, decides $N_2$. The only remaining numbers are $L_1, S_2, Z_2$; and the only freedom here is in the choice of $S_2$, since the conditions (i) $n\alpha_1\alpha_2 \lesssim 1$ and (ii) that $L_1$ is fully represented in $\mathscr{P}_2$, decide $L_1$ given $S_2$. The model of § 3.1 chooses $S_2 = 10000$, giving $L_1 = 100$ per block. $S_2 = 20000$ would allow $L_1 = 200$ per block, but if $L_1$ is in fact substantially larger than 100, it will be necessary to interpose another layer between the $\mathscr{P}_1$ and the $\mathscr{P}_2$ of § 3.1. (The anatomy of the hippocampal formation suggests that, in the most direct application of this theory, an extra layer of this kind is actually present.)

The general conclusion from the arguments outlined here is that, provided $L_1$ and $N_1$ are roughly as in § 3.1, the rest of the memory will have roughly the prescribed dimensions. The specifications of § 3.1 can be changed, and the general equations of § 2 provide rough guides to the consequences of such changes. If $L_1$ is actually much larger than the value suggested, an extra layer is necessary to transform it into a signal which is acceptable to $\mathscr{P}_2$. Detailed calculations must await the discovery of some quantitative anatomical information.

### 3.3. *Remarks concerning threshold setting*

3.3.1. *Subtraction and division*

The computations of § 3.1 assumed that inhibition is capable of division and of subtraction. It was proposed by Marr (1970, § 4) that inhibition applied to pyramidal cell dendrites will be subtractive in effect, but that inhibition concentrated at a soma is capable of performing a division. Neither function has been demonstrated to occur.

The model (§ 3.1) does not depend upon the ability to set both a subtraction and a division threshold, but its performance is impaired if only one of these is allowed. If only subtraction is allowed, equations S 5 of §§ 3.1.3 and 3.1.4 become

$$R_i = T_i \quad (i = 3, 2 \text{ respectively}).$$

If only division is allowed, they become

$$R_i = sf_i \quad (i = 3, 2 \text{ respectively}).$$

The equations for the projection $\mathscr{P}_1 \rightarrow \mathscr{P}_2$ have been recomputed for the cases where only a subtraction or only a division function is allowed, and the results appear in table 11. It will be seen that the results, especially for division alone, are much inferior to those when both are allowed.

TABLE 11. COMPARISON OF PERFORMANCE USING PURE SUBTRACTION AND PURE DIVISION THRESHOLDS WITH PERFORMANCE USING A COMBINATION OF THE TWO

Figures are for one block of $\mathscr{P}_1 \rightarrow \mathscr{P}_2$ as in tables 1 and 2. $T$ denotes the subtraction threshold; $f$, the division threshold. 50 000 events have been stored. * denotes no solutions involving between 10 and 1000 active cells. $A_i$, $B_i$ as in text, and figure 6.

| | subtraction | | division | | combination | |
|---|---|---|---|---|---|---|
| input $A_0/A_1$ | $T$ | $B_0/B_1$ | $f$ | $B_0/B_1$ | $(T, f)$ | $B_0/B_1$ |
| 10/0 | | * | | * | (4, 1.0) | 49/354 |
| | | | | | (5, 1.0) | 23/95 |
| | | | | | (6, 1.0) | 8/48 |
| 30/0 | 9 | 80/169 | | * | (9, 1.0) | 80/26 |
| | 10 | 61/48 | | | (10, 1.0) | 61/8 |
| | 11 | 43/12 | | | (11, 1.0) | 43/2 |
| 50/0 | 13 | 104/132 | 1.0 | 121/393 | (13, 1.0) | 104/5 |
| | 14 | 94/47 | | | (14, 1.0) | 94/2 |
| | 15 | 81/15 | | | (15, 1.0) | 81/1 |
| | 16 | 67/5 | | | (16, 1.0) | 67/0 |

3.3.2. *Changing $f$ during recall*

It can be seen from tables 5 to 7 that during the recovery of a simple representation by the collateral effect, best results are obtained if $f$ is raised for each new cycle. In the simple model which was used to make the computations, recovery, if it happens at all, will take place within about three cycles—that is, three successive applications of the collateral effect. In a physiological memory of this type, the cycles as such will not exist in this discrete sense: recovery will be a smooth process. But it will happen quickly, if at all, and will proceed best if $f$ is increased gradually throughout it. The fact that recovery will occur so quickly means that the 'program' for increasing $f$ can without undue loss be the same for all inputs. (This would, for example, not have been so if borderline cases had tended to spend a large number of cycles near the borderline, since $f$ would then sometimes have had to be held for some time at (say) 0.3.)

In physiological terms, this means that the proportion of basket cell inhibition to inhibition applied to the $\mathscr{P}_3$-cell dendrites should initially take some small value—say corresponding to a value $f \approx 0.3$—and should be raised during recall until $f$ is near 1.0. This increase can take place at the same rate and from the same initial value for all recall problems. The likely time-course of the change is of the order of 0.25 s, allowing 50 to 100 ms for each cycle, and the whole operation must be carried out subject to the (negative feedback) condition that a roughly constant number of $\mathscr{P}_3$-cells is kept active. There are various methods by which this could be done, though I can find no single one which seems to be particularly preferable to the others. One method, for example, is to employ an external agency which gradually increases basket cell activity in $\mathscr{P}_3$. The subtractive inhibitory level is then set at an appropriate level by the usual negative feedback through $\mathscr{P}_3$-cell collaterals and an inhibitory interneuron (the $G$-cells).

### 3.4. *The return from the memory*

The analysis of the projection back to the neocortical pyramidal cells is straightforward. If, say, each pyramid devotes 10 000 synapses to the memory, an expected 22 will be active in each

learnt event. These synapses need to be Hebb modifiable synapses, facilitated by simultaneous pre- and post-synaptic activity. Inhibition needs to be applied to these dendrites so that the cells fire only when all their active afferent return synapses have been modified. In view of the small number active, they probably need to be close together, and perhaps a little larger than other synapses.

### 3.5. *Scanning during recall*

Simple memory was originally suggested by the need for a direct form of storage which would enable common subevents to be discovered. Addressing the memory with a subevent will cause events to be recalled that contained most of the addressing subevent. Whole events presented to the memory are unlikely to cause recall of other whole events, since any two events will probably differ substantially.

It therefore appears that to use the memory for storage, whole events should be presented to it. Using it for recall requires that subevents should address it, which in turn implies some categorization of the current internal description even at this early stage. The notion that, in order for recall to take place, only a small part of the current internal description should have access to the memory, is close to an idea of attention.

The two problems raised by this are, first, how are common subevents picked out; and secondly, how are they copied out of the memory during the codon formation for new classificatory units? The first problem is the partition problem (Marr 1970, § 1.3.3). Simple memory shows how this problem can be approached, since the ability of a subevent to pick out a related event despite a fair amount of noise shows that test subevents do not have to be all that accurately chosen. Rather general, and perhaps innate, techniques for scanning the current internal description will lead to the discovery of many subevent clusters. The scanning process itself may well be subject to neocortical control. The teaching of scanning techniques—how to 'look' at things—may be a very important factor in the development of a child, since it will have a great influence on the classificatory units that the child will form.

The second problem is more technical and easier to give some kind of answer to. Presumably, when a subevent causes recall of a previous event, it is 'marked' in some way— that is associated (in the technical sense) with a 'marker' input from some special centre. This centre also has a measure of the 'importance' to the organism of this kind of information. When a subevent cluster of sufficient size and importance has been formed, this centre will (perhaps during sleep) call the information out from the memory during a period when codon formation is possible. This can be done simply by addressing the memory with the marker event. The markers have to be fairly simple stereotyped inputs, which can be reproduced when required, and which call up (by association) the subevents that they mark. The obvious candidates for 'marker' inputs, in view of the 'importance' parameter necessary for this function, are the rather primitive firing configurations which may perhaps be associated with the subjective experience of a fairly strong emotion.

The problems outlined in this section will form the subject of a later paper.

## 4. A THEORY OF HIPPOCAMPAL CORTEX

### 4.0. *Introduction*

In this section is presented the analysis of hippocampal cortex that follows from its interpretation as a region in which the simple representations of many events are formed. The discussion is restricted to the consideration of local properties of the cortex of various parts of the

hippocampal formation, and includes a brief classification of cortical cell types, based on the results of this paper and of Marr (1970). An interpretation of the macroscopic intrinsic and extrinsic connexions of the hippocampal formation will appear in the paper on hippocampal input–output relations.

### 4.1. *The morphology of the hippocampal formation*

#### 4.1.0. *Gross morphology*

Most of the following description of the structure of the hippocampal formation is derived from information about the mouse (Cajal 1911; Lorente de No 1934) and the rat (Blackstad 1956; White 1959). There is, however, a remarkable uniformity in the structure of the hippocampus in mammals (Lorente de No 1934), so that the divisions made in the mouse are easily recognizable in man. The only important histological difference is in the size of the elements involved: man's hippocampus is larger in every way than that of the mouse. The homology of the afferent and efferent paths in the two species is less good, since the many slight differences in the sizes of the relevant tracts combine to give overall pictures which are considerably different. Those aspects of the present theory which relate only to histology may however be applied to the hippocampal cortex of most mammals.

Blackstad (1956) and White (1959) have recently used morphological information to classify the various regions of the hippocampal region in the rat. Their findings agree closely, and the present paper will usually follow the terminology of Blackstad. According to that author, the hippocampus admits of the following subdivisions:

| | | |
|---|---|---|
| (1) area entorhinalis | (a.e.) | |
| (2) parasubiculum | | |
| (3) presubiculum | (pres.) | |
| (4) area retrosplenialis e | | |
| (5) subiculum | (sub.) | |
| (6) cornu ammonis | (CA) | (the hippocampus proper), |
| was divided into | CA 1 | |
| | CA 2 | |
| | CA 3 | |
| | CA 4 | by Lorente de No (1934) |
| (7) fascia dentata | (FD) | |

The division is illustrated in figure 7: Blackstad (1956) gives the explicit criteria for distinguishing the borders between the different regions (1) to (7). Regions (6) and (7) are those most characteristic of the hippocampal formation. The subdivision of (6) CA into CA 1 to CA 4 is based on variations in the structure of the hippocampal pyramids. CA 4 is in many ways distinct from the rest of the CA, and it will be discussed separately in § 4.4, together with the FD (7).

#### 4.1.1. *The histology of the cornu ammonis (CA)*

CA is composed principally of a layer of large pyramidal cells, whose axons constitute the efferent tracts from the hippocampus. Many of these cells are extremely large, and their dendritic trees usually span the whole thickness of the CA. They are arranged in a particularly neat row, and it is the bodies of these cells which give the hippocampus its characteristic appearance. Figure 8 illustrates their arrangement in the cortex.

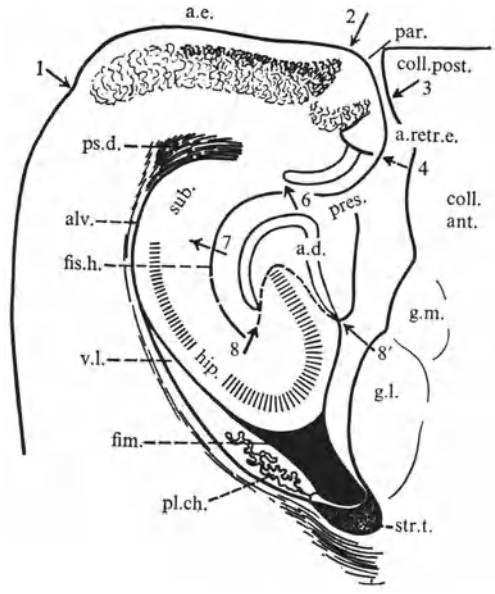



FIGURE 7. Diagram of the hippocampal region in the rat, based on horizontal silver-impregnated sections. The posterior end of the hemisphere is at the top of the figure, the medial side at the right. Arrows show the limits between the areas, which are abbreviated as follows: parasubiculum (par.), presubiculum (pres.), subiculum (sub.), hippocampus (hip.), fascia (area) dentata (a.d.). Other structures shown are ps.d. dorsal psalterium, alv. alveus, fis.h. fissura hippocampi, v.l. lateral ventricle, fim. fimbria, pl.ch. choroid plexus, str.t. stria terminalis, g.l. lateral geniculate body, g.m. medial geniculate body, coll. ant. and post. the anterior and posterior colliculus, and a.retr.e. area retrosplenialis e. (Fig. 2 of Blackstad 1956.)

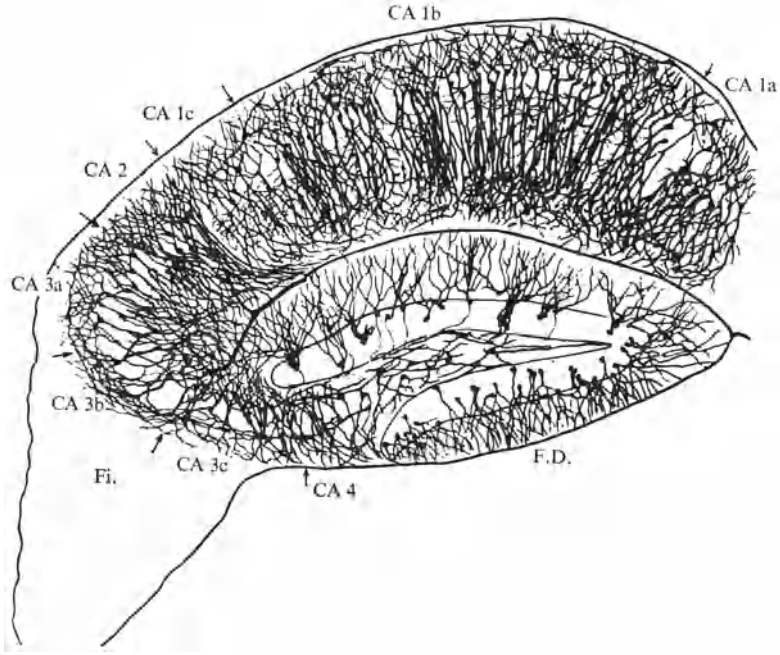



FIGURE 8. Longitudinal section of the adult mouse brain, Cox method. Fi is the fimbria: the divisions are those of Lorente de No (his Fig. 5, 1934).

Hippocampal cortex is commonly regarded as having the four layers shown in figure 9. The bodies of the hippocampal pyramidal cells lie in the Stratum Pyramidale (S. Pyr.), and their basal dendrites span the Stratum Oriens (S. Oriens). Their apical dendrites rise through the Stratum Radiatum (S. Rad.), where they may split into two or more shafts, and arborize freely in the Stratum Moleculare (S. Molec.). The region between the S. Rad. and S. Molec. is often called the Stratum Lacunosum (S. Lac.). Lorente de No (1934) combined information from his own studies with that obtained by Cajal (1911) and earlier authors to give the following description of the cell types in these layers.

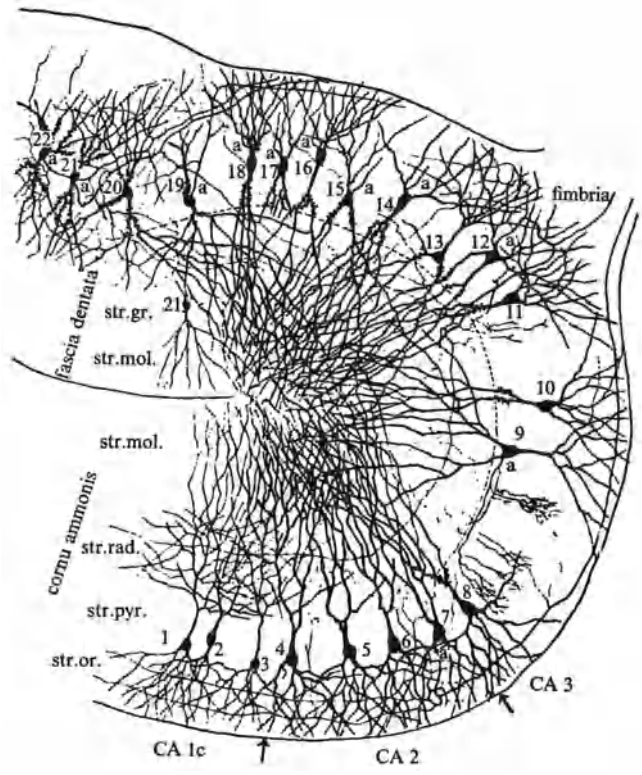



FIGURE 9. Types of pyramids in fields CA 1, CA 2, CA 3, CA 4. 1 to 3 are pyramids of CA 1; 4 to 7 of CA 2; 9 a pyramidal basket cell of CA 3. Only axons of cells 12, 19, 21, 22 have been included in the drawing. Twelve-day-old mouse, Golgi method. (Lorente de No 1934, Fig. 9.)

*Stratum Pyramidale*

(*a*) *Pyramidal cells.* These vary slightly in appearance from region to region, but figure 9 illustrates their basic uniformity. All pyramidal cells of this class send an axon out of the hippocampus. Those in CA 4 have a modified form, which is explained later.

(*b*) *Pyramidal basket cells.* Their bodies and dendrites are similar to those of the pyramidal cells, but their axons are completely different: they travel horizontally and form baskets round the somas of the pyramidal cells (cell 9, figure 9). There are no basket cells in CA 4, and those in CA 3 do not receive synapses from the so-called mossy fibres (i.e. axons of the granule cells of the FD).

(*c*) *Cells with ascending axon.* Their bodies and descending dendrites are similar to those of the

pyramidal cells, but the ascending dendrites leave the soma directly, and are not branches off a single shaft. The axon arborizes chiefly in S. Rad. (cell 1 of figure 10).

*Stratum Oriens*

(*d*) *Horizontal cells with ascending axons* have dendrites which remain in S. Oriens: the axons ascend to S. Molec. and arborize there (cell a, figure 12).

(*e*) *Polygonal cells with ascending axons* are similar to (*d*) except in two respects: their axons sometimes emit collaterals in S. Rad., and they send a dendrite to Ss. Rad. and Molec. (cell 5 of figure 11).


FIGURE 10

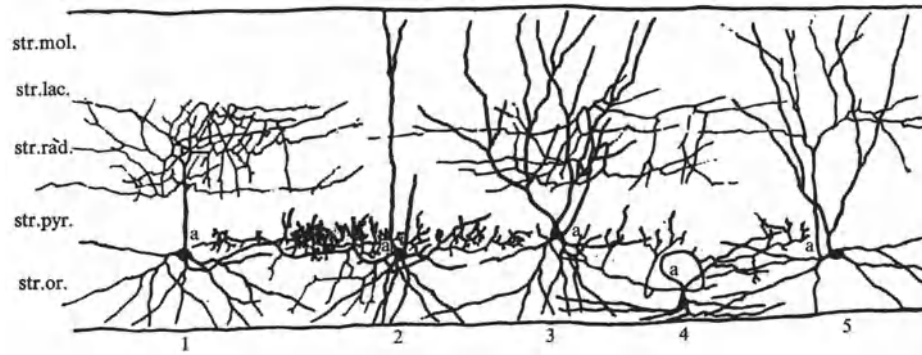



FIGURE 11

FIGURE 10. Types of cell with short axon in CA 1. Twelve-day-old mouse, Golgi method. (Lorente de No 1934, Fig. 7.)
FIGURE 11. Types of cell with short axon in CA 1. Twelve-day-old mouse, Golgi method. (Lorente de No 1934, Fig. 8.)

(*f*), (*g*) *Basket cells* are of two types, one with horizontal dendrites remaining in S. Oriens, and one with a dendrite ascending to S. Molec. (cells 4 of figure 10, 2 of figure 11, and b of figure 12).

(*h*) *Horizontal cells with axon in S. Rad.*, whose dendrites remain in S. Oriens (cell 1 of figure 11).

(*i*) *Horizontal cells with horizontal axon* are globular with dendrites remaining in S. Oriens, and axons ramifying in S. Oriens and occasionally also in S. Pyr. (cells 2 and 5 of figure 10, cell 4 of figure 11).

*Strata Radiatum and Lacunosum*

Cajal (1911) described the S. Lac. separately in the rabbit, where the Schaffer collaterals are especially distinctly grouped; but in the mouse, cat, dog, monkey and in man, the S. Rad. and S. Lac. are not obviously distinct (Lorente de No 1934). They contain the following types of cell:

(*j*) *Cells with axon ramified in S. Rad.*, of which there are four types, being all combinations of

two kinds of dendritic and two axonal distributions. Some dendrites reach S. Mol., others remain in S. Rad. and S. Lac.; some axons ramify only in S. Rad. and S. Lac., others give branches to S. Pyr. (e.g. cells 3, 6, 7 of figure 10).

(*k*) *Cells with ascending axon ramified in S. Mol.*, after branching in S. Rad. and S. Lac. The dendrites ramify in Ss. Lac., Rad., Pyr. and even Oriens (cells e to m of figure 12).

(*l*) *Horizontal cells of S. Lac.* have axonal and dendritic distributions both in S. Lac., the region of the Schaffer collaterals (see below) (cell 3 of figure 10).

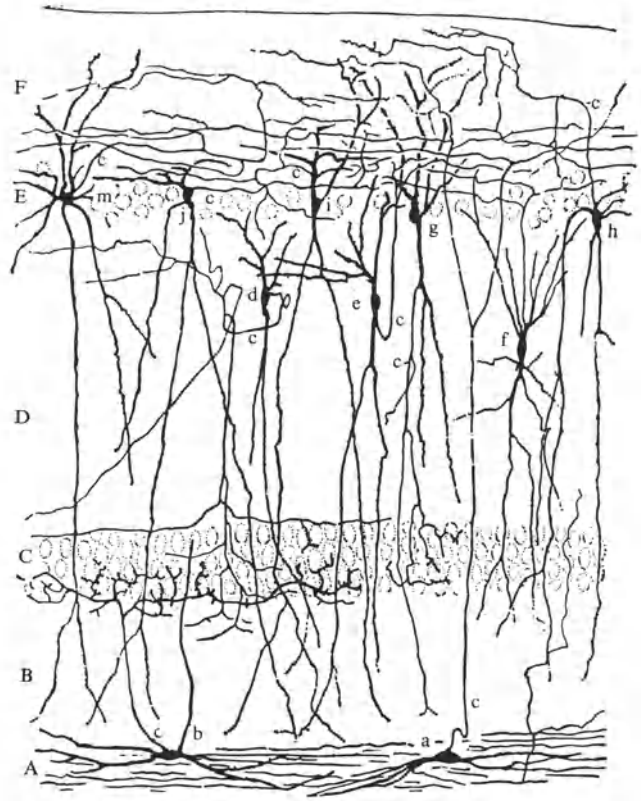



FIGURE 12. Various short-axon cells of the CA. Six-day-old rabbit, double-silver chromate method. (Cajal 1911, Fig. 476.)

*Stratum Moleculare*

The S. Molec. contains several cells with short axon, typical of a cortical molecular layer.

(*m*) *Cells with short axon*, and

(*n*) *Horizontal cells*,

both of which seem to be rather difficult to stain.

### 4.1.2. *The histology of the fascia dentata* (*FD*)

Cajal (1911) gave a full description of FD, which he divided into three layers, the molecular, granular, and polymorph layers. The most notable elements of the cortex are the granule cells, whose bodies, like those of the hippocampal pyramids, are neatly packed and arranged in a granular layer (see figure 13). These cells have supporting cells analogous to those found in CA: they are described on the next page.

*Molecular layer*

(*a*) *Displaced granule cells* look, and will be treated, like granular cells displaced a little into the molecular layer (cell a, figure 13).

(*b*) *Short-axon cells*, of which there are two main types. The more superficial (figure 14, f and g) have delicate dendrites, mostly horizontal or descending. Their axons are extremely thin and terminate locally, in the outer part of the molecular layer, with a considerable ramification.

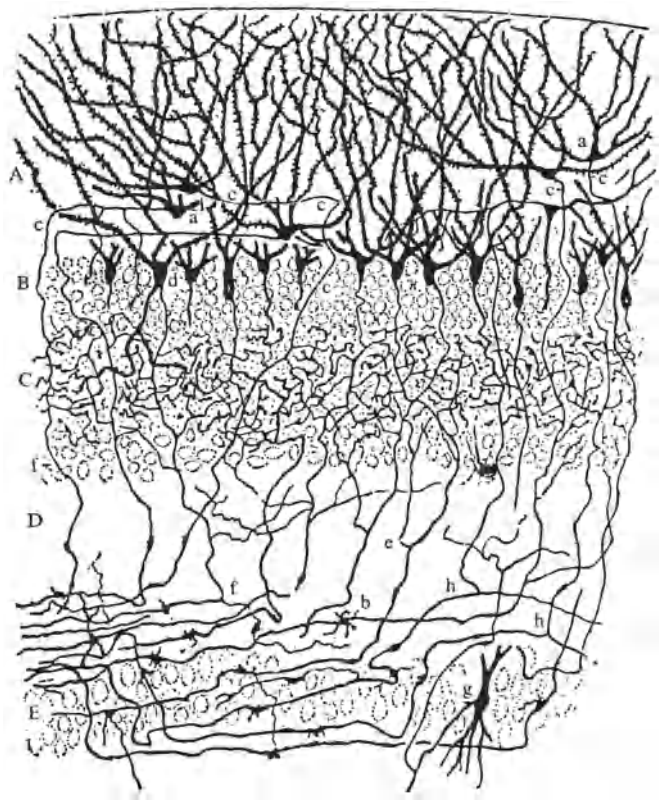



FIGURE 13. The FD in the region of the hilus of the CA. One-month-old guinea-pig, Golgi method. (Cajal 1911, Fig. 478.)

The deeper cells are larger, and occupy the lower portion of the layer (figure 14e). They possess dendrites which spread and divide in all directions—even crossing the granule layer to reach the polymorph layer. Their axons are larger than those of the more superficial cells; they arborize freely in different directions, while remaining in their original layer.

*Granular layer*

Cajal (1911) regarded the granule cells of the FD as a variant of the cortical pyramidal cells. Figure 13 contains many examples: it will be seen that they lack basilar dendrites, and send about four or five dendrites up through the molecular layer. Their axons are thin, and become the so-called mossy fibres of CA 4 (see below). As they cross the polymorph layer, they give off four or five collaterals, which terminate there. These axons hardly ever give out collaterals after they have crossed the polymorph layer.

*Polymorph layer*

(*c*) *Pyramidal cells with ascending axon* (figure 15). These cells possess basilar dendrites, which give them a pyramidal shape. Their apical dendrites rise in the manner shown, and their axons eventually ramify horizontally into the granular layer. The cells have obvious similarities with the pyramidal basket cells of the hippocampus proper. Occasionally, but rarely, pyramidal cells are seen that send their axon to the alveus.

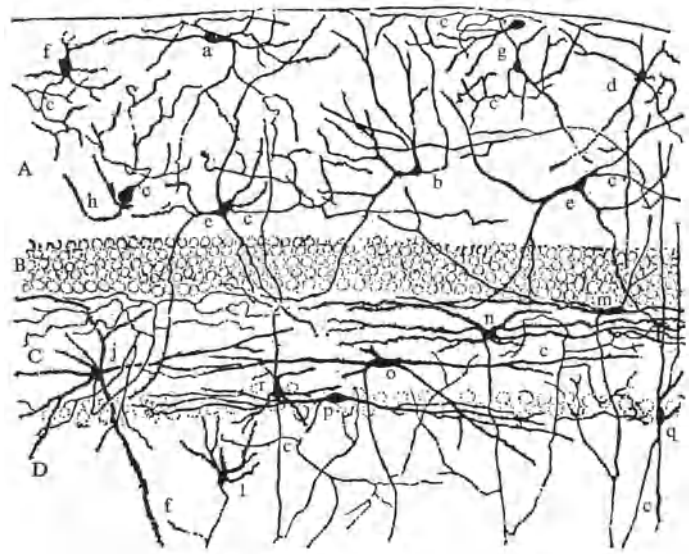



FIGURE 14. The FD. One-month old rabbit, Cox method. (Cajal 1911, Fig. 477.)

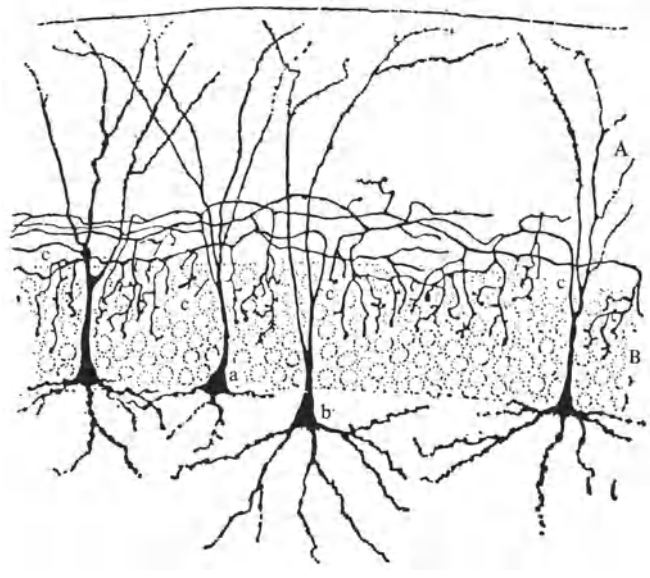



FIGURE 15. The FD. One-month-old rabbit, Cox method. (Cajal 1911, Fig. 480.)

(*d*) *Cells with ascending axon,* which crosses the granular layer and ramifies horizontally. They have various kinds of dendritic distribution (figure 16, e and f; i and o are basket cells).

(*e*) *Cells with descending axon* have long horizontal dendrites which never cross the granular layer. Their axons become fibres in the alveus (figure 16g, j).

(*f*) *Short-axon cells* with local axonal and dendritic distributions: they are found throughout the lower part of this layer (figure 15h).

(*g*) Various star and fusiform cells found low in this layer send their axons eventually to the alveus.

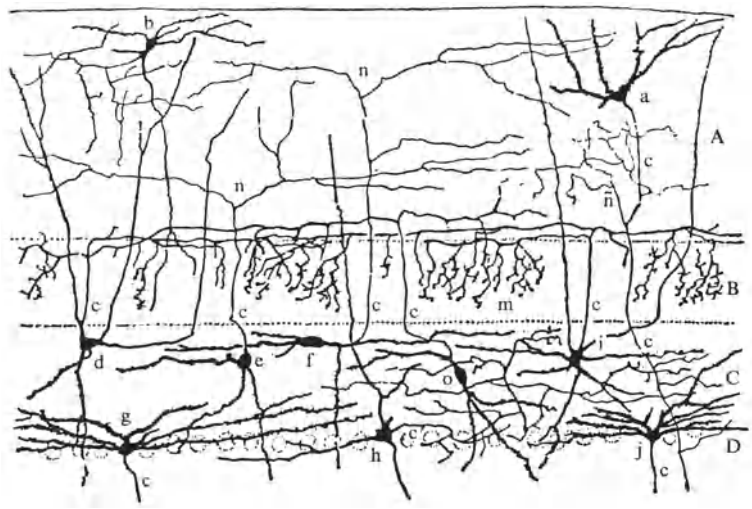



FIGURE 16. The FD. Eight-day-old rabbit, Golgi method. (Cajal 1911, Fig. 481.)

4.1.3. *The principal association systems of the hippocampus*

The present investigation will not concern itself with the relationship between the hippocampi of the two sides of one animal, and consequently little information about the various highly organised commissural connexions will be required (see § 4.5.2). There are four principal systems for association in the hippocampus, and they are dealt with separately.

(i) *The mossy fibres.* The FD granule cell axons become the mossy fibres of the hippocampus. These axons run from FD along CA 4 and CA 3 near the pyramidal cell bodies. They synapse with the dendritic shafts in these regions, producing the distinctive thorns which show up so well in Golgi preparations (figure 9) (Cajal 1911). Few if any penetrate beyond the boundary between CA 3 and CA 2. There are two crucial points to note about these fibres: first, they form the only efferent pathway for the dentate granule cells; and secondly, they specifically avoid the pyramidal basket cells of the hippocampus. These cells thus lack the characteristic thorns (Lorente de No 1934). In CA 4, mossy fibres form the main source of afferent synapses with the pyramidal cells there, and CA 4 contains no basket cells.

(ii) *The Schaffer collaterals* are thick collaterals of the pyramidal cells in CA 3 and CA 4. They travel away from the dentate fascia, and rise through S. Rad. as they go. They synapse in S. Lac. with the pyramidal cells of CA 2 and CA 1 (Schaffer 1892).

(iii) *The axon collaterals of CA* 1 *and CA* 2. The Schaffer collaterals are a transverse association system, joining CA 3 and 4 to CA 1 and 2. CA 1 and 2 also possess a predominantly longitudinal

association system consisting of collaterals synapsing with pyramidal cells in S. Lac.-Molec. These join CA 1 and 2 with other parts of CA 1 and 2. The associations stay more local in CA 1 than in CA 2, but are clear in both cases (Raisman, Cowan & Powell 1965, in the rat).

(iv) *Local associational paths.* It is evident from the descriptions of Cajal (1911) and of Lorente de No (1934) that most hippocampal pyramidal cells have axons which give off collaterals. These probably end locally if they do not contribute to (ii) or (iii), but they have not yet been studied closely. It is necessary therefore to bear in mind that, at least on a local level, the hippocampus is provided with an extremely rich system of interconnexions. It seems to be a general rule in the hippocampus and dentate fascia that different afferent systems terminate both in specific regions and in specific layers of the cortex, not by a random ramification (Blackstad 1956; Raisman *et al.* 1965).

The hippocampal pyramidal cells are extremely large, and so are likely to have at least as many afferent synapses as large pyramidal cells in the motor cortex of the same animal.

### 4.2. *The hippocampal pyramidal cells*

#### 4.2.0. *The basic model*

The pyramidal cells of sections CA 1, CA 2 and CA 3 of the mammalian hippocampus will be regarded as being populations of cells in which simple representations of various input events are formed. It is proposed that these cells are closely analogous to the cells of $\mathscr{P}_3$ in the model proposed in § 2 and analysed in § 3.

The theory of §§ 2 and 3 requires that, if a cell participates in the simple representations set up in a simple memory of about the specified dimensions, it should have the following properties:

- P 1 Its input fibres should be suitable.
- P 2 The activity $\alpha_{CA}$ of the ammonic pyramids should be small: $0.01 \geqslant \alpha_{CA} \geqslant 0.001$ with $\alpha_{CA}$ probably nearer 0.001.
- P 3 Each cell possesses very many ($\gtrsim 50000$) afferent Brindley synapses from the previous layer of cells, and many ($\gtrsim 10000$) Hebb (or Brindley) synapses from other cells of the CA.
- P 4 Synapses from fibres likely to be co-active should be placed near one another.
- P 5 There should exist an extensive collateral system in CA, giving rise to the collateral synapses of P 3, which allow the completion of the simple representations of partially specified input events.
- P 6 There should exist appropriate supporting cells to supply the required inhibition.
- P 7 There should exist a means of clearing information from these cells when it is re-stored—either as associations or as associations or as new classificatory units—in the neocortex.

Points P 2 to P 7 are discussed separately in the following paragraphs: P 1 is dealt with in a later paper.

#### 4.2.1. $\alpha_{CA}$

If the hippocampus is involved in storing information in the proposed way, the number of events it can store depends upon the size of each input event, and upon the number of cells used for each. The smaller is $\alpha_{CA}$, the greater is the capacity, and the more powerful is the collateral effect. $\alpha_{CA}$ is bounded below by about 0.001, a figure which arises out of the necessity to be able to detect those cells which are active (§ 2.3.4). It should not be very difficult to determine $\alpha_{CA}$ by experiment.

4.2.2. *Modifiable synapses*

The competing virtues of the three possible kinds of modifiable synapse of figure 2 have already been discussed. Model 1 was rejected on the ground that each cell would need to be used for more than one input; and the climbing fibre model 3 on the grounds that it needs additional cells, and will not select such suitable cells as model 2 will. It was therefore concluded that model 2, using Brindley synapses, was the preferred choice for all cells in a simple memory. The central feature of Brindley synapses is that they are initially excitatory, and can therefore be used themselves to decide at which cells there should be facilitation (Brindley 1969). This powerful trick solves the problem of selecting the most suitable cells for storing a given input (cf. codon formation, Marr 1970).

There are two practical difficulties associated with the use of Brindley synapses to select CA pyramidal cells for a simple representation. The first arises out of the usual problems associated with a large dendritic tree. It has been pointed out (Marr 1970, § 5.1.4) that in the absence of climbing fibres, it is unreasonable to suppose that synaptic modification is consequent upon simultaneous pre- and post-synaptic activity when these activities are far apart from each other: for example, the spike frequency in an axonal initial segment probably has rather little direct effect upon a synapse 1 mm away at the tip of an apical dendrite of the same cell. Conditions for synaptic modification are therefore likely to hold only locally in a dendrite. This will, however, not be a great disadvantage if input fibres are arranged in such a manner that those that are often coactive tend to lie near one another. It is interesting in this connexion to note that there exists a very marked lamination in the hippocampal afferent system (Blackstad 1956).

The second difficulty is related to the first, and concerns the setting of the thresholds of the CA pyramids. The first time any input is presented to the memory, the appropriate threshold can be computed easily: it is simply a multiple of the power of the unmodifiable component of a Brindley synapse. But after a number of events have been learnt, a non-zero fraction of the CA pyramidal cell afferent synapses will have been facilitated. The thresholds must rise to counteract this effect, and so the amount of inhibition applied to the CA pyramids has to be increased with the number of events that are stored there. Furthermore, if (as seems likely) synaptic modification occurs as a result of a decision process in a local region of dendrite, this inhibition must be applied to such local regions: it is, for example, no use increasing the inhibition at the soma in order to prevent the modification of a synapse at the extremity of an apical dendrite.

The use of Brindley synapses, in output cell selection as well as in codon formation, therefore requires that the amount of inhibition applied to the post-synaptic dendrite, for a given size of input event, should increase with the number of events that the memory has learned. The most satisfactory way of achieving this seems to be to drive the inhibition by collaterals of, in this case, the CA pyramidal cell axons (§ 2 and Marr 1970, § 4.3.1). The cells which achieve this inhibition will be identified in § 4.3.

The conclusion which may be drawn from these arguments, together with those of § 4.3, is that the inhibition level at the CA pyramids can be made to vary in a way which makes it possible for their afferent excitatory synapses to be Brindley synapses. These synapses are in principle the best choice for the function which the present theory assigns to the CA pyramids, and hence the following prediction is made. *Excitatory fibres from the area entorhinalis should terminate on the pyramidal cells of CA* 1 *to CA* 3 *by Brindley synapses.*

4.2.3. *Collateral synapses from other CA pyramids*

The collateral effect (§ 2.4) is an important means by which the simple representation of an incompletely specified input may be completed. The manifestation of this effect in the CA requires that collateral synapses between CA pyramids are modifiable. The synapses included in this discussion are those belonging to reciprocated collateral systems. They do not include either the mossy fibres, or the Schaffer collaterals, both of which are projecting collaterals to which there do not exist reciprocal counterparts: these collateral systems are dealt with in § 4.5.

Collateral synapses should ideally be Hebb synapses: that is, they should initially be ineffective, but should be facilitated by the conjunction of pre- and post-synaptic activity (see § 1.3 for the distinction between Hebb and Brindley synapses). Modification conditions are therefore the same as for the standard CA afferents, except that collateral synapses should probably lack the power to set up modification conditions by themselves.

It is interesting that most collateral synapses to the CA pyramids are found in the S. Rad. (Lorente de No 1934): it seems likely that the importance of the collateral effect is one of the main reasons for the huge development of this part of the dendrite in the CA pyramids. Spencer & Kandel (1961) have shown that the apical dendritic shafts of the CA pyramids can sustain an action potential. It is therefore reasonable to assume that the modification of synapses in S. Rad. could depend on the coincidence of pre-synaptic activity and a burst of post-synaptic action potentials. This would be appropriate on the assumption that decisions about synaptic modification are taken locally in the apical dendritic tree for two reasons. First, spikes will travel at a high rate down through S. Rad. only when that cell is being used to record an input event (though the same activity may lead to the recall of another event): hence post-synaptic depolarization will exist only at the correct times. Secondly, during the recall of an event through the collateral effect, only dendrites in S. Rad. will be exposed to collateral excitation: thus the areas in S. Molec. where the majority of afferents terminate will not be exposed to post-synaptic depolarization, and so inappropriate synaptic modification will not occur there. Both these arguments show that the situation in which the placing of the afferent and collateral synapses was reversed—i.e. where most afferents made synapses in S. Rad.—would be unworkable.

There may be two true reciprocating collateral systems in CA 1 to 3; one distributing its collaterals longitudinally among cells of CA 1 to 2, the relevant fibres rising from S. Oriens and terminating in S. Rad. (Lorente de No 1934); and one being composed of local axon collaterals, many of which distribute in S. Oriens (Lorente de No 1934). Many of the collaterals in the second group will be involved in driving inhibitory threshold controlling cells (§ 4.3). Finally, it must be noted that the associational paths between the hippocampal cortex of each side of the brain must be composed largely of fibres of collateral status. There is evidence that many of these fibres synapse in the contralateral S. Rad. (Lorente de No 1934; Blackstad 1956). (See §§ 4 5.1 and 4.5.2.)

4.2.4. *Numerical predictions*

There are so many unknowns in the equations computed in § 3 that only the most tentative estimates can be made for the expected values of the various parameters. It is probably useful to have some idea of the values compatible with the present form of simple memory theory, if only because if any are shown to be greatly different, it will immediately become clear that others which are related to them must also be different. The following rough values are therefore

given, with the accompanying reservation that they should be regarded only as guides to the orders of magnitude of the various parameters.

(i) $\alpha_{CA}$ is near 0.001.
(ii) $S_{CA} \gtrsim 50000$.
(iii) The number of collateral synapses at a CA pyramidal cell $\gtrsim 10000$.
(iv) $Z_{CA}$, the contact probability of the afferent fibres, is of the order of 0.1.

4.2.5. *Clearing the simple memory*

The final point with which this section deals concerns the role of the CA pyramidal cells in the transfer of information from simple memory to the neocortex.

The alternative ways of losing information from the simple memory are probably either by a gradual decay applied to all information held therein, or by the selective destruction of a simple representation as the information it represents is transferred to the neocortex. Neither method seems particularly satisfactory: the first would mean that the combination of informations acquired at greatly different times more or less requires that the earlier part has been put into neocortex (a store which, if not actually permanent, is imagined to decay with a rather long half-life). The successful combination probably requires that the earlier has since been rehearsed. The second method is more difficult to make convincing, since the nature of simple memory is such that synapses can be involved in the storage of more than one event: hence the cancelling of one trace has the unwanted side effect of weakening the records of a number of other largely unrelated events.

There seem to be no immediate reasons why either mechanism should be preferred to the other, but the first requires what are probably simpler assumptions about the modification conditions at the hippocampal pyramidal cell synapses.

## 4.3. *Short-axon cells in the cornu ammonis*

4.3.0. *Introduction*

According to the present theory, the CA contains no codon cells. It follows that none of the short-axon cells found there are excitatory, and that they carry out all the functions required of inhibitory threshold controlling cells. Hippocampal cortex is in this respect unusual: the cerebellar cortex certainly contains short axon excitatory cells (the granule cells, Eccles, Llinas & Sasaki 1966), and the cerebral neocortex probably does (Martinotti cells, Marr 1970).

4.3.1. *The functions of inhibition*

The present theory requires that the thresholds of the CA pyramids be controlled in a very careful manner. Suppose that synaptic modification is an all-or-none process, and that $p, q$ represent respectively the strengths of the unmodified and modified states of a Brindley synapse, where $0 < p < q \leqslant 1$. Then $[p, q]$ is the analogue of the plausibility range for output cells (Marr 1970, §4.1.3).

The three principal tasks of the pyramidal cell threshold-setting mechanisms are as follows:

T 1. *The storage of events:* when an event $E$ is presented, synaptic modification must take place at those cells which have the greatest number of active afferents.

T 2. *The recognition of subevents:* when a subevent $X$ is presented, those cells must fire which have the greatest fraction $f$ of active afferent modified synapses, provided that the number of such synapses exceeds some number, $T$.

T 3. *The completion of events:* given the firing of a number of hippocampal pyramidal cells, those cells must fire at which the greatest fraction of active afferent collateral synapses have been modified, provided that the number of such synapses exceeds some number $T'$.

These criteria have to be fulfilled without any other instructions, if possible: that is, the mechanism for performing T 1 should naturally perform T 2 when the current input subevent has occurred in a previous event. Collateral synapses tend to lie in S. Rad., where they have their own special inhibitory cells, so T 3 can to some extent be taken separately. The three tasks are discussed below.

### 4.3.2. *The storage of events*

The crucial factor in the storage of events is that the correct conditions for synaptic modification prevail in the pyramidal cell dendrites. Excitation there is due to two components: one, of fixed size, due to the unmodifiable excitatory component of the Brindley synapses; and one, whose size increases with the number of events stored in the memory, due to the fraction of active synapses that have already been facilitated.

The first component is a standard multiple of the number of active afferent fibres, and can reasonably be expected to be counteracted by local inhibitory cells in the hippocampal cortex. The function of these cells is to provide inhibition in the pyramidal cell dendrites such that when no events have been learned, only those dendrites which receive more than a certain number of active synapses are depolarized enough to modify their active afferent synapses. (The necessary number of such synapses is the threshold which appears in table 3.) This inhibition can be provided by cells whose axonal and dendritic distributions are subject to the kinds of sampling techniques outlined by Marr (1969). The obvious candidates for such cells in the hippocampal cortex are the components of cells (*c*) and (*e*) due to their ascending dendrites; cells (*i*) (for this function in S. Oriens); (*j*); (*l*); (*m*); and (*n*) (see §4.1.1).

The second component must increase with the number of events stored in the memory, and again must act on the dendrites of the pyramidal cells, where it must affect the formation of postsynaptic conditions for synaptic modification. It was argued in §2 that the simplest way of achieving this is by having inhibitory cells driven by axon collaterals of the hippocampal pyramids (analogous to the upper dendritic tree of the cerebellar Golgi cells). The following cells of §4.1.1 are interpreted as performing this function: the components of cells (*c*) and (*e*) due to their descending dendrites; (*d*); (*h*); and (*k*). This is an important function for which, fortunately, many of the described cells have appropriate axonal distributions. It remains for electron microscope studies to show whether the dendrites of any of these cells receive synapses from the pyramidal cell axon collaterals.

### 4.3.3. *The recognition of subevents*

It was shown in §§2, 3 that the most sensitive indicator of whether a given cell has previously recorded a subevent similar to the current one is the *fraction* of the active synapses which have been modified. This is computed by a division which Marr (1970, §4.1.6) has argued may be associated with inhibition applied to the soma of a pyramidal cell. The requirement set out in the discussion there of output cell theory was that the amount of inhibition applied to the soma should vary with an estimate of the total number of active fibres: and this is obtained by dendritic sampling by many inhibitory cells, whose synapses converge at the soma. Such cells are for this reason usually

called basket cells, and are present in the hippocampus with suitable axonal distributions (cells (*b*), (*f*) and (*g*) of §4.1.1). Andersen, Eccles & Løyning (1963) have shown that they are inhibitory, but the question of whether they effectively perform a division has not yet been investigated.

The second component of §4.3.2 is also needed for the recognition of learnt subevents.

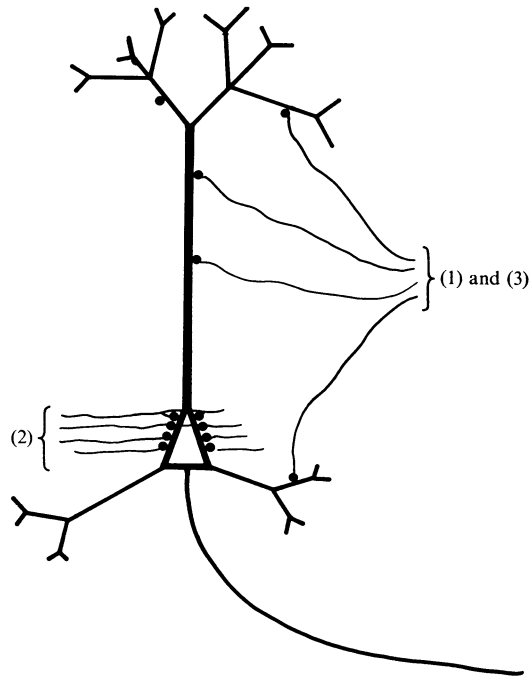



FIGURE 17. Three functions of inhibition: (1) Remove $pK$ where $[p, q]$ is the plausibility range. *S*-cells (i.e. cells *c*, *e*, *i*, *j*, *m*, *n*, of §4.1.1: *l* for the Schaffer collaterals). (2) Divide by $K$ to obtain the fraction $f$ of the active synapses that have been modified. Basket cells (*b*, *f*, *g* of §4.1.1). (3) Raise $p$ to some value $p'$ such that: (*a*) the correct number of cells have outputs in the range $[p', q]$: $p'$ depends on $E$; (*b*) the correct modification conditions are implemented (cells *c*, *d*, *e*, *h*, *k* driven by pyramid collaterals (§4.1.1)).

### 4.3.4. *The completion of a simple representation*

According to §2.4, the principal mechanism available for the completion of a subevent $X$ is the collateral effect, which can recover the simple representation of the event $E \vdash X$ even though $X$ is small (§3.1). For this, collaterals of the pyramidal cells should synapse with other pyramidal cells (in S. Rad.) through Hebb (or Brindley) synapses. Recovery of a simple representation by the collateral effect has been discussed at length in §3.1, where it was seen that best results are achieved if the division threshold (basket inhibition) can be gradually increased during recall. The subtractive inhibition must be decreased in a corresponding way, so as to keep the number of active cells roughly constant.

Subtractive inhibition requires inhibitory synapses applied to S. Rad., and for this the cells (*c*) of §4.1.1 would be suitable. Cells (*h*) have the appropriate dendritic and axonal distributions for the division function. Many of the cell types referred to in §4.3.2, however, have axons ramified

in S. Rad. and S. Lac. as well as in S. Molec. This suggests that synapses in S. Rad. and S. Lac. may also be Brindley synapses, and hence that selection of CA cells depends on their suitability judged from the point of view of the collateral effect as well as of the exogenous afferents.

Although there are various ways by which the proportion of somatic to dendritic inhibition might be changed during recall, the available information does not help one to decide if this is in fact done. One possibility is that the transmitter at basket synapses tends to be degraded rather slowly, causing the effect of these synapses to increase gradually during stimulation. The negative feedback circuit through the other cells would ensure that dendritic inhibition is decreased in an appropriate way.

The three functions performed by the inhibitory cells of the CA are summarized in figure 17; the cells thought to be responsible for each are listed in the legend.

### 4.4. *The fascia dentata*

#### 4.4.0. *Introduction*

The granule cells of the FD will be regarded essentially as extensions to the dendritic trees of the CA pyramidal cells. It is proposed that simple representations are set up in FD in the same way as in CA 1 to CA 3, but that instead of the FD granules sending their own axons elsewhere, they synapse with what may be regarded as 'collector' cells in CA 4 and CA 3. The collector cells send axons elsewhere, and a collateral effect probably operates amongst them.

There are various ideas behind this interpretation of the FD granule cells. The first is that the proposed scheme will result in a saving in the total number of cells transmitting simple representations elsewhere, and hence in savings elsewhere in the numbers of cells and synapses needed to deal with them. It has been seen that the storage capacity for simple representations in a population of cells depends on the activity $\alpha$ of that population; and that $\alpha$ is likely to be bounded below by about 0.001. Hence above a certain point, it is unprofitable to increase the size of the population carrying simple representations, the certain point being in the region of $10^5$ cells. If the amount of afferent information to be dealt with requires more cells than this, something like the proposed theory for the FD becomes the natural scheme to adopt.

The second idea concerns $\alpha_{FD}$, the activity of the FD cells. Once it has become unnecessary for a collateral effect to operate among the cells of a simple representation, the lower bound on $\alpha_{FD}$ ceases to be dictated by the constraint that only about 10000 synapses will be available for the collateral effect. The value of $\alpha_{FD}$ can be pushed down to the bound dictated by the weaker constraints that $\alpha_{FD}$ can be detected by other cells all of whose synapses may be devoted to the task—by the local inhibitory cells, and the proposed collector cells. This notion implies that the collector cells should possess potentially powerful afferent synapses from FD granules, an implication which receives support from the huge size of the mossy fibre synapses in CA. Thirdly, the activity in the population of collector cells must be comparable to that in the rest of CA, so that a collateral effect is possible there.

Finally, it is worth noting that the present theory supports the opinion of Cajal (1911), based on histological evidence, that the dentate granules are a variant of the hippocampal pyramids in CA 3. Lorente de No (1934) remarks (p. 147) that, in the monkey and in man, the similarity between CA and FD is outstanding.

4.4.1. *The FD granule cells*

In the present theory of the FD, essentially the same remarks apply to the granule cells as were made about the CA pyramids, except that there may be no collateral effect amongst them. (It may be replaced by a collateral effect among the cells of CA 4 and CA 3 to which the granules project.) The granule cells (figure 13) are therefore regarded as being like CA pyramids without an S. Rad., S. Lac. or S. Oriens. Their principal afferents from elsewhere should terminate in Brindley synapses: all synapses from local short axon cells should be inhibitory, and should terminate in unmodifiable synapses. The inhibitory synapses on the granule cell dendrites should have a subtractive effect, and those on the soma should perform a division (§4.3 and Marr 1970, §4). The activity $\alpha_{FD}$ should be very small, probably less than $\alpha_{CA}$. Synapses likely to be coactive should be juxtaposed, and the afferent contact probability is probably in the region of 0.1, and may be greater than that found in the CA.

The present theory gives no grounds for supposing that any granule cells should not possess afferent basket synapses (or an equivalent grouping of inhibitory synapses just above the soma). The special cells noticed by Cajal (cell a, figure 16) are therefore not explained by this theory, unless they are found to be inhibitory and to have a local axonal distribution, or to be extremely rare.

4.4.2. *Short-axon cells in the FD*

The requirements for inhibition in the FD are the same as in the CA 1 to CA 3, and the arguments put forward in §4.3 need not be repeated. It remains only to summarize the different functional elements required in the dentate cortex, and to identify them with the cells described by Cajal (1911). The next three headed paragraphs correspond to the sections 4.3.2 to 4.3.4 on short axon cells in the CA.

*The storage of events.* It was seen in §4.3.2 that two components of inhibition are required to ensure that the correct numbers of synapses are modified by an incoming event. The first varies only with the number of active afferent fibres, and is performed by short axon cells with local dendritic fields. Such cells estimate the amount of local afferent fibre activity, and send inhibition to the granule cell dendrites (cells *b* of §4.1.2, including only those parts of the activities of cells e of figure 14 that are due to dendrites in the molecular layer). The second component of inhibition must increase with the number of events stored in the memory. It should be supplied by cells whose axons ramify in the molecular layer, but whose dendrites are exposed mainly to activity in granule cell axon collaterals. The polymorph layer, below the granule cell bodies, receives most of their collaterals: the natural candidates for these inhibitory cells are *b* and some of *d* of §4.1.2.

*The recognition of subevents* requires basket cells and the cells of the last paragraph. Basket cells are present in the FD (cells *c* and others of *d* of §4.1.2).

*The completion of subevents* relies on the collateral effect. Although it is thought that this principally occurs in CA 4 and CA 3, it is worth noting that some FD granule cells do send axon collaterals to the molecular layer of the FD, where the appropriate inhibitory mechanisms are already available.

*Remarks.* The only cells left unaccounted for are certain inhabitants of the polymorph layer (cells, *e*, *f*, *g* of §4.1.2). It seems likely that these cells, found principally in the lower parts of the polymorph layer, should properly be regarded as components of CA 4: the long axon cells as 'collectors' (see later) and those with short axon as the usual inhibitory threshold controlling cells.

There is some evidence (Raisman *et al.* 1965) that septal afferents to the FD terminate in the polymorph layer, though this is not firmly established. If it is true, and if the polymorph cells are largely inhibitory, the finding suggests that the septal nuclei play rather a special role in controlling the FD.

4.4.3. *CA* 3, *CA* 4 *and the mossy fibres*

Lorente de No (1934) described the large cells of CA 4 as modified pyramidal cells. They differ in two major respects from the pyramids of CA 3: first, no basket plexus envelops their somas; and secondly, they receive mossy fibre synapses over much of their dendrites, not (as in CA 3) over small sections of dendrites near the soma.

Since no basket plexus envelops the somas of the CA 4 modified pyramids, it follows that the mossy fibres fail to drive basket cell inhibition at these cells. This interesting characteristic is preserved by the mossy fibres in CA 3, where they conspicuously avoid synapsing with the pyramidal basket cells. No other hippocampal afferents share this feature.

In that part of CA 4 which is closest to FD, almost the whole of the modified pyramids' dendrites seem to be covered with long spines: the number appears to decrease slightly towards CA 3. At the border between CA 3 and CA 4, two things happen: the pyramids suddenly start sending a dendritic stem to the molecular layer of the CA, so the number of their afferent fibres that are not mossy increases sharply; and the basket plexus appears (Lorente de No 1934).

It was proposed in § 4.4.0 that the cells of CA 4 are essentially collector cells for the FD granules, in which an output representation of FD activity is set up and transmitted elsewhere. Thus if mossy fibre synapses are modifiable, they are Brindley synapses, and the setting up process proceeds in the usual way. For this, inhibition is required in CA 4, so that only the correct, small proportion of CA 4 cells is used each time. Short-axon cells of the required kind have been described by Lorente de No (1934, p. 132). The situation is in outline the same as for the ordinary pyramids of CA 1 to CA 3, and the remarks of § 4.3.1 about the setting-up process apply here.

One of the two anomalies concerning the mossy fibres—that they produce very large synapses (Hamlyn 1962) and are not associated with basket inhibition—can be explained by assuming that $\alpha_{FD}$ is extremely low. For this means that $P$(CA 4 & FD), the probability that a (randomly chosen) CA 4 pyramid and an FD granule fire simultaneously, is extremely small—less than $P$(CA 3 & CA 3) for example—and hence that the mossy fibre synapses should be larger than the CA 3 to CA 3 collateral synapses. The fact that the mossy fibres do not drive basket inhibition may mean that these synapses are not modifiable.

### 4.5. *Collaterals and their synapses in the hippocampus*

4.5.1. *Collaterals in the CA*

All hippocampal pyramidal cells send collaterals to S. Oriens (Lorente de No 1934), of which those from CA 2 seem to be the longest. Most give off ascending collaterals which ramify locally in S. Rad., and many also produce a major long-distance collateral to S. Lac. This last category includes the Schaffer collaterals from CA 3 and 4 to CA 1 and 2, and the longitudinal collaterals which arise from cells in CA 1 and 2, and from those cells of CA 3 which have no Schaffer collaterals (Lorente de No 1934).

The collateral effect proper (§ 2.4) is thought to be associated principally with the local axon

collaterals which ramify in S. Oriens and S. Rad. If S. Oriens and S. Molec. are largely independent (a conjecture suggested by their great distance apart), the collateral effects to which each gives rise could be largely independent. Collaterals in S. Oriens are also expected to drive recurrent inhibition (§§ 4.3.2 and 4.3.3).

The long-distance collaterals probably serve another function, analogous to that proposed for the mossy fibres. The axons of the cells of CA 3 and 4 project in the rat to the septal region only; those of CA 1 and 2 project to the anterior thalamus, the mammillary bodies, and to the septum (Raisman, Cowan & Powell 1966). Thus the cells of CA 3 and 4, and hence also of FD, have access to the mammillary bodies and the anterior thalamic nuclei only through the Schaffer collaterals. It is not known to what extent the CA 1 and 2 longitudinal collateral system is a reciprocal one, so it is not possible to say what kind of collateral effect these fibres produce. The efferent projections from CA 1 and 2 are to a certain extent topographically organized (Raisman *et al.* 1966), so the only way one part of (say) CA 2 can influence cells to which another part projects is probably through the longitudinal association path. Such associational effects may require that the relevant collateral synapses are Hebb (or Brindley) synapses, and that the cortex is supplied with suitable inhibitory interneurons (e.g. cells *l* of § 4.1.1 for the Schaffer collaterals).

The afferent fibre systems to the hippocampus are also to some extent topographically organized (Raisman *et al.* 1965). It is therefore possible that a subevent may be fed into CA 3 and 4 alone: this subevent may previously have been associated with a simultaneous subevent in CA 1 and 2, but this may now be absent. The input to CA 3 and 4 can, through the Schaffer collaterals, evoke the original activity in CA 1 and 2 by stimulating cells there and relying on a local collateral effect (in the usual way). Provided (*a*) that the activities $\alpha$ in CA 1 to 4 are low enough for this simple kind of association to work (in conjunction with a local collateral effect), and (*b*) that the Schaffer collateral synapses are strong enough to allow rather few active facilitated synapses to stimulate a cell in CA 1 and 2, these collaterals could initiate this kind of associative recall. The higher the probability that a given Schaffer collateral synapse has been modified, the higher the number of facilitated collateral synapses that needs to be active at a CA pyramid in order for that cell to fire.

Hamlyn (1962) and Andersen (1966) describe the Schaffer collateral synapses as having a size between that of the usual spine synapses, and that of the mossy fibre synapses. This suggests that the probability that a Schaffer collateral synapse has been modified lies between the values for the other two kinds of synapse: i.e. if the probabilities that an ordinary collateral, a Schaffer collateral, and a mossy fibre synapse have been modified are $p_c$, $p_s$, $p_m$ respectively, one would expect that $p_c > p_s > p_m$.

### 4.5.2. *Commissural connexions*

Blackstad (1956) found that most hippocampal commissural fibres are very fine, and terminate in the Ss. Oriens and Rad., with a certain number from the contralateral area entorhinalis to S. Lac.-Molec. He was unable to determine the origins of many of these fibres, but from his evidence, and that of Raisman *et al.* (1965), it would seem that the projections are probably homotopic in CA 2 to 4, and are certainly homotopic and very symmetrical in CA 1.

The details of these projections are unimportant at the present crude level of theory: it is important only to note that, since the connexions are probably reciprocal, they probably allow a standard collateral effect (§ 2.4) between the hippocampi of the two sides. It is in accordance with

the theory that those fibres which terminate above S. Pyr. do so in S. Rad. rather than in S. Molec.; and with the notion that S. Molec. and S. Oriens are independent that they should distribute both above and below S. Pyr.

4.5.3. *The FD*

Cajal (1911) and Lorente de No (1934) both describe the collaterals of the dentate granule cells. They synapse with the dentate polymorph cells (as required by §4.4.2), and to some extent they ramify in the molecular layer. This would enable something of the usual collateral effect to take place among the dentate granules.

Blackstad (1956) describes massive degeneration in the inner one-quarter to one-third of the molecular layer after contralateral lesions, but is uncertain of the origin of the fibres responsible. Raisman *et al.* (1965) have some evidence which implicates the contralateral septum, but suspect there may be a projection from the contralateral CA 1.

## 4.6. *A brief functional classification of cell types*

4.6.0. *Introduction*

The distinction between archi- and neocortex is thought to reflect a difference in their functions. Archicortex is essentially *memorizing cortex*, in the sense that a given area of archicortex is likely to contain one or more layers of a simple memory. It typically contains cells resembling the hippocampal pyramids or the dentate granules, without climbing fibres. Neocortex, on the other hand, though undoubtedly used a great deal for simple associational storage, can probably be regarded as *classifying cortex*. Its operation depends on climbing fibres, and its success depends upon the truth of the fundamental hypothesis (Marr 1970, §1.6.4).

In the following sections 4.6.1 and 4.6.2 are listed the principal types of cell which the theories predict in memorizing (**M**) and in classifying (**C**) cortex. In general, archicortex is memorizing cortex, and neocortex can do both. Special additional considerations probably apply to those neocortical regions with special structure (e.g. primary sensory areas). This classification much abbreviates the analysis (§4.7) of the rest of the hippocampal formation.

4.6.1. *Memorizing cortex*

**M 1.** Large pyramidal cells without climbing fibres, with baskets. These cells usually form simple representations (i.e. can support a collateral effect): they have Brindley afferent synapses, and probably some dendritic independence. It is useful to refer to them as memorizing cells.

**M 2.** Star cells, and small pyramidal cells without climbing fibres, with baskets, are like **M 1**. They may be used with baskets in a simple memory, where subevents not wholly included in a learnt event are used to address that event, and are also included in the term 'memorizing cell'.

**M 3.** Star cells or small pyramids, without baskets, without climbing fibres, with small dendrites and ascending axons, are codon cells, used only at the first stage of a simple memory. Perhaps with modifiable synapses (Brindley), their principal function is to reduce $\alpha$.

**M 4.** Short-axon cells, without afferent baskets, without climbing fibres, with small dendrites, driven mainly by **M 1** or **M 2** cell collaterals, and with ascending axons. These cells are inhibitory. They control **M 1**, **M 2** or **M 3** cell dendritic thresholds for synaptic modification, and the level of subtractive inhibition during recall.

**M 5.** Short-axon cells like **M 4** only with local axons and dendrites. They synapse with **M 1**, **M 2** or **M 3** cells, and are inhibitory.

**M 6.** Basket cells, driven by the same afferents as drive **M 1, 2** or **3** cells, and sending inhibitory synapses to the somas of these cells. Basket cells may also receive synapses from **M 1** to **3** cell axon collaterals, since this would be one way of raising $f$ during recall.

**M 7.** Fusiform cells lying deep in the cortex, with a liberal dendritic expansion and local axonal arborization, typically to **M 3** or **M 1** and **2** cell dendrites. They are inhibitory threshold controlling cells, like **M 4**, which operate by negative feedback to the cells whose thresholds they control, and by direct sampling of afferents (cf. cerebellar Golgi cells).

4.6.2. *Classifying cortex* (Marr 1970)

**C 1.** Pyramidal cells with afferent climbing fibres and basket synapses, are cells representing classificatory units.

**C 2.** Star cells, or granule cells, without baskets, without climbing fibres, with small dendrites and often an ascending axon, are codon cells. They are driven mainly by afferents to that region of cortex, and some may have modifiable afferent synapses.

**C 3.** Cells whose axons become climbing fibres.

**C 4.** Short-axon cells other than **C 2**, with local axonal and dendritic ramification: they are inhibitory.

**C 5.** Basket cells, similar to **M 6.**

**C 6.** Fusiform cells with single ascending and descending dendritic shaft, usually lying deep in the cortex, and possessing an axon that goes to white matter without emitting any collaterals. These cells are probably cortical indicator cells of some kind, and some may project to archicortex.

### 4.7. *The histology of various hippocampal areas*

The letters (e.g. **M 3**) accompanying the following descriptions of the histology of allocortical regions refer to the cell classifications of §4.6. No detailed justifications of these diagnoses are given, since the arguments used for such justifications have all appeared in §4.

4.7.1. *The area entorhinalis* (a.e.)

The a.e. was studied by Cajal (1911) and by Lorente de No (1933), who reviewed and revised Cajal's work. The following summarizes the account given by Lorente de No (1933), which combines his and Cajal's work. Roman numerals indicate cortical layers, taken after Lorente de No.

I. Plexiform layer, with the usual short-axon cells (**M 5**). The axons here are mainly ascending axons from deeper layers (e.g. from layer V), and association fibres from other fields arriving through the plexiform layer.

II. Layer of star cells (**M 2**): their axons are thick and go to the white matter after giving off many collaterals. There are also various short-axon cells, some of which may synapse with the star cell somas (**M 5, M 4,** possibly **M 6**).

III. Layer of superficial pyramids (**M 2**). These cells have many dendrites in I, no branching in II, and a dense basilar dendritic field. The cingulum afferents to a.e. seem to end among these basilar dendrites (White 1959). The axon sends collaterals mainly to I and III (some to II and V) and goes to the white matter. Various short axon and miscellaneous other types of cell (**M 4** to **7**) are also found (III includes Cajal's (1911) layer 4°).

IV. Layer of deep pyramids, with thin unbranched dendritic shaft and immense basilar dendritic plexus (**M 2**). In this layer it is indigenous dendrites, rather than foreign axons, which

arborize and ramify. Their axons project to the white matter giving off many collaterals to I, II, III and V. The ascending collaterals rise vertically. Horizontal cells are also found here, probably including basket cells, and various cells with ascending axon (**M 4** to **7**). No collaterals of any extrinsic afferents terminate in this layer.

V. Small pyramidal cells with recurrent axons (**M 3**). Their axons send collaterals to I, II, III and V but not to IV. In IV, however, the dendrites ramify profusely, and the ascending axons synapse with them (probably) forming their main source of afferents. Globular cells with long dendrites inhabit layers V and VI, their axons arborising densely in layer V or VI (**M 7**). Spindle cells with short axons and local dendrites (**M 4, 5**) are also found. According to Cragg (1965), it is the fibres from ventral temporal neocortex which terminate here, in the cat.

VI. Layer of polymorph cells: there are many types, none particularly surprising; globular, polygonal, and those left over from V. They have various combinations of axonal and dendritic distributions (**M 3** to **7**).

4.7.2. *The presubiculum*

Cajal (1911) is the only author who has written about the presubicular histology, though Lorente de No (1934) was clearly familar with this area from his own observations (p. 137). It appears that on histological grounds, the hippocampal formation should be divided into three large regions, the Regio Entorhinalis, Regio Presubicularis and Regio Ammonica (Lorente de No 1934, p. 137). The Regio Entorhinalis and the Regio Presubicularis, in spite of many changes—particularly the introduction of star cells to layer II of a.e.—have the same fundamental plan. The Regio Ammonica starts with the introduction of the Ammonic pyramids in layer II of the prosubiculum, and continues into CA and FD. Thus the subiculum may be regarded as transitional cortex (Lorente de No's Subiculum *b*). (Cajal took what Lorente de No calls presubicular cortex (Sub. *a*) for his description of the human subiculum.)

The division of the hippocampal formation into three large areas, as suggested by Lorente de No on histological grounds, will be adopted here. The argument will essentially be that the Regio Entorhinalis and the Regio Presubicularis prepare information from many different sources for its simple representation in the CA and FD. It seems probable that each collection of cells in the Regio Entorhinalis and the Regio Presubicularis should be treated as preparing information from a separate source: the different shapes of the cells reflect the particular statistical quirks of the different kinds of information. The layer $\mathscr{P}_2$ of § 3.1 is a rough model for all of them.

The lack of detailed information about the Regio Presubicularis prevents its detailed discussion. The presubiculum of Cajal (1911) is presented as a typical example of presubicular cortex.

I. Plexiform layer, extremely wide, and containing many afferents to CA and FD. Its outer zone is composed almost entirely of such fibres, but the inner part contains the terminal bushes of ascending dendrites from layers described below, and so is a true plexiform layer. This region presumably contains the usual short-axon cells (**M 5**), but they seem to be difficult to stain with the Golgi method (Lorente de No 1934).

II. Layer of small pyramids and fusiform cells (**M 2, 3, M 7**?). The axons of many of these cells descend to the white matter, some ending locally. The dendrites of all seem to be confined to layers I and II.

III. Deep plexiform layer. (Lorente de No might have combined II and III as he did in a.e.) This layer is thick, with relatively few cells; small and medium pyramids (**M 2, 3**?) and various other cells (**M 4, 5, 7, 6**?). It contains an extremely dense plexus, and apparently, the layers I

to III receive here the terminal ramification of the massive pathway to the presubiculum carried by the cingulum.

IV. Large and medium pyramidal cells (**M 1, 2?**): the smaller pyramids are probably on average lower in the cortex, and their basilar dendrites generate a dense horizontal plexus. The large ones seem to have a more irregular dendritic arrangement (though information is very sparse, and these statements are inferences from Lorente de No's (1934) incidental remarks). All pyramidal cell axons go to the white matter. The large pyramids of this layer become layer III of the prosubiculum, and seem to be associated with the existence of Martinotti type cells (**M 3**) beneath them.

V. Fusiform and triangular cells, similar to those found in other cortical areas (**M 4, 5, 7**), and cells with ascending axon (**M 3**). No details are available.

4.7.3. *The subiculum* (Prosubiculum + Sub. *b* of Lorente de No)

It is convenient in this section to use the terminology of Lorente de No (1934, p. 134).

The subiculum lies next to CA 1, into which it gently merges. Only a very small region (Lorente de No's Sub. *b*) can be said to have a distinctive structure in that the presubicular pyramids have disappeared, but the prosubicular pyramids have not yet appeared. The huge terminal ramification of the cingulum is strictly confined to the presubiculum, and does not spill over into the prosubiculum (White 1959; Raisman *et al.* 1965; Cragg 1965).

I. An extremely wide plexiform layer, containing the perforant tract from a.e. to CA and FD. The lower zone is a true plexiform layer, and contains horizontally running collaterals of some of the fibres running overhead. There are the usual short-axon cells (**M 5**).

II. Modified ammonic pyramids (**M 1**). The apical dendrites lack S. Lac. and S. Rad., which ceases abruptly at the edge of CA 1. The basal dendrites are horizontal, and none descend to III. There are also many short axon and basket cells (**M 4, 5, 6**).

III. Prosubicular pyramids: the upper cells have no side branches in III to their dendritic shafts, but the lower ones do. None have any in II; all have them in I. Thus the cells of II avoid the plexus in III, and the cells of III avoid the plexus in II. These cells are probably **M 1**. Again, there are various short axon and basket cells (**M 4** to **6**).

Many pyramidal cells in the prosubiculum send axon collaterals to CA 1 and CA 2. Most axons enter the alveus of the CA, and thence enter the fimbria.

IV has two strata: (*a*) of globular cells, of which there are various kinds. Those whose axons pass to the white matter are probably **M 1**, and those with ascending axon are probably **M 3**; and (*b*) of Martinotti (**M 3**) type cells with local dendrite and ascending axon. These seem to be associated with the prosubicular pyramids, and to die out with them, which suggests that their axons do not rise above III. It may be these axons which cells of II are anxious to avoid.

Layer IV, especially IV b, becomes very thin towards the CA. III becomes very wide, and the cells seem to turn into Ammonic pyramids as IV b disappears (Lorente de No 1934, p. 129, figure 11). The prosubiculum thus merges into and becomes the CA, which has already been described.

## 5. Neurophysiological predictions of the theory

### 5.0. *Introduction*

In this section are summarized the most important predictions which follow from the notion that simple memory provides a model for the archipallium in general, and for the hippocampus in particular. They are presented in two parts; the first summarizes the general model for archicortex, and the second deals with the detailed predictions for the hippocampus.

The statements are made with varying degrees of firmness, which are indicated by the number of stars accompanying each (after Marr 1970, §7). Three stars indicates a prediction whose disproof would show simple memory theory to be an inappropriate model for archicortex; a no-star prediction is a strong hint and nothing more: one and two star statements lie between these extremes.

### 5.1. *The general model for archicortex*

Whereas neocortex is capable both of classifying and of memorizing inputs, archicortex is capable only of memorizing them***. The variety of the functions performed by archicortex is achieved in part by the application of its basic memorizing ability to widely different kinds of information. Two examples of the uses to which archicortex may be put are free simple memory (in which the memory projects to its own input cells), and directed simple memory (in which it does not).

The central feature of archicortex is a collection of so-called memorizing cells, identified as that class of cell which is most numerous and whose axons project elsewhere. Such cells will have at least two kinds of afferent synapses***: excitatory afferent synapses with Brindley modification conditions***; and unmodifiable inhibitory afferent synapses***. The dendrites of memorizing cells are often independent**, modification conditions being decided locally**.

The inhibition applied to memorizing cells performs at least two principal functions: one is to control the synaptic modification conditions in the memorizing cell dendrites during the learning of events**; and the other is to control the cells' thresholds during the recall of previously learned events***. Cells for the first function apply inhibition to the dendrites of the memorizing cells**, and are driven either by memorizing cell axon collaterals, or by afferent collaterals, or both (by analogy with the cerebellar Golgi cells). They act so as to maintain the number of memorizing cells involved in learning each new event at a roughly constant level**.

Cells for the second function are of two types**; basket cells, performing a division**, and stellate cells, synapsing with the dendrites, performing a subtraction**. The stellate cells act to remove from the output signal some of the excitation due to the unmodifiable component of the Brindley synapses*. The basket cells and stellate cells are driven by the main afferent system to the memorizing cells (through unmodifiable excitatory synapses)**, and perhaps also by memorizing cell collaterals*. It is appropriate in certain circumstances to raise the division threshold of the memorizing cell during recall of a learnt event**. There are various circuits capable of achieving this.

Archicortex may contain codon cells, perhaps with modifiable afferent synapses. If so, and if the synapses are modifiable, then they are Brindley synapses**, and are accompanied by the same kinds of inhibitory housekeeping cells as are memorizing cells**. They are often small and numerous**, and are necessary when the activity ($\alpha$) of the input fibres is too high for the learning capacity required of the memorizing cells***.

It is the lack of climbing fibres which deprives archicortex of the clustering ability underlying

the classification process in neocortex**. Archicortex is therefore bad at the kind of classification of which neocortex is probably capable***.

This outline of the processes carried out in archicortex gives rise to a rough classification of archicortical cell types. These have been labelled **M 1** to **M 7**, and are not set out here since they have been summarized in the appropriate way in § 4.6.1. For the purposes of this section, they may be regarded as owning two stars, except where overridden by the statements made above.

### 5.2. *The hippocampal cortex*

Star ratings in this section test the proposition that the various divisions of the hippocampal formation form components of a simple memory.

The pyramidal cells of CA 1 to 3 and the granule cells of the FD are memorizing cells, in the sense of § 4.6.1***. Their main afferents therefore terminate by means of Brindley modifiable synapses***. All other cells there are probably inhibitory**, and certainly many are***. These cells are concerned with the formation of simple representations, in the sense of § 3**. That is, the activities of these populations are low** (near 0.001) and there is an extensive collateral system** which uses Hebb (or Brindley) modifiable synapses**. The collaterals aid the completion of simple representations during recall**. The performance of regions of the CA (e.g. say CA 2) is qualitatively similar to that of the layer $\mathscr{P}_3$ in the explicit model of § 3.1**.

The star cells of the entorhinal area are also memorizing cells***, and are qualitatively analogous to the layer $\mathscr{P}_2$ of the model of § 3.1**. Various predictions follow from these remarks, in particular that they possess Brindley modifiable afferent synapses***. Many other cells in various archicortical areas have been discussed, and the predictions concerning them follow the general lines of § 5.1. In the following lists, the various cells are classified according to the terminology of § 4.6.1; the firmness of the classification is indicated; and the references specify the relevant pieces of text.

5.2.1. *Cornu ammonis: CA 1 to 3*

| cell type | described (§) | stratum | class | reference(§) | stars |
|---|---|---|---|---|---|
| pyramid | 4.1.1 (*a*) | pyr. | **M 1** or **2** | 4.2 | *** |
| pyr. basket | 4.1.1 (*b*) | pyr. | **M 6** | 4.3.3 | *** |
| asc. axon | 4.1.1 (*c*) | pyr. | **M 4, 5** | 4.3.2 | ** |
| horizontal | 4.1.1 (*d*) | oriens | **M 4** | 4.3.2 | ** |
| polygonal | 4.1.1 (*e*) | oriens | **M 4, 5** | 4.3.2 | ** |
| basket | 4.1.1 (*f*) | oriens | **M 6** | 4.3.3 | *** |
| basket | 4.1.1 (*g*) | oriens | **M 6** | 4.3.3 | *** |
| horizontal | 4.1.1 (*h*) | oriens | **M 4** | 4.3.4 | ** |
| horizontal | 4.1.1 (*i*) | oriens | **M 5** | 4.3.2 | ** |
| various | 4.1.1 (*j*) | rad. & lac. | **M 5** | 4.3.2 | *** |
| asc. axon | 4.1.1 (*k*) | rad. & lac. | **M 4** | 4.3.2 | *** |
| horizontal | 4.1.1 (*l*) | rad. & lac. | **M 5** | 4.3.4 | *** |
| short axon | 4.1.1 (*m*) | molec. | **M 5** | 4.3.2 | *** |
| horizontal | 4.1.1 (*n*) | molec. | **M 5** | 4.3.2 | *** |

One kind of cell can fall into two classes if it possesses two kinds of dendritic or axonal distribution.

There may be an afferent system capable of changing the ratio of somatic to dendritic inhibition at the CA pyramids. This would increase the amount of basket inhibition during recovery of a simple representation. No-star estimates of the values of the relevant parameters for CA appear in § 4.2.4.

5.2.2. *The fascia dentata*

| cell | described (§) | layer | class | reference (§) | stars |
|---|---|---|---|---|---|
| granule | 4.1.2 | granular | **M 1** or **2** | 4.4.1 | *** |
| displaced gran. | 4.1.2 (*a*) | molec. | **M 1** or **2** | 4.4.1 | ** |
| short axon | 4.1.2 (*b*) | molec. | **M 4, 5** | 4.4.1 | *** |
| pyr. basket | 4.1.2 (*c*) | polymorph | **M 6** | 4.4.1 | *** |
| asc. axon | 4.1.2 (*d*) | polymorph | **M 4** | 4.4.1 | ** |
| desc. axon | 4.1.2 (*e*) | polymorph | } probably CA 4. | | |
| short axon | 4.1.2 (*f*) | polymorph | | | |
| star, etc. | 4.1.2 (*g*) | polymorph | | | |

$\alpha_{FD}$ is probably rather low (near 0.001)*.

5.2.3. *CA* 3, *CA* 4 *and the mossy fibres*

The pyramids of CA 4 are 'collector' cells for the output of FD granule cell activity*, (§§4.4.0, 4.4.3). They may have Brindley modifiable afferent synapses from FD granule cell axons*, being the short-axon cells of CA 4 the necessary class **M 4** and **M 5** cells*. The mossy fibre synapses in CA 3 may be Hebb or Brindley synapses*. The large size of the mossy fibre synapses suggests that $\alpha_{FD}$ is very low*—certainly lower than $\alpha$ for the other hippocampal afferents** (§4.4.3).

5.2.4. *Hippocampal collateral systems*

All short hippocampal pyramidal cell collaterals to other hippocampal pyramids end in Hebb or Brindley modifiable synapses**. Those collaterals which are reciprocated can take part in the collateral effect**. Those which do not are concerned with associating simple representations formed in different regions of the hippocampus (§4.5.1), these being completed by local reciprocating collaterals*. Examples of the second sort are the mossy fibres**, and the Schaffer collaterals**. Examples of the first kind are local collaterals**, and perhaps commissural connexions (§4.5.2). There should be $\gtrsim$ 10000 collateral synapses at each CA pyramidal cell**. Local collaterals joining hippocampal pyramids tend to make synapses in S. Rad.* (§4.5). There may be a collateral effect in FD (§4.5.3).

5.2.5. *Area entorhinalis*

| cell | described (§) | layer | class | references (§) | stars |
|---|---|---|---|---|---|
| short axon | 4.7.1 I | I | **M 5** | 4.7.1 | *** |
| star | 4.7.1 II | II | **M 2** | 4.7.1 | *** |
| various | 4.7.1 II | II | **M 4, 5, 6?** | 4.7.1 | ** |
| pyramid | 4.7.1 III | III | **M 2** | 4.7.1 | *** |
| various | 4.7.1 III | III | **M 4–7** | 4.7.1 | ** |
| pyramid | 4.7.1 IV | IV | **M 2** | 4.7.1 | *** |
| various | 4.7.1 IV | IV | **M 4–7** | 4.7.1 | ** |
| pyramid | 4.7.1 V | V | **M 3** | 4.7.1 | *** |
| globular | 4.7.1 V | V | **M 7** | 4.7.1 | ** |
| spindle | 4.7.1 V | V | **M 4, 5** | 4.7.1 | ** |
| polymorph | 4.7.1 VI | VI | **M 3–7** | 4.7.1 | * |

5.2.6. *Presubiculum*

| cell | described (§) | layer | class | references (§) | stars |
|---|---|---|---|---|---|
| short axon | few seen | I | **M 5** | 4.7.2 | *** |
| pyramids | 4.7.2 II | II | **M 2** or **3** | 4.7.2 | *** |
| fusiform | 4.7.2 II | II | **M 7** | 4.7.2 | none |
| various | 4.7.2 III | III | **M 2–7** | 4.7.2 | (little information) |
| pyramids | 4.7.2 IV | IV | **M 1?, M 2** | 4.7.2 | ** |
| fusiform } triangular } | 4.7.2 V | V | **M 4, 5, 7** | 4.7.2 | * |
| asc. axon | 4.7.2 V | V | **M 3** | 4.7.2 | ** |

This region has been studied even less than the others.

5.2.7. *Prosubiculum* (of Lorente de No)

| cell | described (§) | layer | class | reference (§) | stars |
|---|---|---|---|---|---|
| short axon | 4.7.3 I | I | **M 5** | 4.7.3 | *** |
| pyramid | 4.7.3 II | II | **M 1** or **2** | 4.7.3 | *** |
| short axon | 4.7.3 II | II | **M 4, 5** | 4.7.3 | ** |
| basket | 4.7.3 II | II | **M 6** | 4.7.3 | *** |
| pyramid | 4.7.3 III | III | **M 1** or **2** | 4.7.3 | *** |
| pyramid | 4.7.3 IV | IVa | **M 1** or **2** | 4.7.3 | *** |
| Martinotti | 4.7.3 IV | IVb | **M 3** | 4.7.3 | *** |
| short axon | 4.7.3 | III, IV | **M 4–6** | 4.7.3 | ** |

The prosubicular pyramids are probably **M 1**** since they send collaterals to CA and axons to the fimbria.

I wish to thank Professor G. S. Brindley, F.R.S. for his helpful criticisms, and Mr S. J. W. Blomfield for many discussions. The following kindly gave me permission to reproduce figures from other papers: Dr T. W. Blackstad and the Wistar Press for figure 2; Professor R. Lorente de No and Akademie-Verlag GmbH for figures 8 to 11; and C.S.I.C. Madrid for figures 12 to 16. The work was supported by Trinity College, Cambridge.

*Note added in proof*, 15 *April* 1971

Lømo (1971) has published evidence for the facilitation of the perforant path—Dentate granule cell synapses in the rabbit. His findings are necessary but not sufficient to prove this theory's prediction (§5.2.2) that these synapses are facilitated by simultaneous pre- and post-synaptic depolarization.

## References

Andersen, P. O. 1966 Correlation of structural design with function in the archicortex. In *Brain and conscious experience* (ed. J. C. Eccles), pp. 59–84. Berlin: Springer-Verlag.

Andersen, P. O., Eccles, J. C. & Løyning, Y. 1963 Recurrent inhibition in the hippocampus with identification of the inhibitory cell and its synapses. *Nature, Lond.* **198**, 540–542.

Blackstad, T. W. 1956 Commissural connections of the hippocampal region in the rat, with special reference to their mode of termination. *J. comp. Neurol.* **105**, 417–538.

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*, Tome II. Madrid: C.S.I.C.

Cragg, B. G. 1965 Afferent connections of the allocortex. *J. Anat.* **99**, 339–357.

Cragg, B. G. 1967 The density of synapses and neurones in the motor and visual areas of the cerebral cortex. *J. Anat.* **101**, 639–654.

Eccles, J. C., Llinás, R. & Sasaki, K. 1966 Parallel fibre stimulation and the responses induced thereby in the Purkinje cells of the cerebellum. *Expl Brain Res.* **1**, 17–39.

Hamlyn, L. H. 1962 The fine structure of the mossy fibre endings in the hippocampus of the rabbit. *J. Anat.* **96**, 112–120.

Lømo, T. 1971 Potentiation of monosynaptic EPSP's in the perforant path—dentate granule cell synapse. *Expl Brain. Res.* **12**, 46–63.

Lorente de No, R. 1933 Studies on the structure of the cerebral cortex. I. The Area Entorhinalis. *J. Psychol. Neurol. (Lpz.)* **45**, 381–438.

Lorente de No, R. 1934 Studies on the structure of the cerebral cortex. II. Continuation of the study of the Ammonic system. *J. Psychol. Neurol. (Lpz.)* **46**, 113–177.

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

Marr, D. 1970 A theory for cerebral neocortex. *Proc. Roy. Soc. Lond.* B **176**, 161–234.

Raisman, G., Cowan, W. M. & Powell, T. P. S. 1965 The extrinsic afferent, commissural and association fibres of the hippocampus. *Brain* **88**, 963–996.

Raisman, G., Cowan, W. M. & Powell, T. P. S. 1966 An experimental analysis of the efferent projection of the hippocampus. *Brain* **89**, 83–108.

Schaffer, K. 1892 Beitrag zur Histologie der Ammonshornformation. *Arch. mikrosk. Anat.* **39**, 611–632.

Shariff, G. A. 1953 Cell counts in the primate cerebral cortex. *J. comp. Neurol.* **98**, 381–400.

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

White, L. E. Jr. 1959 Ipsilateral afferents to the hippocampal formation in the albino rat. I. Cingulum projections. *J. comp. Neurol.* **113**, 1–41.

**David Willshaw**

Commentary on

## *Simple Memory: A Theory of the Archicortex*

This is the third, and last, of David Marr's series of three theoretical papers on the neurobiology of learning and memory (Marr 1969, 1970, 1971). In this paper, he proposes a theory for the functioning of the mammalian hippocampus — one of the most important but least understood parts of the brain.

The hippocampus is one of the phylogenetically older parts of the brain (hence:archicortex). It is found in mammals, and a related structure exists in birds. The mammalian hippocampus has a simple and regular structure, and specific circuits have been identified within it. It has afferent and efferent pathways to many parts of the neocortex, and these interconnections are fairly well characterized.

It has proved difficult to assign positively any definite function, or functions, to the hippocampus. Nonetheless, various proposals have been made. At the time Marr wrote this paper, the startling results from such patients as HM, who became amnesic after undergoing bilateral hippocampectomy for the relief of epilepsy, suggested a role for the hippocampus in memory (Scoville and Milner, 1957). More recently, the idea has been developed that a "cognitive map" is built in the hippocampus (O'Keefe and Nadel, 1978). This is based on the finding that there are "place units" in the rat hippocampus — neurons that fire when the animal is at a specific place in the environment.

Marr had previously proposed (1970) that the neocortex is the site of long-term associative storage of information, the information being stored in a form that retains the essential details and removes the superfluous. In the hippocampus paper, he argues that it would be inefficient to store the raw associations directly, before the salient features had been extracted; furthermore, neocortical interconnectivity is not sufficiently complete to allow any arbitrary association to be stored. Marr proposes that there is a special temporary memory store — the hippocampal formation.

The central question is concerned with the architecture required for this temporary memory, and whether it matches the known structure of the hippocampus. As in his other papers on learning and memory, Marr's method of working is to constrain the problem by a number of assumptions as to the likely values of some of the parameters of the system. These values either were derived intuitively (e.g., the number of items to be stored) or had some biological basis (e.g., the number of synapses on a nerve cell). To these assump- tions are added a number of computational constraints that must hold if the memory is to perform satisfactorily. He concludes that there

must be an intermediate layer of cells between the input and output layers of the memory. In the present day parlance of Connectionism, this would have a natural interpretation as a layer of hidden units. Having derived a complete specification of this three-layer system, he goes on to relate this three-layer model to the known facts of the hippocampus and its connections to the neocortex.

In this paper, Marr's use of a set of constraints to derive the minimal structure for the given problem reaches a most sophisticated level. However, his attempt is not wholly satisfactory, since there is an inconsistency in the argument, which leaves his case for a three-layer model not proven. He therefore relies more heavily on his view of hippocampal circuitry than is stated explicitly. In effect, he views the problem from two different perspectives: (a) that the structure of the memory proposed is necessary on computational grounds and (b) that it has to have this structure because this is the way that the hippocampus was built. This double perspective can be seen in light of his subsequent development of the importance of the computational, the algorithmic and the implementational levels of explanation (Marr, 1982).

Although he does characterize in some detail the individual properties of the cells that are meant to form the layers of his model, only a loose corres pondence is made between the subdivisions of the hippocampus (together with the associated neocortical circuitry) and the layers of his model. The most extensive discussion is concerned with the nature of the cells of the output layer of the memory, which are identified with the pyramidal cells of the hippocampus. He does not distinguish between the various elements of the Dentate Gyrus-CA3-CA1 trisynaptic circuit, the existence of which was known at the time (Andersen et al., 1971). This may have been a foresighted omission, given that the notion of the trisynaptic circuit itself is now in the process of change with the discovery of other extrinsic pathways of the hippocampus (Squire et al., 1989). His major contribution is in his discussion, at a cellular and subcellular level, of the properties that the individual elements of his model must have. In particular, he proposes various types of *dual strategies* for setting the thresholds of the cells (which have never been properly investigated since), which are required for efficient storage and retrieval in the biologically realistic cases of incompletely connected networks. The requirement that synapses be modified by simultaneous presynaptic and post-synaptic activity, after the style of Hebb (1949), predates the discovery of hippocampal long-term potentiation (Bliss and Lømo, 1973), although he does add a note in proof about Lømo's earlier paper (1971) that showed synaptic facilitation in the perforant path — dentate gyrus.

In summary, David Marr presents a somewhat abstract interpretation of the hippocampus as a temporary memory store. The strength of his analysis lies not in the translation of his formal model into neurobiological terms, but rather in his discussion of what types of local circuitry are required to perform the various computations that are needed for the memory to function efficiently.

It is unfortunate that this paper is not more widely read or understood. It

required considerable effort to come to terms with the inaccessible style that is characteristic of his earlier writings; but I found that the effort was well worthwhile. Even 20 years after publication, Marr's theory remains the most complete computational model of the hippocampus.

This short commentary is based on a recent review of the computational basis of Marr's theory of archicortex (Willshaw and Buckingham, 1990). We also describe the results of analysis and of computer simulations that were designed to compare the performance of two-layer and three-layer models.

## REFERENCES

Andersen P, Bliss TVP, Skrede K (1971): Lamellar organization of hippocampal excitatory pathways. *Exp Brain Res* 13:222-238

Bliss TVP, Lmo T (1973): Long-lasting potentiation of synaptic transmission in the dentate area of the anesthetized rabbit following stimulation of the perforant path. *J Physiol* 232:331-356

Hebb DO (1949): *The Organization of Behavior.* New York: John Wiley and Sons

Lømo T (1971): Potentiation of monosynaptic EPSP's in the perforant path— dentate granule cell synapses. *Exp Brain Res* 12:46-63

Marr D (1969): A theory of cerebellar cortex. *J Physiol* 202:437-470

Marr D (1970): A theory for cerebral cortex. *Phil Trans Roy Soc* B 176:161-234

Marr D (1971): Simple memory: theory for archicortex. *Phil Trans Roy Soc B* 262:23-81

Marr D (1982): *Vision.* New York: Freeman

O'Keefe J, Nadel L (1978): *The Hippocampus as a Cognitive Map.* Oxford: Oxford University Press

Scoville WB, Milner B (1957): Loss of recent memory after bilateral hippocampal lesions. *J Neurol Neurosurg Psychiat* 20:11-12

Squire LR, Shimamura AP, Amaral DG (1989): Memory and the hippocampus. In: *Neural Models of Plasticity*, Byrne J, Berry W, eds. NY: Academic Press

Willshaw DJ, Buckingham JT (1990): An assessment of Marr's theory of the hippocampus as a temporary memory store. *Phil Trans Roy Soc B* (in press)

*Professor*
*Centre for Cognitive Science*
*University of Edinburgh*
*Edinburgh, Scotland UK*

**Bruce McNaughton**

Commentary on

## *Simple Memory: A Theory of the Archicortex*

I regard it as a significant honor to be able to comment here, from a neurobiologist's perspective, on the impact of David Marr's theoretical neural network models on our understanding of the biology of associative memory, in particular in the mammalian hippocampal formation and neocortex. While there is some truth to Willshaw and Buckingham's (1990) suggestion that some of us have cited Marr's papers rather more widely than we have understood them, his three papers (Marr 1969, 1970, 1971) on the cerebellum, neocortex and archicortex (hippocampus) have been guiding lights both to myself and to a number of other experimental neuroscientists. (It is unfortunately also the case that Marr's ideas are sometimes more widely exploited than they are cited.) Marr's approach, in its mathematical rigor, was always difficult, and often obscure to the non-mathematician. This, unfortunately, led to his theories being less widely appreciated (or understood) among neurobiologists than they might otherwise have been; however, the value of Marr's models for neurobiological studies lies not so much in their mathematical sophistication or overall correctness in detail (they are almost certainly wrong), but for the far-reaching explanatory power of their relatively simple individual components. It is the broad conceptual framework provided by these models, rather than their correctness in detail, that will insure Marr his important place in the historical development of our understanding of how biological neural networks actually operate. Looking back to the sparsity of the experimental database from which Marr developed his ideas, it is astounding the extent both to which these insights have been substantiated, and to which they have brought order to a number of otherwise disconnected data on the anatomy, biophysics and information transmission of the mammalian hippocampal formation and its relations with the neocortex. Contrary to Willshaw and Buckingham's (1990) statements, many of Marr's predictions have, in fact, been followed up. In the following I shall attempt to illustrate this with a few of the more salient examples.

*Synaptic Modification*

Marr was the first theoretician to attempt to make use, in the context of a detailed, neurobiologically constrained model, of Hebb's postulate that synapses should be enhanced under conditions of conjoint pre- and postsynaptic depolarization. At the time that he wrote, the first experiments by Lømo, and subsequently by Bliss, Gardner-Medwin and Lømo, were beginning to reveal that hippocampal synapses exhibited a plasticity of sufficient duration to be considered as a candidate for associative memory; however, it was not until much later that the first evidence was obtained that Hebb's

principle might be implemented in this process (McNaughton, Douglas and Goddard, 1978), and later still before this was fully confirmed and understood mechanistically (Collingridge et al., 1983; Gustafsson et al.,1987; Harris et al., 1984; Kelso et al., 1986; Wigström et al 1986). A substantial body of literature has accumulated that is at least consistent with the idea that this process does, indeed, reflect the experimental activation of mechanisms that normally subserve at least the initial registration of associative memories (see McNaughton and Morris, 1987, for overview). Most of the available data indicate that the characteristics of the main modification process are consistent with what Marr called "Brindley" synapses (which have a non-modifiable excitatory component) rather than binary "Hebb" synapses, although this question is by no means closed.

*Pattern Completion*

In the archicortex paper, Marr suggests that the completion of stored events from fragmentary or noisy input information should be the primary function of the "simple memory" system he envisioned for the hippocampus. This fundamental idea has proven to be of immense value in the design of neurophysiological and behavioral experiments, and two lines of investigation now suggest the fundamental correctness of this assertion. In the rodent hippocampus, the "events acted upon relate primarily (or at least most obviously) to the animal's representation of space. Individual pyramidal cells are selectively active in limited regions and orientations within the animal's known environment. Although these "place fields" are determined by the animal's orientation with regard to the distal visual landmarks, removal of any subset of these landmarks has little or no effect on the spatial information transmitted by these cells (O'Keefe, 1976; O'Keefe and Conway, 1978). More direct evidence for pattern completion in hippocampal circuits was recently obtained in studies (Mizumori et al., 1989b) in which the discharge rates and spatial selectivities of CA3 pyramidal cells were severely curtailed by temporary inactivation of a modulatory input from the medial septum, which is necessary to maintain the excitability of CA3 cells. Pyramidal cells in CA1, whose major source of modifiable excitatory input is CA3, were almost completely unaffected. Somehow, the highly reduced subsets of spatial representations conveyed from CA3 were sufficient to enable complete spatial representations in CA1.

*Inhibitory Control of Global Threshold During Storage and Recall*

Perhaps the most insightful and powerful of Marr's ideas was his suggestion that inhibitory interneurons should control both the threshold for synaptic modification during storage, and, by means of a division operation, the output threshold for principal cells during associative recall (pattern completion). The former notion has been verified in a number of studies that have shown that the modification of hippocampal synapses is largely regulated by GABAergic inhibition (Wigström and Gustafsson, 1983; Sharfman & Sarvey, 1983; Larson et al., 1986). The latter idea, although more difficult to verify, has some experimental support. Inhibition mediated by the chloride dependent $GABA_A$ channel is fundamentally a division operation (for elaboration, see McNaughton and Barnes, 1990, and McNaughton and Nadel, 1989). Because the chloride equilibrium potential is almost the same as the resting potential, the effect

of inhibition (relative to rest) is primarily to increase membrane conductance. Because the soma voltage change is roughly the outward excitatory synaptic current ($i_m$) divided by membrane conductance ($g_m$), and because resting conductance is small, a division operation is implemented ($DV_m = i_m/g_m$). Secondly, in the studies cited above by Mizumori et al. (1989b) in which CA1 output was preserved in the face of reduced and degraded CA3 input, the activities of basket inhibitory interneurons were reduced in proportion to the reduced CA3 input. This appears consistent with Marr's idea that inhibitory cells should sample the activity in the input fiber population and feed forward a proportional division signal. Also consistent is the fact that, in all hippocampal subfields, most inhibitory cells receive direct excitation from the same modifiable excitatory inputs that project to the principal cells. As suggested by the idea of setting the output threshold globally, these cells need not be numerous, and indeed, in the hippocampus, they constitute only a small population relative to the principal cells. It is also known that the behavioral conditions under which the density of afferent activity from entorhinal cortex to hippocampus is greatest are also the conditions under which hippocampal inhibitory cells are most active. In further support of this idea, the probability of inhibitory cell output is graded with stimulus intensity (i.e., number of active afferents), whereas the principal cells do not normally fire until the intensity is high enough to activate many more afferents than would be coactive in a typical physiological event (Mizumori et al., 1989a).

Another interesting consequence of the threshold setting hypothesis is that, unlike principal cells, which care about exactly *which* afferents are active in an event, the inhibitory cells should care primarily only about *how many* are active (McNaughton and Nadel, 1990). This clearly characterizes the differences in spatial firing characteristics between hippocampal pyramidal and basket cells.

Finally, although Marr did not consider in detail the dynamics of his proposed 'input normalization', there is one logical consequence of this scheme which provides considerable insight into the dynamics of the feed-forward inhibitory networks of the hippocampus. In order for the division operation to be effective, the division signal arriving at the principal cell soma must arrive with or before the excitatory synaptic signal from the current event; however, the inhibitory signal must cross two synapses, whereas the excitatory signal need cross only one. To compensate for this, the inhibitory system appears to have evolved an extremely rapid response mechanism. When hippocampal afferent fibers are electrically activated, inhibitory cells fire well before principal cells (Ashwood et al., 1984; Buzsaki, 1984; Douglas, McNaughton and Goddard, 1983; Mizumori et al., 1989a) so that the inhibitory conductance in the principal cells is already activated before most principle cells reach threshold.

*The Necessity for Keeping a Low*

Marr proposed that the simple memory system must satisfy the dual constraints of maximizing the event storage capacity, while at the same time preserving enough information from each event to ensure reliability. These constraints essentially dictated the size of the required networks, and the proportion $a$ of cells that could be used in the representation of any given event. Marr proposed that the value of $a$ should lie between 0.01 and 0.001, and be

roughly constant across events. To translate this into actual neuronal firing rates, take as the 'time-step' the apparent time constant of most hippocampal and cortical neurons, which is on the order of 0.01 sec. The corresponding average firing rates then become between 1.0 and 0.1 Hz, values that are quite low by the standards of most cortical neurons. It turns out that these are about the typical mean discharge frequencies for hippocampal principal cells recorded in alert rats during the performance of spatial learning tasks dependent on the integrity of the hippocampus (O'Keefe, 1976; McNaughton et al., 1983). This 'sparse' encoding of events is also manifested in the exquisite spatial selectivity exhibited by hippocampal pyramidal cells. In extended spaces, a typical cell fires intensely only in a highly restricted region of the animal's known accessible environment, a region typically covering on the order of 0.01 to 0.001 of the total area (these values vary somewhat depending on the size of the environment and other factors). It is also of interest that this sparse coding scheme appears to be a unique characteristic of the hippocampus. In both the entorhinal cortical input and the subicular output structures, spatial coding is considerably more highly distributed, and a (mean firing rate) is correspondingly substantially higher (Barnes et al., 1990).

Marr proposed a rule of thumb for the relationship between $a$ and the number of events $n$ to be stored:

$$n\, a_{i-1}\, a_i \leq 1$$

This ensures that when $n$ inputs have been learned, not all of the synapses have been modified. Using Marr's proposed parameters, this translates to between about 60% and about 10% modified synapses at full capacity, depending partly on how much information is to be made available for retrieval. Above these values, information storage would be unreliable, a given subevent recalling either too many active output cells, or the wrong ones (this is quite analogous to the psychological concept of interference). The prediction of these considerations is that simple memory will fail if the above constraint on the number of modified synapses is exceed. This is exactly the behavioral consequence of artificially increasing the proportion of modified synapses in the hippocampus by high-frequency stimulation of the main input pathways bilaterally. Such stimulation induces a long-term enhancement (LTE/LTP) of a significant proportion of perforant path synapses. This enhancement persists for several weeks, during which time there is both a disruption of recently stored spatial memories and an inability to store new ones (McNaughton et al., 1986; Castro et al., 1990). It is also entirely consistent with Marr's notion of the hippocampus as a temporary memory system that electrically induced synaptic enhancement decays over time, at least at these synapses.

*The Collateral Effect*

Marr suggested that pattern completion occurred in the pyramidal layers via a "collateral effect". The fundamental idea was that modifiable excitatory collateral synapses would assist recall over several cycles of recurrent excitation. After input of an appropriate subevent, additional cells belonging to the original stored pattern would be activated on succeeding cycles. The "collateral effect" mechanism has now come to be known as "recurrent autoassocia-

tion" (Kohonen, 1972, 1978) and, in one form or another, figures importantly in a number of connectionist style models. Although the implementation of a collateral effect in the hippocampus has yet to be verified experimentally, CA3 has an abundant system of modifiable recurrent collaterals which could perform this function. Also, Marr made two predictions about the dynamics of the collateral effect which seem to be approximately supported by modern data. First, he supposed that about three cycles of the collateral effect should be sufficient to complete the representation. When the hippocampus is actively processing inputs, there is an oscillating cycle of excitation and inhibition known as the theta rhythm, whose mean period is about 140 msec, and to which all hippocampal cell output is phase locked. If one assumes that the completion effect must be going on during the quarter cycle when excitation is increasing, this allows about 35 msec. In the CA3 recurrent system, the combination of conduction delay and synaptic delay amounts to about 6 to 8 msec. This would thus be sufficient for about four to six cycles; only slightly more than Marr predicted. The second prediction was that there should be some external mechanism which gradually increases inhibition during the collateral effect, to make sure only the correct cells were included in the output. The medial septal nucleus, which paces the theta rhythm, has a strong modulatory effect on inhibitory interneurons. As predicted, the activity of the inhibitory cells does increase during the rising excitatory phase of the theta rhythm.

*Orthogonalization of Similar Input Vectors*

One of the most powerful of Marr's concepts was the idea that memory capacity could be maximized if representations that were rather similar at the input could be recoded by a separate group of cells in such a way as to minimize the overlap in the output. In his cerebellum paper, Marr assigned this function to the cerebellar granule cells, which he called "codon" cells. In the cerebellar paper, the basic idea was to project the input vector onto a higher dimensional space (there are about 40 billion granule cells in the human cerebellum) and then use this orthogonalized representation as input to the memory cells. In the models for neocortex and archicortex, it was considered to be more economical if codons were not hard-wired, but could be created on demand through the use of modifiable synapses. In this way only those codons (subevents) which actually occurred in the experience of the animal would be required. It turns out that the initial projection from the entorhinal cortex into fascia dentata does involve a projection into a higher dimensional space. There are about 105 entorhinal projection cells, and about 106 granule cells in the fascia dentata. This projection terminates in modifiable synapses (probably of the Brindley variety). Moreover, single neuron recording studies of physiologically identified granule cells indicate that $a$ in the granule cell population is among the lowest of any hippocampal subfield (Mizumori et al., 1989a). Thus, although the question requires more systematic investigation, Leonard (1990) has obtained preliminary evidence for pattern separation in the hippocampal output cells in CA1.

*Readout from Simple Memory During Sleep*

One of the boldest of Marr's predictions was that readout from simple memory should occur during sleep. This idea was first developed in the neo-

cortex paper. Marr argued that "information from which a new classificatory unit is to be formed will often come from a simple associative store," (i.e. hippocampus) "not from the environment ... the most natural way of selecting a location for a new classificatory unit was to allow one to form wherever enough of the relevant fibers 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."

It is unclear whether Marr was aware that at the time this was written, there was a growing psychological literature on the possible role of sleep, particularly REM sleep (Leconte and Bloch, 1971), in memory consolidation (for reviews see Fishbein and Gutwein, 1977; Horne and McGrath, 1984; Smith, 1985). Certainly the basic idea seems to have fallen out from the premises of the model. Recently, some very exciting neurophysiological studies have produced strong support for the plausibility of Marr's idea that information is transferred from temporary (hippocampal) to permanent (cortical) memory during sleep. Pavlides and Winson (1989) studied the effects of selective spatial experience on the subsequent activity of hippocampal "place" cells during sleep. They recorded from pairs of place cells with nonoverlapping place fields. During the waking episode, they exposed the animal to the field of one member of the pair but not to the field of the other. They then removed the animal to a neutral location and allowed it to fall asleep. During the sleep episode, there was a large increase in the output activity of the cells that had been exposed to their place fields, in particular, the occurrence of high-frequency bursts increased, and the interspike intervals during bursts decreased. This are exactly the sort of activity that would be most likely to lead to synaptic modification in target cells. The effect was present in all phases of sleep, but was greatest in REM sleep. This phenomenon thus seems to fit precisely the requirements suggested by Marr's sleep hypothesis.

*Closing the Loop*

In the foregoing, I have tried to illustrate the astounding prescience of Marr's neurobiological models, and the deep influence his basic ideas either have had or should have on the interpretation of experiments directed towards understanding the different roles of the hippocampus and neocortex in associative memory. Fortunately for the field of computational vision, but unfortunately for the neurobiology of memory, Marr turned his attention away from these problems before completing his theory with a model for the input-output relations between hippocampus and neocortex. He clearly must have thought deeply about this issue, because a forthcoming paper on it was promised but apparently never completed. Many neurophysiologists and neuroanatomists agree that this issue represents the single most important area of almost complete ignorance in the field at present, and Marr's keen insight could very profitably have been applied to this problem. It is amusing to speculate whether, given the rather dramatic increase in our knowledge about the organization of cortical and archicortical memory systems over the past decade, Marr might have turned his attention back once again to these fundamental issues.

## REFERENCES


Ashwood TJ, Lancaster B, Wheal HV (1984): In vivo and in vitro studies on putative interneurons in the rat hippocampus: Possible mediators of feed-forward inhibition. *Brain Res* 293:279-291

Barnes CA, McNaughton BL, Mizumori SJY, Leonard BW, Lin L-H (1990): Comparison of spatial and temporal characteristics of neuronal activity in sequential stages of hippocampal processing. *Prog Brain Res* 83:287-300

Bliss TVP, Gardner-Medwin AR (1973): Long-lasting potentiation of synaptic transmission in the dentate area of the unanaesthetised rabbit following stimulation of the perforant path. *J Physiol* 232:357-374

Bliss TVP, Lmo T (1973): Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetised rabbit following stimulation of perforant path. *J Physiol* 232:331-356

Buszáki G (1984): Feed-forward inhibition in the hippocampal formation. *Prog Neurobiol* 22:131-153

Castro CA, Silbert LH, McNaughton BL, Barnes CA (1989): Recovery of spatial learning deficits after decay of electrically induced synaptic enhancement in the hippocampus. *Nature* 342:545-548

Collingridge GL, Kehl SJ, McLennan H (1983): Excitatory amino acids in synaptic transmission in the Schaffer collateral-commissural pathway of the rat hippocampus. *J Physiol (Lond)* 334:33-46

Douglas RM, McNaughton BL, Goddard GV (1983): Commissural inhibition and facilitation of granule cell discharger in fascia dentata. *J Comp Neurol* 219:285-294

Fishbein W, Gutwein BM (1977): Paradoxical sleep and memory storage processes. *Behav Biol* 19:425-464

Gustafsson BH, Wigström WC, Abraham WC, Huang Y-Y (1987): Long-term potentiation in the hippocampus using depolarizing current pulses as the conditioning stimulus to single volley synaptic potentials. *J Neurosci* 7:774- 780

Harris EW, Ganong AH Cotman CW (1984): Long-term potentiation in the hippocampus involves activation of N-methyl-D-aspartate receptors. *Brain Res* 323:132-137

Hebb DO (1949): *The Organization of Behavior*. New York: John Wiley and Sons

Horne JA and McGrath MJ (1984): The consolidation hypothesis for REM sleep function : Stress and other confounding factors—a review. *Biol Psychol* 18: 165-184

Kelso SR, Ganong AH, Brown TH (1986): Hebbian synapses in hippocampus. *Proc Natl Acad Sci USA* 83:5326-5330

Kohonen T (1972): Correlation matrix memories. *IEEE Transactions on Computers*, C-21, Verlag

Kohonen T (1978): *Associative Memory: A system theoretic approach.* New York: Springer-Verlag

Larson J, Wong D, Lynch G (1986): Patterned stimulation at the theta frequency is optimal for the induction of hippocampal long-term potentiation. *Brain Res* 368:347-350

Leconte P, Bloch V (1971): Déficit de in rétention d'un conditionnement après privation de sommeil paradoxal chez le rat. *C R Acad Sci (D) (Paris)* 273:86-88

Leonard B (1990): The contribution of velocity, spatial experience, and proximal visual complexity to the location-and direction-specific discharge of hippocampal complex-spike cells in the rat. Unpublished doctoral dissertation, University of Colorado, Boulder

Lømo T (1966): Frequency potentiation of excitatory synaptic activity in the dentate area of the hippocampal formation. *Acta Physiol Scand (Suppl)* 68:128

Marr D (1969): A theory of cerebellar cortex. *J Physiol* 202:437-470

Marr D (1970): A theory for cerebral cortex. *Phil Trans Roy Soc B* 176:161-234

Marr D (1971): Simple memory: theory for archicortex. *Phil Trans Roy Soc B* 262:23-81

McNaughton BL, Barnes CA (1990): From cooperative synaptic enhancement to associative memory: bridging the abyss. *Sem Neurosci* (in press)

McNaughton BL, Barnes CA, O'Keefe J (1983): The contributions of position, direction, and velocity to single unit activity in the hippocampus of freely-moving rats. *Exp Brain Res* 52:41-49

McNaughton BL, Barnes CA, Rao G, Baldwin J, Rasmussen M (1986): Long-term enhancement of hippocampal synaptic transmission and the acquisition of spatial information. *J Neurosci* 6:563-571

McNaughton BL, Douglas RM, Goddard GV (1978): Synaptic enhancement in fascia dentata: co-operativity among coactive afferents. *Brain Res* 157:277-293

McNaughton BL, Morris RGM (1987): Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends Neurosci* 10:408-415

McNaughton BL, Nadel L (1990): Hebb-Marr networks and the neurobiological representation of action in space. In: *Neuroscience and Connectionist Theory*, Gluck MA, Rumelhart DE, eds. New Jersey: Lawrence Erlbaum Associates

Mizumori SJY, McNaughton BL, Barnes CA (1989a): A comparison of supramammillary and medial septal influences on hippocampal field potentials and single-unit activity. *J Neurophysiol* 61:15-31

Mizumori SJY, McNaughton BL, Barnes CA, Fox KB (1989b): Preserved spatial coding in hippocampal CA1 pyramidal cells during reversible suppression of CA3 output: evidence for pattern completion in hippocampus. *J Neurosci* 3915-3928

O'Keefe J (1976): Place units in the hippocampus of the freely moving rat. *Exp Neurol* 51:78-109

O'Keefe J, Conway DH (1978): Hippocampal place units in the freely moving rat: why they fire where they fire. *Exp Brain Res* 31:573-590

Pavlides C, Winson J (1989): Influences of hippocampal place cell firing in the wake state on the activity of these cells during subsequent sleep episodes. *J Neurosci* 9:2907-2918

Smith C (1983) Sleep states and learning: a review of the animal literature. *Biobehav Rev* 9:157-168

Scharfman HE, Sarvey JM (1983): Inhibition of post-synaptic firing in the hippocampus during repetitive stimulation blocks long-term potentiation. *Soc Neurosci Abstr* 9:677

Wigström H, Gustafsson B (1983): Large long-lasting potentiation in the dentate gyrus *in vitro* during blockade of inhibition. *Brain Res* 275:153-158

Wigström H, Gustafsson B (1984): A possible correlate of the postsynaptic condition for long-lasting potentiation in the guinea pig hippocampus in vitro. *Neurosci Lett* 44:327-332

Wigström H, Gustafsson B, Huang Y-Y, Abraham WC (1986): Hippocampal long-term potentiation is induced by pairing single afferent volleys with intracellularly injected depolarizing current pulses. *Acta Physiol Scand* 126:317-319

Willshaw DJ, Buckingham JT (1990): An assessment of Marr's theory of the hippocampus as a temporary memory store. *Phil Trans Roy Soc B* (in press)

*Professor,*
*Division of Neural Systems, Memory and Aging,*
*University of Arizona, Tucson, Arizona 85724*

*Proc. Roy. Soc. Lond.* B. **176**, 161–234 (1970)

# 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 $\lambda$, 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 $\lambda$ over $\mathfrak{A}$ induces a distribution $\lambda'$ over $\mathfrak{A}'$ obtained by looking only at the fibres $a_1, \ldots, a_r$. $\lambda$ is called the *projection* onto $\mathfrak{A}'$ of $\lambda$. If $\mathfrak{X}$ is a subevent space, then the phrase '$\lambda'$ is the distribution induced over $\mathfrak{X}$ by the input' refers to the $\lambda'$ induced from the full input probability distribution $\lambda$ by projecting it onto $\mathfrak{X}$. If $\mathfrak{B}$ is any subset of $\mathfrak{A}$, then the *restriction* $\lambda|\mathfrak{B}$ of $\lambda$ 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.

$\{ \mid \}$ 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 $F$,
$E \cap \mathrm{F}$ = the intersection of $E$ and $F$,
$E \backslash F$ = the set of elements which are in $E$ but not in $F$,
$E \triangle \mathrm{F}$ = the set of elements which are in exactly one of $E$ and $F$,
$E \subseteq \mathrm{F}$ 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 $\mu$, $\nu$ 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 $\mu$ given $\nu$ 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 $\mu_i$ with weights $w_i$, is

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

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 $\sigma$ 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 $\mu$ be the probability distribution which describes what the memory expects. Then the amount of information the memory requires to store $\sigma$ 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 $\sigma$, are related by the following result. Assuming the memory can store $\sigma$, then:

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

*Proof.* If the memory can store $\sigma$, $\mu(E) = 0 \Rightarrow \sigma(E) = 0$, 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 $\sigma$. It can either store it straight, at cost $h(\sigma:\mu)$, or it can change its internal structure to a new distribution, $\mu'$ say, and store the signal relative to that. The amount of information required to change the structure from $\mu$ to $\mu'$ 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 $\mu'$ 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 $\mu$ 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 $\lambda$ 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.

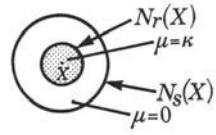



FIGURE 1. An illustration of the form of redundancy being discussed: the probability distribution $\mu$ induced by the environment over $N_s(X)$ has non-zero values only in $N_r(X)$.

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 $\mu$ 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 $\kappa$ 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 $\lambda$ 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 $\lambda$), 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 $\Omega_i$ 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 $\Omega_i$ 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 $\Omega_i$ 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 $\Omega_i$ 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 $\Omega_i$, then $E_j$ is probably also an $\Omega_i$. 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 $\Omega$, is how many of the events like $E$ are definitely known to be an $\Omega$. 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 $\Omega$, therefore constitutes *evidence* that $E$ should also be classified as an $\Omega$. It will be clear that the more such events there are, the stronger the case for classifying $E$ as an $\Omega$.

It is appropriate to make two general remarks about evidence. The first concerns the absolute weight of evidence provided by $\Omega$-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, $\Omega$.

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

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

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 $\Omega$, (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 $\Omega$: 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 $\Omega$, 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 $\Omega$ given that $c = 1$.

*Definition.* The pair $\langle c, P(\Omega|c)\rangle$ is called the *evidence* for $\Omega$ 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 $\theta$ of

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

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

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 $\Omega$ 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 $\Omega$? 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 $\Omega$ 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 $\Omega$ has occurred. The system wishes to know what is the probability that $\Omega$ 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 $\Omega$. 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 $\quad \sum_i I(T_i|T) \geqslant \sum_i I(T_i|T_0)$,

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 $\Omega$ 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 $\Omega$, 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 $\Omega$, 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 $\Omega$ 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 $(R, \theta)$-codons, and most of the evidence for a single classificatory unit has the same values of $R$ and $\theta$.

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* $\Omega$

$\mathcal{P}$ 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 $\Omega$-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 $\Omega$, there exist special regions in which for some reason, $\Omega$ does not hold. Provided the region in which $\Omega$ does not hold is itself a cluster within the larger $\Omega$-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 $\Omega$, by collecting evidence for 'not $\Omega$'—evidence against $\Omega$—within either $\mathfrak{X}$, or a space related to $\mathfrak{X}$. The form of the analysis is exactly the same as for $\Omega$, except that the classificatory unit for 'not $\Omega$' must be capable of overriding that for $\Omega$. It is of course important for the successful diagnosis of $\Omega$ that diagnostic spaces for $\Omega$ and for 'not $\Omega$' should both be appropriate, and both have evidence functions of suitable quality: but the mechanism which discovers the diagnostic space $\mathfrak{X}$ for $\Omega$ can clearly be used to discover the appropriate space for 'not $\Omega$'.

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 $\Omega$. 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 $\Omega$. 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 $\lambda$ 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 $\Omega$ 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 $\Omega$. 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 $\Omega$ 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 $c_1 \in \mathfrak{P}_1$ sends a synapse to a given cell $c_2 \in \mathfrak{P}_2$ 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 $c_2 \in \mathfrak{P}_2$ 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 $c_2$ 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

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

### 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 $\mathfrak{J}$ be a set of events of $\mathfrak{X}$—for example, $\mathfrak{J}$ may be the set of events with the property $\Omega$. Let $\mathfrak{V}_r$ be the collection of all subsets of $\{a_1, \ldots, a_N\}$ of size $r$.

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

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

*Proof.* If $\mathfrak{V}_r$ discriminates $\mathfrak{J}$ from $\mathfrak{X}$, any $\mathfrak{W}_r$ s.t. $\mathfrak{W}_r \supset \mathfrak{V}_r$ also discriminates $\mathfrak{J}$ from $\mathfrak{X}$. If $\mathfrak{J}$ can be discriminated by $\mathfrak{V}_r$, then $\mathfrak{J}$ 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 $\mathfrak{V}_r$. Finally, $\mathfrak{J}$ is always discriminated by $\mathfrak{V}_L = \{E | E \in \mathfrak{X}\}$. Hence there exists a unique lower bound $R$ s.t. $\mathfrak{J}$ is discriminated from $\mathfrak{X}$ by all $\mathfrak{V}_r$ for $r \geqslant R$.

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

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

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 $\lambda$ 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 $\lambda$).

3.2.2. Let $\mathfrak{F} = \{E | (L - |E \cap F|) < \rho\}$ where $F$ is some fixed event of $\mathfrak{X}$, and $\rho$ is a positive integer. That is, $\mathfrak{F}$ is the $\rho$-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 $\rho$-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 $\Omega$ be the

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

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

(that is $\rho$ 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 $\rho$-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 $(R,\theta)$-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 $\Omega$ 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 $\Omega$, 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 $\Omega$. 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 $\Omega$.

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 $\Omega$.

4.1.3. *The plausibility range*

Let $[0, q]$ be the range of relevance for evidence associated with $\Omega$. 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 $\Omega$ extremely unlikely. Let $p$ be that value of $\mathcal{P}(\Omega|E)$ at and below which it is impossible that $E$ ever is an $\Omega$; then $[p, q]$ is called the *plausibility range* associated with $\Omega$, 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 $\Omega$ 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 $\mathbf{\Omega}$ 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 $E$. $\mathbf{\Omega}$ 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 $\Omega$.

### 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 $\nu$, 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

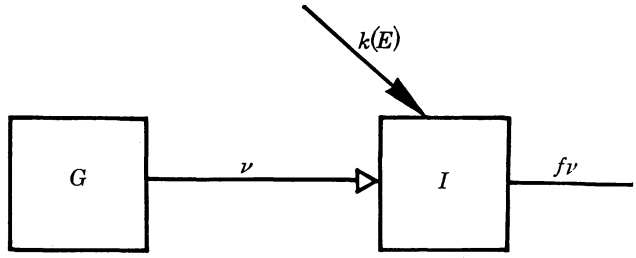



FIGURE 2. A model for division. The spike generator $G$ emits spikes at a rate $\nu$ and the inhibitor $I$ allows a fraction $f$ to be transmitted, where $f \propto k^{-1}(E)$. Spikes are therefore emitted at a rate $f\nu \propto \nu k^{-1}(E)$.

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 $\Omega$ 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 $\Omega$ 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 $\Omega$ 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.

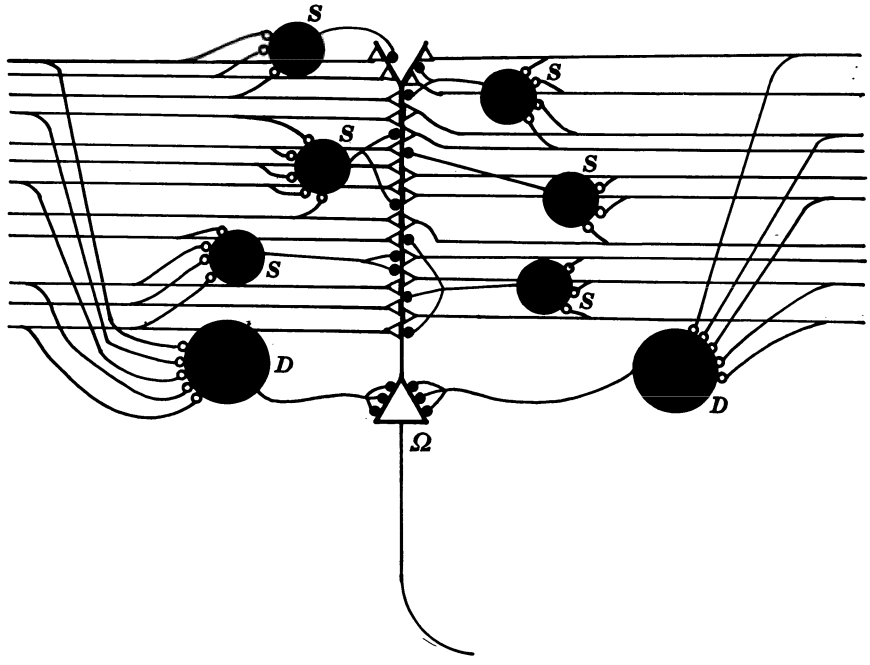



FIGURE 3. The output cell **Ω** has three kinds of afferent synapse: Hebb synapses (open triangles) from evidence cells, and two kinds of inhibitory synapse. Those from the ***S***-cells are spread over the dendritic tree, and perform a subtraction: those from the ***D***-cells, concentrated at the soma, 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 $\Omega$ 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 $\Omega$ 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 $(R, \theta)$-codons. First, observe that figure 4 shows the simplest kind of afferent

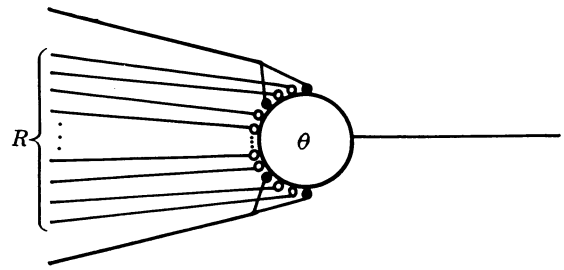



FIGURE 4. An $(R, \theta)$-codon cell. There are $R$ excitatory afferent synapses (open circles), and enough inhibition (filled circles) to give the cell a threshold of $\theta$.

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 $\theta$, which may be determined by some suitably arranged inhibition. Then the cell will emit a signal whenever at least $\theta$ of the $R$ fibres $a_{i_1}, \ldots, a_{i_R}$ are active: hence the set of firing conditions for the cell constitutes an $(R, \theta)$-codon on any event space over fibres which include $a_{i_1}, \ldots, a_{i_R}$. An $(R, \theta)$-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 $(R, \theta)$ can be chosen to construct an $(R, \theta)$-codon which will match any previously specified quality of evidence. Hence the second condition is fulfilled by the family of $(R, \theta)$-codons. The various technical problems which arise when one tries to design a net which will produce $(R, \theta)$-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 $\Omega$, and let $\Omega$ 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 $\Omega$ 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 $(R, \theta)$-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 $(R, \theta)$-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 $(R, \theta)$-codon cells, where the values of $R$ and $\theta$ 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 $\theta$ synapses at each cell of $\mathfrak{P}$, where $\theta$ is the $\theta$ of the $(R, \theta)$-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.

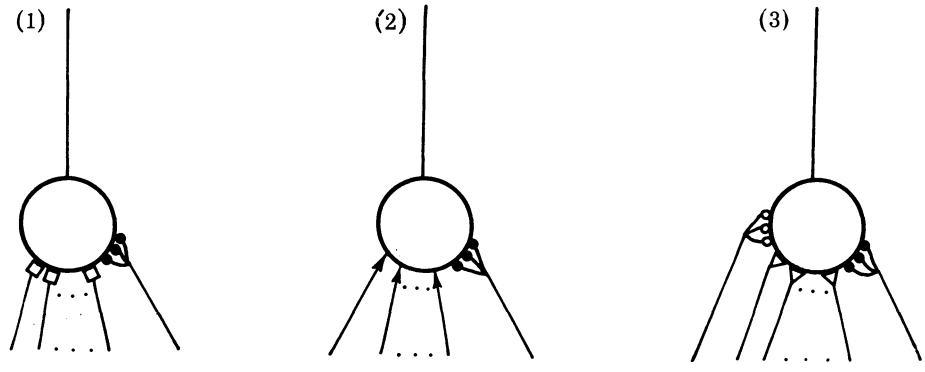



FIGURE 5. Three models for codon formation. (1) Uses synapses which are initially excitatory, but are modified to be ineffective by post- without pre-synaptic activity (open squares), (2) uses Brindley synapses (arrows), (3) uses Hebb synapses (open triangles) and a climbing fibre (open circles). All 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 $\theta$ *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 $\Omega$ 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 $\mathbf{\Omega}$ 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 $\mathbf{\Omega}$ 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 $\mathbf{\Omega}$, 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 $(R, \theta)$-codons are chosen as evidence functions for diagnosis of the property $\Omega$, 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 $\Omega$ 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 $\mathbf{\Omega}$; 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 $\Omega$ 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 $\nu$: then $k^{-1}$ has distribution $\nu^{-1}$ say, with expectation $\overline{k^{-1}}$ ($\neq \bar{k}^{-1}$ in general), and variance $\sigma$ (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 $\sigma$ 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 $\sigma$ and $\sigma/\sqrt{\{n(c_i)\}}$, but since $\sigma$ 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 $R, \theta$ and $r$, where $c_i$ is an $(R, \theta)$-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 $(R, \theta)$-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 $(R, \theta)$-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 $(R, \theta)$-codon to some suitable lower value $\theta'$, 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 $(R, \theta)$-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 $\theta'$, and $z = W/N$ (by analogy with 3.1.2). This is just the usual tail of a binomial distribution. Now as $\theta'$ decreases, the number of $(R, \theta)$-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 $\theta'$ is lowered, the number of $c_i$-cells which $X$ fires increases very fast: so that the difference in $\theta'$ 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 $\theta'$ active afferent synapses.

The problem of the differential weighting of the $P(\Omega|c_i)$ can thus be alleviated as long as $\theta'$ 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 $c_i$-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 $c_i$-cells active constant is adequate for this task. If this technique is used, those few $c_i$-cells with more than $\theta'$ active afferents will have a higher firing rate than those with exactly $\theta'$. Hence they will anyway be given greater weighting at the $c_i$-cell. It would be optimistic to suppose this weighting would be exactly the correct amount, since the factor involved depends on the parameters $N, L, W, R, \theta, r$; but the effect will certainly reduce the errors involved.

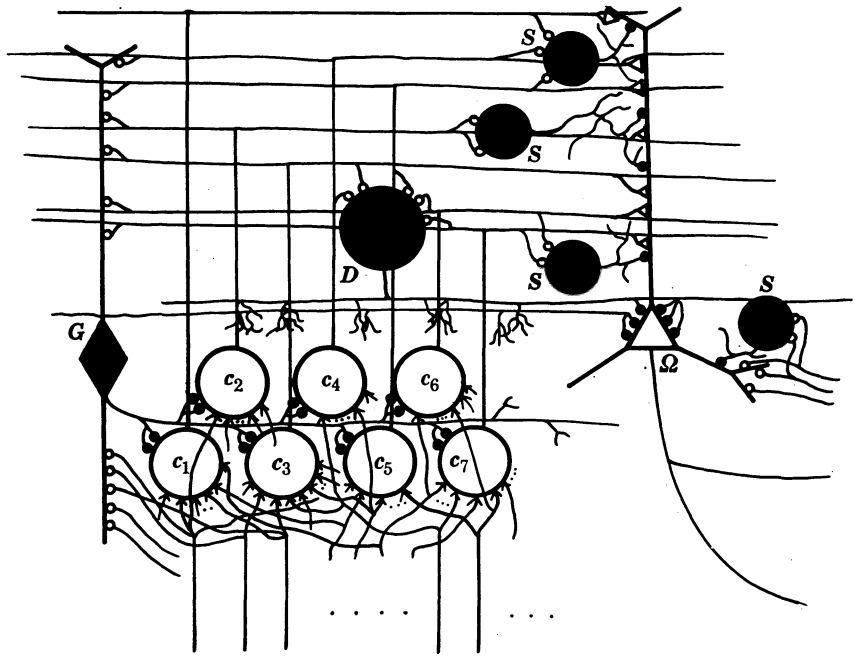



FIGURE 6. The basic neural model for diagnosis and interpretation. The evidence cells $c_1, \ldots, c_7$ are codon cells with Brindley afferent synapses. The $G$-cell controls the codon cell threshold: it uses negative feedback through its ascending dendrite to keep the number of codon cells active roughly 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 $c_i$-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 $G$ control the thresholds of the $c_i$-cells, and their function is to keep the number of active $c_i$-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 $a_j$, and by the $c_i$-cell axons. The final control should be exercised by the number of $c_i$-cell axons active, but a direct input from the $a_j$ axons would provide a fast route for dealing with a sudden increase in the size of the input event.

The $c_i$ axons and the output cell $\Omega$ 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 $\Omega$ is shown with two types of evidence cell afferent: one, through the $c_i$-cells to the apical dendrites, and one (whose origin is not shown) to a basal dendrite.

In practice, the distribution of the $a_j$ 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 $\Omega$-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 $\boldsymbol{\Omega}$, 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 $\boldsymbol{\Omega}$ 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).

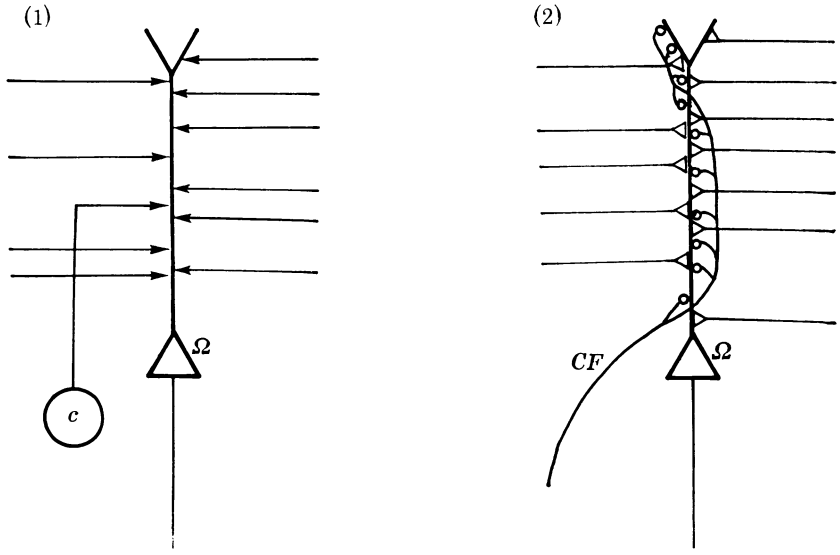



FIGURE 7. Two models for output cell selection. (1) Uses Brindley synapses, (2) uses Hebb synapses and a climbing fibre (*CF*).

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

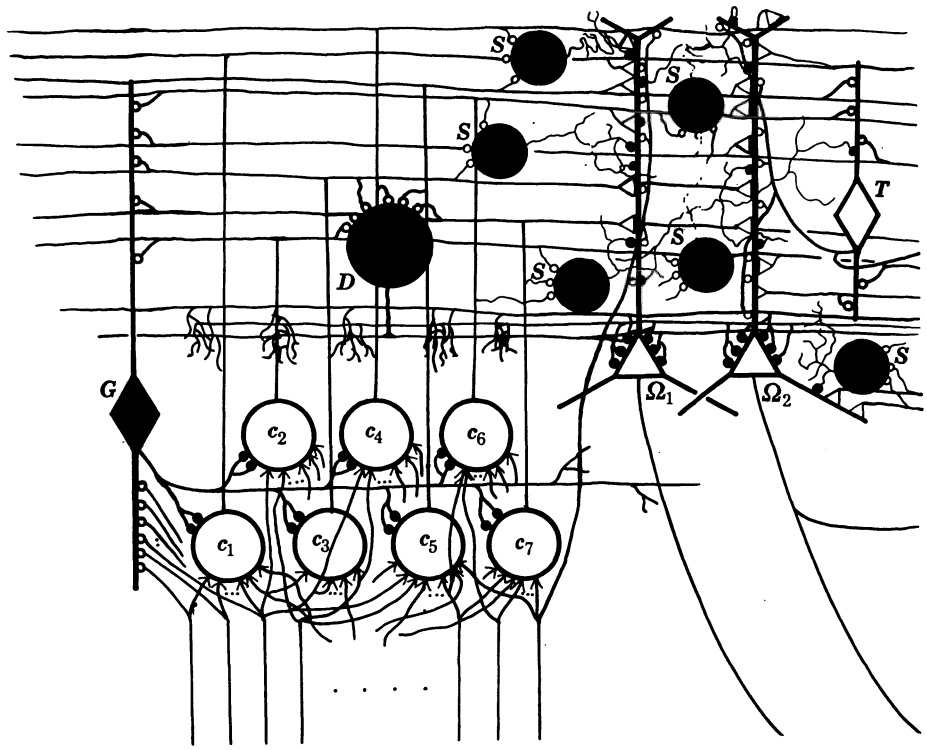



FIGURE 8. The fundamental neural model, obtained by combining the models of figures 6 and 7(2). Two climbing fibres are shown; one from an input fibre, and one from a nearby output selector cell ***T***.

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 $\lambda^*$ (say) over $\mathfrak{Y}$. $\lambda^*$ is the probability distribution the environment induces over $\mathfrak{Y}$, and is derived from the input distribution $\lambda$ over $\mathfrak{X}$. Both $\lambda$ and $\lambda^*$ 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 $\mu$ 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 $\mu$ is a *standard (k, M)-plateau* over $c_1, \ldots, c_M$.

That is, $\mu$ 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* $\mu^*$ is one that can be built up out of plateaux $\mu_i$ 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. $\phi$ cannot in general be regarded as a function from $\mathfrak{Y}$ to $\{0, 1\}$ since $\phi$ may take the value 1 at a time when there is no event in $\mathfrak{Y}$. Some kind of relation between $\phi$ 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 $c_i$ to $\boldsymbol{\Omega}$, and on this view, the strength varies with $P(\phi|c_i)$. The cell $\boldsymbol{\Omega}$ really diagnoses $\phi$ if this is so, but it will be shown in §5.2.3 that if the structure of $\lambda^*$ over $\mathfrak{Y}$ is appropriate, this will be adequate.

The other possible view is that $\phi$ acts as a pointer for $\boldsymbol{\Omega}$. On this model, the effect of $\phi$ 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 $\mu$ denote the standard $(k, M)$-plateau on $\{c_1, \ldots, c_M\}$, $M < K$, and let $\nu$ 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 $\mu$ from the distribution $\lambda^* = \frac{1}{2}(\mu + \nu)$ even though the climbing fibre $\phi$ did not. This is the crucial property which the system possesses.

In general, if $\phi = c_1$, $\phi$ 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 $\phi$ 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 $\lambda$ 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, $\phi$ 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 $\Omega$. 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 $\Omega$ 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 $\lambda$, *and* on those events to which (by (2)) $\Omega$ 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 $\omega_1^i$ of $\boldsymbol{\omega}_1$ are simply the projections onto the $c_i$ of the restriction $\lambda | N_\theta(\boldsymbol{\omega}_0)$ of $\lambda$ to $N_\theta(\boldsymbol{\omega}_0)$. Then $\boldsymbol{\omega}_1$ represents the state of the synapses from the $c_i$ to $\Omega$ if $\Omega$ responds only to the events in $N_\theta(\boldsymbol{\omega}_0)$.

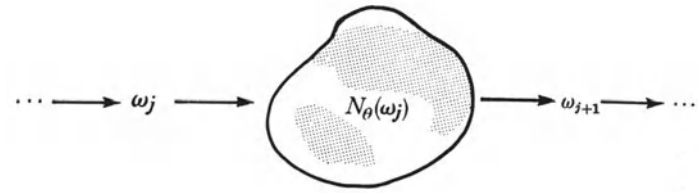



FIGURE 9. The state vector $\boldsymbol{\omega}_j$, which describes the strengths of the afferent synapses to the output cell $\Omega$, determines the set $N_\theta(\boldsymbol{\omega}_j)$ of events to which $\Omega$ will respond. This in turn determines a new state vector $\boldsymbol{\omega}_{j+1}$.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

The situation is thus that in the state $\boldsymbol{\omega}_0$, the cell $\Omega$ 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 $\Omega$, 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 $\Omega$ 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 $\Omega$ 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 $\mu^*$, $\mu^* = w_1\mu_1 + \ldots + w_\rho\mu_\rho$, 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 $\lambda^*$. 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 $\lambda$. The answer that it does if and only if a line of steepest ascent leads

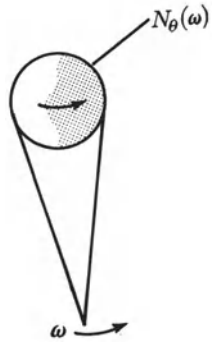



FIGURE 10. The state vector $\boldsymbol{\omega}$ determines the set $N_\theta(\boldsymbol{\omega})$ of events to which $\boldsymbol{\Omega}$ responds. The environmental probability distribution over $N_\theta(\boldsymbol{\omega})$ is stippled where it has non-zero values. $\boldsymbol{\omega}$ changes 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 $\lambda$ 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 $\boldsymbol{\omega}_1 = (\omega_1^1, \ldots, \omega_1^K)$, $\omega_1^i$ varies with the expected probability that $c_i = 1$ in $N_\theta(\boldsymbol{\omega})$ under $\lambda$. 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 $\Omega$ 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 $\Omega$ 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 $(mR, \theta)$-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 $\mu$m 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.

# Jack D. Cowan

## Commentary on

## *A Theory for Cerebral Neocortex*

David Marr's paper "A Theory for Cerebral Neocortex" published in 1970, is the third in a trilogy of papers, the first being "A Theory of Cerebellar Cortex"; the second, "Simple Memory: A Theory for Archicortex" was published in 1971. The three papers are closely related, and it is difficult to comment on the 'neocortex' paper without some discussion of the others, particularly of the 'cerebellum' paper.

The first paper proposes that the cerebellar cortex is a network that learns associations. It was once (mistakenly) characterized to me by Minsky and Papert as a perceptron-like theory, but in fact it is very similar to the associative nets first studied by W.K. Taylor (1956), which can be trained to associate and recognize classes of patterns (see Fig. 1).

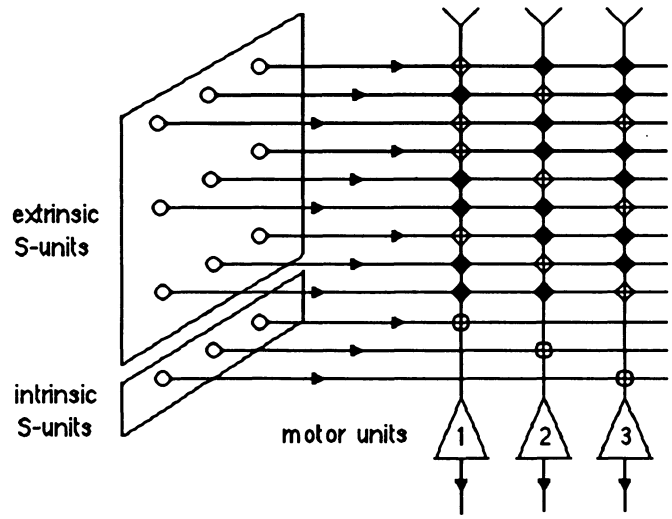



Fig. 1. Taylor net. This net uses analog neurons with modifiable weights and can be trained to associate differing sets of stimulus patterns via synaptic facilitation.

The basic neural machinery is that of synaptic facilitation, either heterosynaptic, in which coincident activation of weak extrinsic synapses and a strong

suprathreshold intrinsic synapse leads to strengthened extrinsic synapses, or else homosynaptic, in which correlated pre- and postsynaptic activity strengthens the synapse (see Fig. 2).

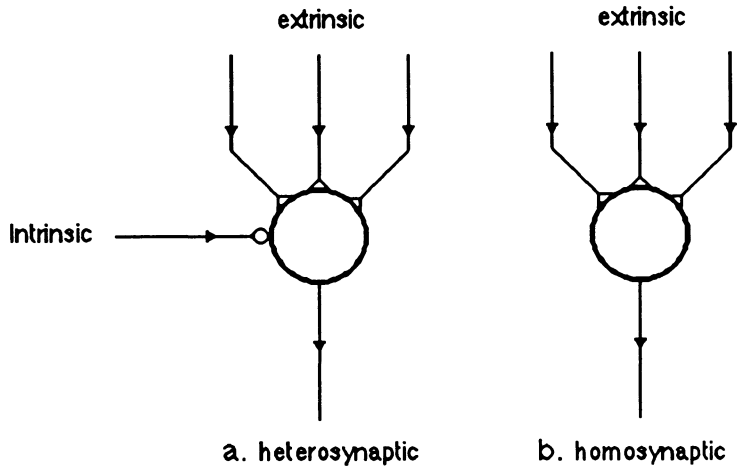



Fig. 2. Facilitation of extrinsic synapses (a) by coincident activation of an Intrinsic synapse, and (b) by correlated activation of extrinsic synapses and the post-synaptic cell.

This machinery was used by many investigators throughout the 1960s, but it was Marr who first recognized the need to encode input patterns carefully for optimal storage, and to control carefully the retrieval of stored patterns. Marr's work is closely related to a little known paper by P. H. Greene (1965) entitled "Superimposed Random Coding of Stimulus-Response Connections," itself based on earlier work by C. N. Mooers (1949,1954) involving punched cards. Greene introduced the network shown in Figure 3 for the storage and retrieval of information. The basic idea is that forming random subsets of incoming patterns (in modern terms a form of "hash" encoding) provides an economical and efficient representation of the information contained in the patterns, and also minimizes any overlap between differing patterns. The similarity between this network and Marr's model of the cerebellar cortex is striking. It requires only inhibitory neurons to control thresholds, and intrinsic climbing fibers to train synapses (both suggested but not implemented by Greene), to complete the picture. Greene recognized the associative memory aspects of the network but did not pursue this, although he suggested that recurrent connections from the motor units back to the sensory units might allow the network to generate sequences of stored responses. It was Marr who introduced climbing fibers and inhibitory interneurons to control both

the storage and the retrieval of information in such an architecture, and who followed up a suggestion of J. Szentagothai's (c. 1964), that the cerebellar cortex learns.

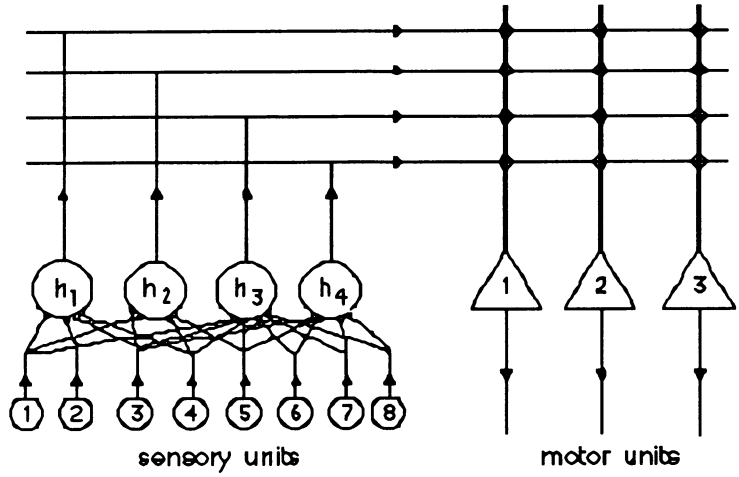



Fig. 3. Greene's network, a neural realization of Mooers' "Zato-coding" scheme for storing and retrieving multiple descriptors on punched cards.

Marr's theory of the hippocampus differs from that of the cerebellar cortex in several respects. First, the granule cell synapses are themselves modifiable via homosynaptic rather than heterosynaptic facilitation, because internal representations of novel incoming sets of patterns have to be formed. Marr showed that this could only be performed reliably with two layers of granule cells. Second, the output pyramidal cells are excitatory and make recurrent connections with the first layer of granule cells, as well as with neocortical cells; they are also interconnected via modifiable excitatory synapses. This permits the recall of patterns from only a subset of cues. The entire network functions as an associative memory. Interestingly, the addition of excitatory interconnections, together with inhibitory interneurons for threshold control, gives the hippocampus model dynamical properties that even now have not been investigated.

The overall structure of the hippocampal model is shown in Figure 4. In modern terms it comprises a recurrent network with two layers of trainable "hidden" neurons to provide efficient and reliable encoding and classification of sets of input patterns connected to an associative memory store. The cerebellum model can be seen as a specialization of this comprising only a single layer of granule cells with fixed synapses, and an associative memory store

with no recurrent connections, and with another set of conditioning inputs, that is, the climbing fibers.

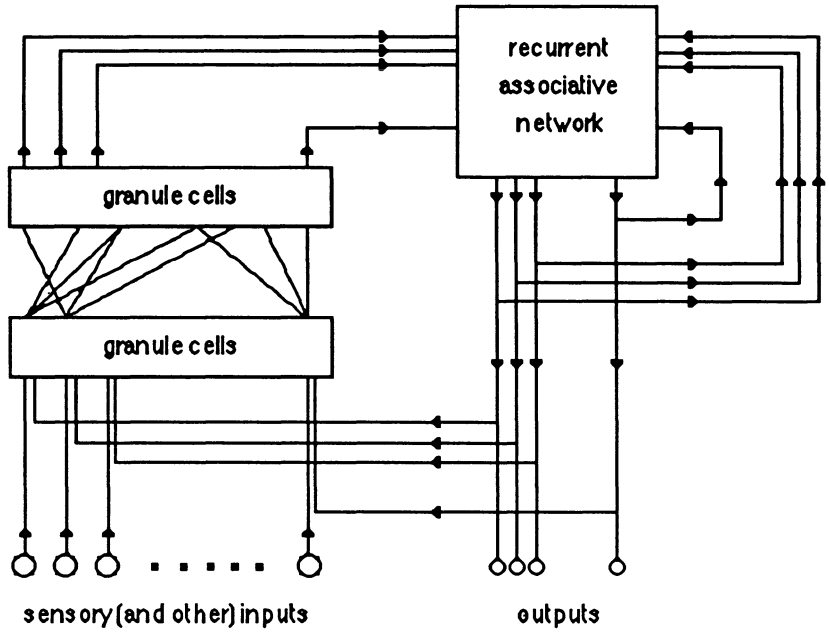



Fig. 4. Marr's model of the hippocampus and associated networks.

The neocortex model is a composite comprising elements of both the cerebellar cortex and hippocampus models. It tackles a much more complicated problem, that of the *ab initio* formation and organization of networks capable of classifying and representing "the world." The cerebellar and hippocampus models both assume this development to have already taken place: in the neocortex model Marr tries to solve this problem. In current terms it represents an early attempt at a theory of unsupervised learning, essentially using some methods of cluster analysis borrowed from numerical taxonomy. The basic idea stems from a broader view of the circuitry shown in Figure 2(A) in which "extrinsic" synapses are facilitated by the coincident activation of a strong "intrinsic" synapse. In Marr's more general view, intrinsic properties are those already learned and utilized by the organism, whereas extrinsic properties are novel. The intrinsic inputs may therefore be thought of as indicating or diagnosing the membership of an extrinsic activation pattern in some class of patterns already learned by the organism, and the synaptic weights themselves may be thought of as indicating the conditional probabilities $P(\Omega|c_i)$ of the extrinsic event $c_1, c_2, c_n$ being in the class $\Omega$. Interestingly, conditional probabilities (and mutual information measures) were hypothesized by A. M.

Uttley some 10 to 20 years earlier (c. 1954-1966), to be computed at synapses. However, Marr's use of conditional probability differs fundamentally from Uttley's in that it does not refer to the frequency of occurrences of events, but rather to their overlap or similarity with intrinsic events. Figure 5 shows the resulting circuit. It embodies what Marr called the diagnosis theorem, namely that for a given evidence cell activated by the event $E = \{c_1, c_2, \ldots, c_n\}$, the best (maximum likelihood) estimate of the probability that $E$ is in the class $\Omega$, is given by the arithmetic mean of its synaptic weights. Marr also proved a related theorem, the interpretation theorem, in case the extrinsic events are only partial subsets of $\{c_1, c_2, \ldots, c_n\}$. Let $X$ be such a subevent, and let $E_i$ be all the possible events that can contain $X$, the completions of $X$. Then the best estimate of $P(\Omega|X)$ is given by the arithmetic mean of the conditional probabilities $P(\Omega|E_i)$, again stored as synaptic weights in a set of evidence cells. This is implemented by the following circuit shown in Figure 6. The similarity between this circuit and the cerebellar cortex circuit described earlier is apparent.

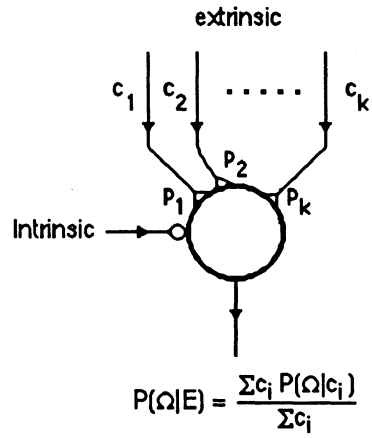



Fig. 5. Granule cell as an "evidence" cell. Its firing rate is proportional to $P(\Omega|E)$ where $E$ is the extrinsic event $\{c_1, c_2, \ldots, ck\}$, and $\Omega$ is a given set of events, i.e., a "class." Note that $P(\Omega|E)$ is the arithmetic mean of the individual synaptic weights $p_i = P(\Omega|c_i)$.

But can the ideas of class formation and so forth work in general? Marr's answer is no, but they will work if the world is spatially coherent, for then classification in terms of similarity is possible, and generalization. What is required is a mechanism by which novel classes can be discovered and developed. This is the basic problem of unsupervised learning. It can be solved by self-organization with homo- and heterosynaptic facilitation acting

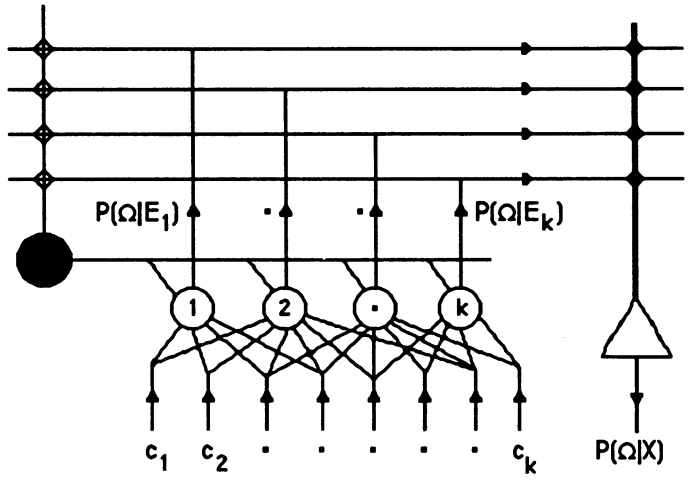



Fig. 6. A neural implementation of the interpretation theorem. A subset $X$ of the event $\{c_1, c_2, \ldots, c_k\}$ activations evidence cells whose thresholds are controlled by an inhibitory interneuron (shown in black), to maintain a roughly constant input to the output cell. Other inhibitory cells controlling the output cell threshold, and providing normalization via $\Sigma c_i$ are not shown.

to strengthen synapses when spatially correlated and coherent inputs are presented. It follows that the machinery outlined above, and described in detail in Marr's papers, will work to discover and develop classifications if the world is reasonably coherent. Marr provided some interesting general insights into these notions, in drawing a parallel with the cluster methods used in numerical taxonomy, in which the differences between objects are first computed, and then clusters or classes are formed that minimize some average measure of these differences. As Marr showed, the network that implements the diagnostic and interpretation theorems can be seen as performing similar computations. Thus, the circuit of Figure 6, when completed with intrinsic inputs to direct or bias the formation of evidence cells and of output cells signaling class membership, and the various interneurons not described here, constitute the first neural implementation of something like a cluster method for the discovery and representation of classes. The more recent work of Kohonen (1982) and Grossberg and Carpenter (1986) should be read in the light of Marr's neocortex paper.

## REFERENCES

Carpenter GA, Grossberg S (1986): Neural dynamics of category learning and recognition: attention, memory consolidation, and amnesia. In: *Brain Structure, Learning and Memory.* AAAS symposium series, Davis J, Newburgh R, Wegman E, eds.

Greene PH (1965): Superimposed random coding of stimulus-response connections. *Bull Math Biophys* 7:191-202

Kohonen T. (1982): Self-organized formation of topologically correct feature maps. *Biol Cybern* 43 1: 599-570

Mooers CN (1949): Application of random codes to the gathering of statistical information. *Zator Co Tech Bull* 31

Mooers CN (1954): Choice and chance coding in information retrieval systems. Trans. IRE Prof. Group on Inf. Theory, *PGIT-4*: 112-118

Taylor WK (1956): Electrical simulation of some nervous system functional activities.In: *Information Theory*, 3, (Ed.) Cherry EC London: Butterworths, 314-328

Taylor WK (1964): Cortico-thalamic organization and memory. *Proc Roy Soc Lond B* 159:466-478

Uttley AM (1956): A theory of the mechanism of learning based on the computation of conditional probabilities. *Proc 1st Int Conf on Cybernetics*, Namur, Gauthier-Villars, Paris

Uttley AM (1959): The design of conditional probability computers. *Infor Cont* (2):1-24

Uttley AM (1966): The transmission of information and the effect of local feedback in theoretical and neural networks. *Brain Res* (2): 21-50

*Professor*
*Department of Mathematics*
*University of Chicago*
*Chicago, Illinois*

Vision Res. Vol. 14, pp. 1377–1388. Pergamon Press 1974.

# THE COMPUTATION OF LIGHTNESS BY THE PRIMATE RETINA

DAVID MARR*
The Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 545, Technology Square, Cambridge, Mass. 02139, U.S.A.

(*Received* 9 *January* 1974)

**Abstract**—It is proposed that one function of the retina is to compute lightness using a two-dimensional parallel algorithm. There are three stages: (1) a centre-surround difference operation; (2) a threshold applied to the difference signal; (3) the inverse of (1), whose output is lightness. The operation of the midget bipolar—midget ganglion channel is analysed in detail, and a functional interpretation of various retinal structures is given. Requirements of the theory are stated concerning the arrangement and connexions of cells, and the signs of the synapses, in the inner plexiform layer.

## SUMMARY

(1) It is proposed that one function of the primate retina is to compute lightness by a method derived from the two-dimensional parallel algorithm of Horn (1974).

(2) The computation consists of three stages: (a) a centre-surround difference operation, computed in approximately logarithmic units, the result being carried by the bipolar cells; (b) an approximately constant threshold applied to this signal; and (c) the inverse transform of (1), performed in the amacrine layer, whose output is lightness. Lightness probably appears at X-cells, which should therefore provide the information for subsequent colour naming.

(3) The operation of the midget bipolar midget ganglion cell channel is analysed in detail. It is shown that the small, stratified amacrine cells are well placed to carry the necessary additive lateral connexions between nearby midget bipolar terminals; and the diffuse amacrine cells, for supplying the necessary subtractive coupling between the two lateral systems in the inner and the outer thirds of the inner plexiform layer.

(4) In particular it is necessary that: (a) A large proportion of the midget bipolar dyad synapses should be with stratified amacrine cells. All synapses in such a dyad complex, including the amacrine/bipolar synapses, must have a computationally positive sign. (b) Diffuse amacrine cells must receive excitation from one layer (from midget bipolar, and possibly from stratified amacrine cells), and must send inhibitory synapses to the other layer, to midget ganglion and to stratified amacrine cells, but probably not strongly, and preferably not at all, to the midget bipolar axon terminals. The synapses from midget bipolar to diffuse amacrine cells need not be accompanied by a reciprocal amacrine/bipolar synapse, whereas those to a stratified amacrine cell should be.

(5) Midget ganglion cells, and perhaps all X-cells, should behave like detectors of lightness. Their centre-surround receptive field organization arises from suitable setting of the d.c. level of the retinal output.

(6) When the illumination falls below a certain minimum level, the lightness computation must be abandoned.

## INTRODUCTION

The assignment of subjective "lightness" and "colour" to visible surfaces is, except in especially restricting circumstances, almost independent of the prevailing illumination; and it has long been thought that our apparent sensitivity to reflectance rather than to luminance depends mainly upon the use of comparative, not absolute, measurements of luminance made by the visual system [Helmholtz, 1867 (1962)]. An anecdotal expression of this opinion may be found in the lecture by Rushton (1972, pp. 27P–31P). Methods that are capable of computing reflectance from measurements of luminance are therefore of considerable interest, and the problem may conveniently be divided into two parts: those situations in which luminance varies gradually across a portion of the visual field; and those in which it changes suddenly. Only the first of these will be discussed in this article.

Gradual changes in luminance are often due to changes in illumination rather than to changes in reflectance. The function of luminance that is obtained by removing from it those components that vary slowly, is therefore a first approximation to reflectance; and it is called "lightness". Because it is close to reflectance, lightness is useful for estimating colour,

* On leave from the M.R.C. Laboratory of Molecular Biology, Cambridge, England.

and Land and McCann (1971) have made this the basis of their work (see also Land, 1964).

Luminance is proportional to the product of reflectance and illumination, and so their logarithms are linearly related. This enabled Land and McCann (1971) to construct a program that computes lightness using essentially linear methods. Their program takes a random path across a luminance array, differentiates log luminance along the path, sets all small values to zero (which removes the effects of slow changes), and reintegrates to obtain the lightness distribution along that path. When enough random paths have been chosen to cover the array completely, the computation is complete. This method is fundamentally unsatisfactory, because situations can be constructed in which the value of the integral obtained between two points depends upon the path used. It is also unsatisfactory as a starting point from which to consider how the nervous system could perform the computation, because in primates, no directional derivatives appear to be computed at least until area 17.

The ugliness of the path-integral method lies in the use of a one-dimensional technique to solve a two-dimensional problem, and it provoked Horn (1974) to search for a suitable two-dimensional technique. To do this, one needs to find a two-dimensional isotropic differential operator, which would be the analogue of differentiating along a path; and an inverse operator, which would be the analogue of the path integral. The lowest-order such differential operator is the Laplacian, $\nabla^2$: Horn expressed $\nabla^2$ as a convolution, and found its inverse.

In order to perform an actual computation using this technique, one needs to construct a discrete version of it. The discrete approximation to the Laplacian is well-known to be a centre-surround operator with a decreasing inhibitory weighting function over the surround (see e.g. Ratliff, 1965, p. 97). The observation that is new in this context, and upon which this investigation rests, is that the discrete approximation to the inverse transform is also very simple. Because of this fact, the two-dimensional method of computing lightness is extremely straightforward and elegant, and it may be expressed in the following way. Let $x$ be the log of intensity at a point $X$ on a photosensitive surface, and let $y_1, y_2, \ldots, y_6$ be the logs of the intensities at neighbouring points $Y_1, Y_2, \ldots, Y_6$ (see Fig. 1). The first stage in the computation is to derive a local difference function using a centre-surround operator. For example, one might compute the local difference $x'$ at $X$ by using the formula

$$x' = x - 1/6 \sum_i y_i. \tag{1}$$

(This difference is computed at all points, so that $y'_1$, $y'_2$, etc. are also obtained.) The second stage of the method is to apply a threshold to the difference $x'$: this is the step responsible for removing the gradual changes, and the symbol $x''$ is used to indicate a thresholded difference signal. The third stage is the inverse of (1), and may be written:

$$x^* = x'' + 1/6 \sum_i y_i^* \tag{2}$$

where $x^*$ is the inverse obtained from $x$, and $y_i^*$ from $y_i$, as illustrated in Fig. 1. Equation (2) states that the answer $x$ at $X$ is obtained by adding the difference at $X$, $x''$, to the average of the answers $y_i^*$ at the neighbouring points. If the boundary conditions are suitable, the set (2) of simultaneous equations is the inverse of the set (1), to within a constant. It has been known for some time that a neural network with reciprocal connexions is in principle capable of solving sets of simultaneous equations of this form; [see Ratliff (1965), pp. 130–142, and especially his careful comments (p. 141) on what to expect of a mathematical model of a neural network].

*Is this method relevant to retinal function?*

The main reason for being interested in Horn's method is that the assumption that it is implemented in the retina makes sense of many of the known facts

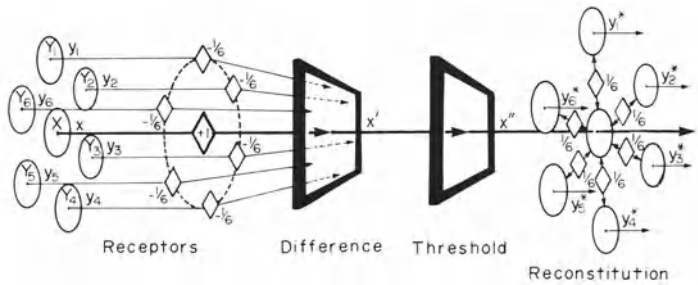



Fig. 1. The discrete form of Horn's two-dimensional method for computing lightness consists of three stages. The first is a centre-surround difference operation [equation (1)], whose output is $x'$. Next, a threshold is applied to $x'$, resulting in $x''$. Finally, the image is reconstituted in the manner described by equation (2), which is implemented here by a set of additive connexions between the solutions $x^*$, $y_i^*$ at neighbouring points. These additive connexions mirror the subtractive connexions that produced the difference signal $x'$.

about the retina, and leads to certain definite and testable predictions. There are however a number of general points that need to be raised before a detailed analysis is seen to be worthwhile. The first question is whether computing lightness is the right problem; it may be, for example, that the estimation of reflectance is in fact carried out by a method which renders irrelevant the distinction between large and small luminance gradients. The only evidence that the distinction is important is the known insensitivity of the eye to small, uniform gradients of luminance, which gives rise to several well-known illusions (see Brindley, 1970, p. 153); and the experiments of Land and McCann (1971).

The second question is, given that computing lightness is probably important, is there any sense in which Horn's method is the best one available? The reasons for choosing a two-dimensional method have already been mentioned: the Laplacian is simply the lowest order two-dimensional operator that could be used. Using it causes constant and linear terms to be removed from the signal: higher order operators, like $\nabla^4$ or $\nabla^6$, would cause higher order terms to be removed, and are therefore to be avoided. A second piece of evidence that the Laplacian is relevant is the centre-surround receptive field organisation of the retinal bipolar cells (see Werblin and Dowling, 1969; Dowling and Boycott, 1966; Boycott and Dowling, 1969). (The one-dimensional path-integral method cannot be used if the difference signal has a centre-surround form: to see this, consider a path that joins two points of equal luminance by travelling along a discontinuous boundary.)

The third question is why does the inverse transform need to be computed at all? Could the brain perhaps deal just in the difference signals? The inverse transform has the characteristic that the solution at one point affects the solution at all other points to some degree, so that to obtain an explicit solution at one point (as would be necessary to specify the colour at that point), something equivalent to the inverse needs to be computed. The fact that we are able to assign "absolute" colours to visible surfaces is therefore evidence that if we use this method at all, then we solve at least an approximation to the inverse at some stage. The hypothesis that both forward and inverse transforms are carried out at a relatively low level is attractive because it simplifies the problem that subsequent mechanisms have to solve without forcing the penalty that local signals have to be combined in a special way later on.

Finally, there is the question of how flexible is the discrete approximation represented by equations (1) and (2). In Fig. 1, the difference function was obtained from the point $X$ and its six neighbours, each with weight 1 6. There is no need to be so restrictive, however: the difference signal can be made up of any number of the local neighbours of $X$, with arbitrary weights, as long as the distribution and weights are exactly reflected in the inverse transform. Hence although it is desirable to avoid collecting higher order operators, the actual computation can tolerate considerable variation in the way it is structured locally.

The rest of this article is concerned with the consequences of assuming that the retina computes lightness, using a parallel method like that of Fig. 1. The general discrete form of the algorithm is as follows: letters like $x$, $y$ denote log of intensity at points $X$, $Y$; $N(X)$ refers to the set of points $Y_i$ in the neighbourhood of $X$; and $w$ is some weighting function on $N(X)$. The difference operation has the form:

$$x' = x - \sum_{Y in N(X)} w(y).y. \tag{3}$$

Letters with a prime, like $x'$, $y'$ refer to differences at $X$, $Y$, obtained in this way. The second step is to apply a threshold to the difference signal, by

$$\begin{aligned} x'' &= x' \text{ if } |x'| > \text{some threshold } t \text{ (say)} \\ &= 0 \text{ otherwise.} \end{aligned} \tag{4}$$

Letters followed by two primes, like $x''$, $y''$, will refer to thresholded difference signals. Finally, the inverse transform

$$x^* = x'' + \sum_{Y in N(X)} w(y) \times y^* \tag{5}$$

is applied. Letters like $x^*$, $y^*$ refer to the output from the whole process: every point $X$ gives rise to an $x^*$. By inspection, for $t = 0$ (3) and (5) are inverses for point sources (with zero boundary conditions). Hence by linearity, they are inverse transformations.

## THE ANATOMY OF THE RETINA

Most qualitative, and some quantitative aspects of the structure of the primate retina are well understood, thanks to the early work of Cajal (1911), and the recent thorough studies by Missotten (1965), Dowling and Boycott (1966), Boycott and Dowling (with Kolb) (1969), and Kolb (1970). The cell types and connexions of the outer plexiform layer are summarized in Figs. 2(a), (b) and (c) and the accompanying legend. The details of the inner plexiform layer are less widely known, and a very brief summary of the cells and synapses described by Boycott and Dowling (1969) is therefore included here. (The word "diffuse", in this context, refers to processes that are distributed perpendicular to the sclera, and the word "stratified" is used to mean layered parallel to the sclera.)

*Amacrine cells [see Fig. 2(a)]*

(A1) Narrow field diffuse amacrine cells, having a diameter of 10–50 nm, average about 25 nm, found all over the retina.

(A2) Wide-field diffuse amacrine cells, having processes that spread out gradually as they descend to the level near the ganglion cell bodies, and spread out there to attain a diameter of up to 600 nm. These cells are particularly likely to synapse with rod bipolar terminals, and are unlikely to contact the ganglion cell bodies.

(A3) Stratified diffuse amacrine cells, having a diameter of 20–50 nm, are restricted to the top, middle, or to the lower third of the inner plexiform layer, but are diffusely distributed within one of them. A given stratified diffuse amacrine

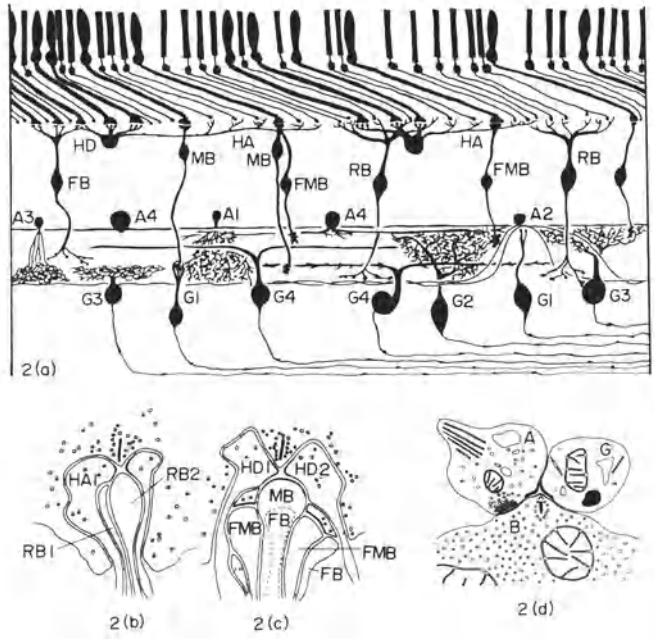



Fig. 2. (a) The general structure of the primate retina, redrawn from Boycott and Dowling (1969, Fig. 98) and Kolb (1970, Fig. 56). The cones contact the horizontal cell dendrites (HD), and three kinds of bipolar cell: midget (MB), flat midget (FMB), and flat bipolar cells (FB). The arrangement of these processes in the cone pedicles is shown in (c) (from Kolb, 1970, Fig. 60). The rods synapse with the horizontal cell axons (HA), and with rod bipolar cells (RB) in the manner shown in (b) (from Kolb, 1970, Fig. 59). The bipolar cell axons synapse with the amacrine cells (A1–A5), and with the ganglion cells (G1–G4): the different kinds of cells are described in the text. (d) (from Dowling and Boycott, 1966, Fig. 14) illustrates a dyad complex. This is composed of synapses from a bipolar cell (B) to an amacrine (A) and to a ganglion cell (G), together with a synapse back from the amacrine process to the bipolar axon. In addition, amacrine–amacrine and amacrine–ganglion synapses are seen.

cell probably makes frequent, but not exclusive contact, with a particular ganglion cell that has its dendrites similarly distributed.

(A4) Unistratified amacrine cells, whose diameter lies between 100 and 1000 nm, extend their processes in the plane immediately corneal to the inner plexiform layer.

(A5) Bistratified amacrine cells, with a diameter of about 100 nm, send horizontally distributed processes to the planes corneal and scleral to the inner plexiform layer.

*Ganglion cells* [*see Fig. 2(a)*]

(G1) The midget ganglion cells are of two kinds, one with terminals in the outer third of the inner plexiform layer, and the other with terminals in the inner third. The middle third seems to be free of midget ganglion cell terminals. This fits with the known distribution of the midget bipolar terminals (see above). There is probably a one-to-one correspondence between midget bipolar and midget ganglion cells.

(G2) Diffuse ganglion cells, dendritic diameters ranging from 30–75 nm, the smaller diameters occurring nearer the fovea.

(G3) Stratified diffuse ganglion cells, like the stratified diffuse amacrine cells, are diffuse within the outer, middle, or inner third of the plexiform layer. There may be more in the outer third than in either of the others. Diameters range from 40 nm near the fovea, to 80 nm in the periphery.

(G4) Unistratified ganglion cells, occurring at all levels, have a diameter of about 200 nm.

*Synapses of the inner plexiform layer*

The most common synaptic complex found in this region of the retina is the so-called dyad synapse [see Fig. 2(d)]. At a dyad synapse, a bipolar cell contacts both a ganglion and an amacrine cell, and close by there is (probably) a further synapse from the amacrine cell back onto the bipolar terminal (Dowling and Boycott, 1966). In addition to the dyad synaptic complex, amacrine to amacrine, and amacrine to ganglion dendrite synapses are seen.

The proportions in which the various types of synapse occur in the human retina are roughly as follows:

the complex of dyad + amacrine to bipolar: 3,
amacrine to amacrine: 1,
amacrine to ganglion: 1,
bipolar to amacrine soma: 1/12
(from Dowling and Boycott, 1966, Table 1).

## THE DIFFERENCE OPERATION

The receptor response is certainly [Kaneko and Hashimoto (1967), Tomita (1968), Naka (1969), Toyoda *et al.* (1969), Werblin and Dowling (1969)], and the horizontal cell response is probably (Werblin and

Dowling 1969), what Rushton calls an $H$-curve, $V/V_{max} = I/(I + K)$ (see e.g. Naka and Rushton, 1966, 1967), $K$ being about 800 quanta/rod/sec for humans. Thus both are roughly linear over a large range. This is supported by the psychophysical findings of Alpern and Rushton (1967), Alpern, Rushton and Torii (1970a–d) (see also Alpern, 1965). Up to the stage just preceding the bipolar signal, therefore, retinal signals are probably linear functions of intensity: but beyond this, they may be logarithmic [see Werblin and Dowling's (1969) bipolar cell records]. Much has been written about how the recorded bipolar signal may be achieved by the observed unusual synaptic structure [Dowling and Boycott (1966), Boycott and Dowling (1969), Dowling and Werblin (1969)]: the bipolar signal in the mudpuppy depends over a large range upon the ratio of the energies incident on the centre and on the surround of its receptive field. In these experiments, the surround stimulus was annular; and it is of some interest to know how the bipolar cell responses behaves for surround stimuli that are not annular, and whether it depends upon the size of the bipolar cell's receptive field. The present theory would prefer the bipolar response to be like equation (1), but does not absolutely require that the component terms be exactly logarithmic, provided that the inverse (3) is tailored closely to the forward transform.

Werblin and Dowling also found that the effect of stimulating the surround of a bipolar cell receptive field is to remove part or all of the hyperpolarisation due to stimulating its centre; but that surround stimuli could not on their own cause positive depolarisation of the bipolar cell. This finding, together with the fact that some bipolar cells have on-centre receptive fields, and some have off-centre ones, shows that the difference signal is split at this stage into its positive and negative parts, which are transmitted down two different channels. Kolb (1970) found that each cone contacts two midget bipolar cells (MB and FMB of Fig. 2): presumably these are the two channels. The separation of the positive and negative parts of the signal gives rise to some complexity in the later parts of this analysis; but it simplifies greatly the question of how a threshold might be applied to the difference signal. If the whole difference signal was carried by a single cell, zero would have to be coded as half the maximum response. Applying a threshold that depends upon the absolute value of the transmitted signal is not easily done under such circumstances. If the signal is split into positive and negative channels, however, it is merely a matter of applying a threshold to the signals in each one.

Because we shall need to refer to the positive and negative parts of the signal, the following notation will be used:

$$\begin{aligned} \mathrm{POS}(x) &= x \text{ if } x > 0 \\ &= 0 \text{ otherwise.} \\ \mathrm{NEG}(x) &= -x \text{ if } x < 0 \\ &= 0 \text{ otherwise.} \end{aligned}$$

and similarly for POS($x'$), POS($x''$), POS($x^*$) etc.

*Thresholding the difference signal*

The heart of the computation is the application of a threshold to the difference signal, for it is this that removes the effects of slow changes in luminance from the image. There is some freedom in how the threshold is applied: for example, $x''$ may be defined as in equation (4), for threshold $t > 0$; or $x''$ may be defined by the two expressions

$$\mathrm{POS}(x'') = \mathrm{POS}(x' - t) \tag{6}$$

$$\mathrm{NEG}(x'') = \mathrm{NEG}(x' + t) \tag{7}$$

which may be slightly easier to implement. The more important question is however what size should the threshold $t$ be?

In order to answer this question, let us consider the effects of doubling the illumination of a scene. The energy received from each point doubles, hence gradients double, and hence the necessary threshold must also double. This is however only a guide, because bright sunlight produces views that are much more contrasty than the (roughly) uniform illumination provided by a thick layer of cloud. It will be important (when adequate transducers become available) to explore how the threshold should vary with lighting conditions: if it is set too high, valuable shading information may be lost. Until then, the linear approximation is the best available, and it is consistent with psychophysical evidence on contrast detection (see below). If the bipolar (difference) signal were logarithmic, a fixed amount $t$ subtracted from all signals in a bipolar axon would have the effect of equations (6) and (7).

## RECONSTITUTING THE IMAGE

The main contribution of this article is to show that retinal structure is well-suited to inverting the difference signal. To reconstitute the image from the bipolar signals, it is necessary to solve a set of equations like those of (5) above. For the sake of precision, this article is limited to the analysis of the message from a single cone, via the midget bipolar, and the midget ganglion cells. The principles are however independent of the particular choice of image resolution.

A straightforward implementation of the equations (5) exists (see Fig. 1, and Horn 1974). In the retina, however, the difference signal is split into positive and negative channels; and the existence of on- and off-centre ganglion cells (Kuffler 1953) suggests that this is also true of the output signal. Because of this, and because the resting discharge frequency of amacrine cells appears to be zero (Werblin and Dowling, 1969), it is taken as a basic constraint on the form of the reconstitution algorithm that the whole computation should be carried out using variables whose signs do not change. This constraint will allow each variable used during the computation to be coded by a nervous element in such a way that zero corresponds to the lowest extreme of its dynamic range. It is fortunately

possible to be quite specific about the implementations that are consistent with the constraint: the argument will be divided into two parts. Firstly the basic constraint is used to select one of the various ways of splitting (5) into two halves; secondly an explicit implementation is exhibited, together with an indication of which of its features are robust enough to support predictions by which the theory may be tested.

*Performing the reconstitution in two halves*

Perhaps the most obvious way in which one might split equation (5) into two halves is the following [$w(y)$ is taken to be $1/n$ for simplicity]:

$$x_1^* = \mathrm{POS}(x'') + 1/2n \sum_{N(X)} y^* \tag{8}$$

$$x_2^* = \mathrm{NEG}(x'') - 1/2n \sum_{N(X)} y^* \tag{9}$$

together with the condition that $x^* = (x_1^* - x_2^*)$. The basic constraint allows us to rule this out, because there is no reason why $x_1^*$ and $x_2^*$ should always be positive. For example, consider a situation where the local average lightness is high, and $x''$ is small but negative. From equation (8), $x_1^*$ is large and positive, and $x_2^*$ is large and negative. Furthermore, if $x''$ is negative, this information is carried by NEG($x''$); the effect of $x''$ in this case is to make $x_2^*$ a little less negative. Therefore $x_2^*$ does not satisfy the basic constraint; and this failure is important, in the sense that a coding of $x_2^*$ that ignored its negative values would cause the reconstitution to fail in certain commonly occurring situations.

In order for a system to be consistent with the basic constraint, it will need to take something like the following form:

$$f(X) = \mathrm{POS}(x'') + 1/n \sum_{N(X)} f(Y) \tag{10}$$

$$g(X) = \mathrm{NEG}(x'') + 1/n \sum_{N(X)} g(Y) \tag{11}$$

where $f$ and $g$ are non-negative functions, that hopefully have some relation to the operator*. At first sight, this pair of equations does not appear to compute anything useful: but observe the following. Subtracting (10) and (11), we obtain:

$$[f(X) - g(X)] = [\mathrm{POS}(x'') - \mathrm{NEG}(x'')] + 1/n \sum_{N(X)} [f(Y) - g(Y)]. \tag{12}$$

Now $[\mathrm{POS}(x'') - \mathrm{NEG}(x'')] = x''$, so that (12) is the same as (5) where the * operator has been replaced by the operator $(f\text{–}g)$. Thus the expression $[f(X) - g(X)]$ is in fact a solution of the equation (5). What this means is that a solution may be obtained from (10) and (11) *provided that the two solutions are coupled in a subtractive way*. In general, the two halves $f(X)$ and $g(X)$ will both be large and positive, since there is nothing negative in either of (10) or (11) to pull them down. The solution is however not disturbed by subtracting a suitable function $h(X)$, provided that it is done to both $f(X)$ and $g(X)$ simultaneously. To satisfy the basic constraint, $h(X)$ must never exceed the smaller of $f(X)$ and $g(X)$: subject to this, a subtractive coupling between $f(X)$ and $g(X)$ is permissible.

Hence we obtain the result that (10) and (11), together with the operations:

$$f(X) \text{ goes to } [f(X) - h(X)] \tag{13}$$

$$g(X) \text{ goes to } [g(X) - h(X)] \tag{14}$$

still represents a solution to (5), as long as the condition

$$f(X) \text{ and } g(X) \text{ are kept positive} \tag{15}$$

also holds. It is perhaps worth pointing out that the determination of $h$ is quite separate from the problem of fixing the d.c. level for the output: nor can variations in $h$ account for the disappearance of stabilized retinal images, since this would correspond to tampering with the difference $[f(X) - g(X)]$: as long as this difference is preserved, the output of the process is a faithful processed copy of the image.

*Implementation details*

The method outlined by equations (10)–(15) leads to an explanation of many features of the inner plexiform layer. There are two basic problems that need to be discussed: firstly, how does one implement the set of linear equations represented by equation (10) or (11); and secondly, how exactly can the subtractive coupling between the two halves (10) and (11) be done?

The first question is the more straightforward. The realisation of equation (10) requires a device with reciprocal connexions to devices that compute the solution at neighbouring points (see Fig. 1). Since the retinal ganglion cells are not pre-synaptic to any other retinal cells, the expression of, for example, $f(X)$ cannot exist only at a ganglion cell, because if it were, it would not be available for the computation at neighbouring ganglion cells. Hence if $f(X)$ is computed in the retina, it will be found either in the bipolar cell axon terminals, or in the amacrine cells, or both.

The two kinds of midget bipolar cell (FMB and MB) terminate in the top and the bottom thirds respectively of the inner plexiform layer (Boycott and Dowling 1969, Kolb 1970). The additive coupling between solutions of $f(X)$ at neighbouring points must therefore be provided by stratified amacrine cells that are driven by, and drive, the bipolar terminals. From Fig. 2, we see that the stratified amacrine cells coupled in this way with the midget bipolar cells must be those of category (A3). The only candidate for the place at which $f(X)$ is computed, which is consistent with retinal structure, is therefore the bipolar cell axon terminals.

The hypothesis, that additive coupling between neighbouring points is provided by the stratified amacrine–bipolar cell interaction, is consistent with the observed synaptic arrangements, because the complex of a dyad synapse plus a return synapse from the amacrine cell to the bipolar cell [Fig. 2(d)] is exactly what it requires. It furthermore follows from this interpretation, that the signs of all the synapses in the dyad complex should be positive, from the point of view of the computation. In practice, this means that the bipolar to stratified amacrine, and the bipolar to midget ganglion cell synapses should be excitatory; and that the stratified amacrine to bipolar synapse should have the same effect on the midget bipolar terminal as stimulation of the centre of that cell's receptive field. It is also necessary that a large proportion of the dyad synapses with a midget bipolar cell involve stratified amacrine cells, and one would expect all dyad synapses that do involve a stratified amacrine cell to be accompanied by a synapse from the amacrine cell back to the midget bipolar.

The final point that concerns the additive interactions is the size of the stratified amacrine cell processes. In order for the inverse computation to proceed accurately, these cells should reflect the dimensions of the horizontal cells that, we suppose, are responsible for the difference operation. The diameter of the (A3) amacrine cells is 20–50 nm, which is consistent with the figure of about 35 nm for the radius of the horizontal cell interaction (Boycott and Dowling, 1969; Kolb, 1970).

*An explicit model for the subtractive coupling*

The second issue concerns the subtractive coupling between the plus and the minus halves of the solution. It is clear that this coupling must be provided by diffuse amacrine cells, probably of category (A1), since the only other candidates (A5) are both large and uncommon; and it is probable that the coupling is achieved through amacrine-amacrine and amacrine-ganglion cell synapses. This coupling can be achieved in a number of ways, but it is helpful to exhibit one of them, because variants on the method are then rather easy to understand.

Equations (10) and (11) may be modified to be the following:

$$f(X) = \mathrm{POS}(x'') + \mathrm{POS}\left\{1/n \sum_{N(X)} [f(Y) - g(Y)]\right\} \tag{10'}$$

$$g(X) = \mathrm{NEG}(x'') + \mathrm{POS}\left\{1/n \sum_{N(X)} [g(Y) - f(Y)]\right\}. \tag{11'}$$

Then $[f(X) - g(X)$ is still a solution. Now write

$$MG_+(X) = \mathrm{POS}[f(X) - g(X)]$$
$$MG_-(X) = \mathrm{POS}[g(X) - f(X)]. \tag{12'}$$

Then one of $MG_+(X)$, $MG_-(X)$ will be positive, and will represent the solution.

The idea behind this formulation is that the subtractive interaction is done in two ways, and it is illustrated in Fig. 3. Firstly, there are inhibitory connexions acting upon the stratified amacrine cells in the top and the bottom thirds of the inner plexiform layer. In Fig. 3, this is provided by the diffuse amacrine cells (A1) that receive excitation from a midget bipolar terminal (or possibly from A3 cells) in one layer, and send inhibitory synapses to the stratified amacrine cells in the other. Secondly, there is direct inhibition driven by the bipolar terminal in one layer, acting on the midget ganglion cell in the other.

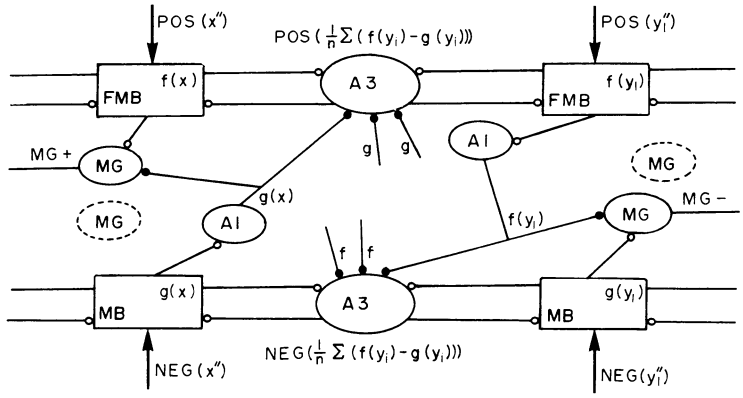



Fig. 3. One way of coupling the two halves of the solution in the inner plexiform layer, as described by equations (10'), (11') and (12'). Open circles represent connexions that are computationally positive: filled circles, those that are negative. Cell names are as in Fig. 2. The association of FMB with POS($x''$) rather than with NEG ($x''$) is arbitrary. The quantities carried by each connexion are shown in the diagram. The principal channel, from $X$, and one neighbour, from $Y_1$, are shown: similar connexions are required between $X$ and the other neighbours.

The important qualitative features of this arrangement are:

(a) The form of the coupling between the two systems of stratified amacrine cells (i.e. the number of terms collected under the POS function), is very flexible; it must however correspond closely in both layers.

(b) The sizes of $f(X)$ and of $g(X)$ (the quantities in the bipolar terminal) are kept positive. This is vital for allowing a POS($x''$) signal to influence the solution even if the solution at $X$ is strongly negative [i.e. $MG_{-}(x)$ is strongly positive. This requirement is so important that it forbids the existence of any significant negative connexion from a diffuse amacrine cell directly onto the bipolar terminal: it is true of all models, because conditions must never totally prevent the signal POS($x''$) from being able to influence the computation].

(c) The diffuse amacrine cells should receive excitatory synapses from the midget bipolar cells, and possibly from the stratified amacrine cells, in one layer: they should send inhibitory synapses to the stratified amacrine and to the midget ganglion cells in the other. (This is a requirement of all models that implement the coupling between the layers, because it has to be provided by diffuse amacrine cells, and the theory requires that the coupling be subtractive.) It is interesting that according to this method, the synapse from a midget bipolar to a diffuse amacrine cell need not be accompanied by a reciprocal synapse, whereas that to a stratified amacrine cell should be.

*Uniqueness*

The arguments that were set out above impose some constraints upon the way in which the reconstitution stage (5) may be implemented, but they fall short of establishing that a particular method must be the one that is used. There are however two arguments, one compelling, and one strong, that greatly constrain the methods that are consistent with retinal structure: they depend upon details that would be difficult to express without having exhibited one method completely. Firstly, should the vertical interaction, between the two halves of the solution, occur before the lateral interactions? In other words, are both halves of the difference signal, POS($x''$) and NEG($x''$), available to both halves of the inverse transform; or is the bipolar signal necessarily contaminated by lateral interactions before it can escape to the other layer? The latter is correct,

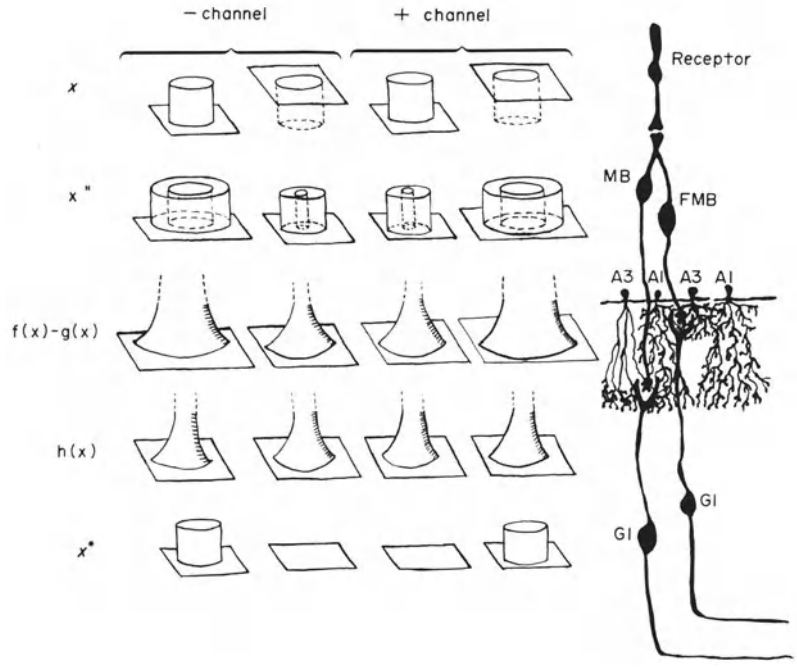



Fig. 4. The computation of lightness using the method described by equations (10)–(15). The figure illustrates the analysis of two stimuli, a light and a dark spot; and both POS and NEG channels are displayed for each one. The intensity information $x$ at $X$ is transformed into a threshold difference signal $x''$. For reconstitution, $f(X)$, $g(X)$, and $h(X)$ are all shown, together with the final output. This particular implementation represents an extremely decoupled technique for computing the reconstitution: $f(X)$ and $g(X)$ represent the values that the midget bipolar terminals would theoretically float up to if, in the model of Fig. 3, the diffuse amacrine connexions were severed. Notice that the response to a central light spot appears at what was initially the negative channel.

because the lateral interactions must be carried by the stratified amacrine cells, which send synapses back to the midget bipolar axon terminals. This means that a formulation is appropriate in which interaction terms between the two layers already contain components due to the lateral interactions.

The second question concerns the degree to which the lateral interactions are "complete" before the vertical interactions begin. In the formulation of equations (10)–(15), the lateral interactions are represented as being complete before the vertical interaction, the subtraction from both layers of $h(X)$, is carried out. This extreme position is undesirable, because $f(X)$ and $g(X)$ will be large compared with $[f(X) - g(X)]$. In the explicit neural model of Fig. 3, the vertical and the lateral interactions proceed simultaneously. The strong argument for this is that the numbers will all be kept small, thus aiding accuracy; but a variety of possibilities lie between these two extremes.

The summarizing diagram shown in Fig. 4 puts the whole model together, and illustrates the computation of lightness in which the inverse is performed in the manner described by equations (10)–(15). The analysis of two stimuli is shown: one, a light spot on a dark background, and the other, a dark spot on a light background. The criteria that were established at various places during the account, as being necessary consequences of the assumption that the retina computes lightness in this way, have been drawn together in the summary.

The above account has dealt with the processing of a single channel of the kind originating from a red or green cone in the fovea; but similar arguments may apply to other channels. I turn now to properties of the retinal input–output relations that are illuminated by these ideas.

## PROPERTIES OF RETINAL GANGLION CELL RESPONSES

### *The Weber–Fechner law*

The Weber–Fechner law (Weber, 1834) relates the increment threshold for a flash linearly to the intensity of the background, and it is true when the flash is long or large, and the background is not too dim (Barlow, 1957, 1958). The present theory may provide an explanation of this curious law: it was seen above that the size of the threshold applied in the lightness computation varies roughly linearly with the ambient luminance, and the increment threshold experiment could measure such a threshold. Hence a linear rule is a consequence of the lightness computation. The constants that arise from the lightness computation may however not be the same as those that occur in the Weber–Fechner law, because the many phasic retinal ganglion cells probably signal quantities other than lightness.

In conditions of low illumination, the difference signals will become unreliable, because they are obtained by integrating over a relatively short time (less than 100 msec). When this happens, the lightness computation must be abandoned. Barlow (1956, 1957) has proposed a satisfying explanation of the increment threshold relation at low intensities: the change in ganglion cell receptive field organisation towards the end of dark adaptation (Barlow, Fitzhugh and Kuffler, 1957) is also to be expected as a result of this change; and the required change in the receptor-horizontal cell interactions may be related to the puzzling behaviour of the electro-retinogram at low intensities (Cone and Ebrey, 1965; Brindley, 1970, pp. 50–56). Illusions like the Craik–Cornsweet illusion, that depend on the lightness computation for their effect, should fail to deceive at these low intensities.

### *Ganglion cell receptive field organization*

According to the present theory, retinal ganglion cells should behave like idealized receptors—receptors of lightness rather than of intensity. Of the three main categories of retinal ganglion cell, the $X$, the $Y$ (Enroth-Cugell and Robson, 1966; Fukada, 1971; Fukada and Saito, 1971; Cleland, Dubin and Levick, 1971, all in the cat; Gouras, 1968 in the monkey), and the $W$ cells (Rodieck, 1967; Fukada, 1971; Stone and Hoffmann, 1972), the $X$ cells seem best suited to carry the lightness signal. The reasons for thinking this are as follows: (1) The lightness computation is quite complex, and to give the result time to settle, probably uses components with quite long time constants. $X$-cells have bad temporal resolution, whereas $Y$ and $W$ cells do not. (2) $X$-cells are the only cells whose response is tonic. (3) $W$-cells probably project to the superior colliculus. (4) $X$-cells have the highest resolution, and are more common toward the fovea (Gouras, 1968). (5) McIlwain's (1964, 1066) periphery effect is weak or absent for $X$-cells, as it should be if the computation is being done correctly.

If this view is correct, the detailed analysis of visual information, and especially of colour (see Zeki, 1973), should rely mostly upon the $X$ input (see Stone and Dreher, 1973), though for reasons perhaps related to after-images, the results of such analysis may be gated by $Y$ information.

The conventional interpretation of the centre-surround organization of the ganglion cell receptive field is that it helps to preserve contrast information over variations in average luminance. According to the present theory, it is a by-product of splitting the retinal output into two channels, the most sensible setting of the zero level being something like the average value of lightness over a region of the retina. This opinion is supported by the results of Maffei and Fiorentini (1972): they showed that the centre-surround organization of geniculate cells makes little use of the centre-surround organization of the ganglion cells, because a geniculate cell surround is driven by ganglion cells whose *centres* project to its surround. Their assertion that the retinal organization is positively "lost" (p. 65) as a result of the geniculate computation is however

unfounded: at most, the retinal organization may be inessential.

*Separation of the three colour channels*

There is some evidence that the rod and the three cone channels are processed independently [Alpern, 1965; Gouras and Link, 1966; Gouras, 1966, 1967; Alpern, Rushton and Torri, 1970(a and d); Westheimer, 1970; Westheimer and Wiley, 1970; and McKee and Westheimer, 1970]. Recent papers (Lennie and MacLeod, 1973; Barlow and Sakitt, 1973; see also Brindley, 1970, pp. 75–86) cast doubt on a number of these findings, however; and although there is very little information available about chromatic interaction in the primate retina (Hubel and Wiesel, 1960; Gouras, 1968), some interaction is visible in primate retinal ganglion cells, and much takes place in the lateral geniculate nucleus (De Valois, 1965; Hubel and Wiesel, 1966).

To what extent is chromatic interaction in the retina consistent with the present theory? There are three ways in which it could be introduced. Firstly, the d.c. level for the output may conveniently be determined by summing locally over all colour channels. Retinal ganglion cells with a centre-surround receptive field organization that show complementary spectral sensitivities in the centre and surround, would be consistent with this: the function of such cells is otherwise rather difficult to understand. "Opponent colour" mechanisms that do no more than form linear combinations of receptor signals are as far from estimating reflectance (and hence colour) as the raw receptor signals.

Secondly, the lightness computation can equally well be performed on linear combinations of the receptor signals. It may be convenient, for example, to use a red + green channel, with two others, rather than the original inputs: such a mechanism would be useful for separating chromatic from spatial information. Thirdly, chromatic interactions needed after the lightness computation could perhaps be pulled back into the retina: a modular design, though probably necessary at some stages of evolution, does not have to be preserved thereafter.

*Acknowledgements*—I wish to thank Dr. B. K. P. Horn for many useful discussions; Profs. M. Minsky and S. Papert for inviting me to the A.I. Laboratory; and Miss H. Kolb, Profs. B. B. Boycott and J. E. Dowling, and the Royal Society for permission to reproduce the illustrations of Fig. 2. Allison Platt and Suzin Jabari kindly prepared the figures. This work was supported in part by the Advanced Research Projects Agency of the Department of Defense, and monitored by the Office of Naval Research under contract no. N00014-70-A-0362-0002.

## REFERENCES

Alpern M. (1965) Rod–cone independence in the after-flash effect. *J. Physiol., Lond.* **176,** 462–472.

Alpern M. and Rushton W. A. H. (1967) The nature of rise in threshold produced by contrast-flashes. *J. Physiol., Lond.* **189,** 519–534.

Alpern M., Rushton W. A. H. and Torii, S. (1970a) The size of rod signals. *J. Physiol., Lond.* **206,** 193–208.

Alpern M., Rushton W. A. H. and Torii S. (1970b) The attenuation of rod signals by backgrounds. *J. Physiol., Lond.* **206,** 209–227.

Alpern M., Rushton W. A. H. and Torii S. (1970c) The attenuation of rod signals by bleachings. *J. Physiol., Lond.* **207,** 449–461.

Alpern M., Rushton W. A. H. and Torii S. (1970d) Signals from cones. *J. Physiol., Lond.* **207,** 463–475.

Barlow H. B. (1956) Retinal noise and absolute threshold. *J. opt. Soc. Am.* **46,** 634–639.

Barlow H. B. (1957) Increment thresholds at low intensities considered as signal–noise discriminations. *J. Physiol., Lond.* **136,** 469–488.

Barlow H. B. (1958) Temporal and spatial summation in human vision at different background intensities. *J. Physiol., Lond.* **141,** 337–350.

Barlow H. B., Fitzhugh R. and Kuffler S. W. (1957) Change of organization in the receptive field of the cat's retina during dark adaptation. *J. Physiol., Lond.* **137,** 338–354.

Barlow H. B. and Sakitt B. (1973) Doubts about scotopic interactions in stabilized vision. *Vision Res.* **13,** 523–524.

Boycott B. B. and Dowling J. E. (1969) Organization of the primate retina: light microscopy (with an appendix by H. Kolb). *Phil. Trans. R. Soc. B.* **255,** 109–184.

Brindley G. S. (1970) *Physiology of the Retina and Visual Pathway* (Physiological Society Monograph no. 6). Edward Arnold Ltd., London.

Cajal S. R. (1911) *Histologie du Système Nerveux.* C.S.I.C., Madrid.

Cleland B. G., Dubin M. W. and Levick W. R. (1971) Sustained and transient neurones in the cat's retina and lateral geniculate nucleus. *J. Physiol., Lond.* **217,** 473–496.

Cone R. A. and Ebrey T. G. (1965) Functional independence of the two major components of the rod electroretinogram. *Nature, Lond.* **221,** 818–820.

De Valois R. L. (1965) Analysis and coding of colour vision in the primate visual system. *Cold Spring Harb. Symp. quant. Biol.* **30,** 567–579.

Dowling J. E. and Boycott B. B. (1966) Organization of the primate retina: electron microscopy. *Proc. R. Soc. B.* **166,** 80–111.

Dowling J. E. and Werblin F. S. (1969) Organization of the retina of the mud-puppy, *Necturus maculosus*: I. Synaptic structure. *J. Neurophysiol.* **32,** 315–338.

Enroth-Cugell C. and Robson J. G. (1966) The contrast sensitivity of retinal ganglion cells of the cat. *J. Physiol., Lond.* **187,** 517–552.

Fukada Y. (1971) Receptive field organization of cat optic nerve fibres with special reference to conduction velocity. *Vision Res.* **11,** 209–226.

Fukada Y. and Saito H.-A. (1971) The relationship between response characteristics to flicker stimulation and receptive field organization in the cat's optic nerve fibres. *Vision Res.* **11,** 227–240.

Gouras P. (1966) Rod and cone independence in the electroretinogram of the dark-adapted monkey's perifovea. *J. Physiol., Lond.* **187,** 455–464.

Gouras P. (1967) The effects of light-adaptation on rod and cone receptive field organization of monkey ganglion cells. *J. Physiol., Lond.* **192,** 747–760.

Gouras P. (1968) Identification of cone mechanisms in monkey ganglion cells. *J. Physiol., Lond.* **199,** 533–547.

Gouras P. and Link K. (1966) Rod and cone interaction in dark-adapted monkey ganglion cells. *J. Physiol., Lond.* **184,** 499–510.
Helmholtz H. (1962) *Treatise on Physiological Optics.* Dover Publications Inc., New York (First edition of *Handbuch der Physiologischen Optik* published in 1867 by Voss, Leipzig).
Horn B. K. P. (1974) On lightness (submitted for publication.)
Hubel D. H. and Wiesel T. N. (1960) Receptive fields of optic nerve fibres in the spider monkey. *J. Physiol., Lond.* **154,** 572–580.
Hubel D. H. and Wiesel T. N. (1966) Spatial and chromatic interactions in the lateral geniculate body of the rhesus monkey. *J. Neurophysiol.* **29,** 1115–1156.
Kaneko A. and Hashimoto H. (1967) Recording site of the single cone response determined by an electrode marking technique. *Vision Res.* **7,** 847–851.
Kolb Helga (1970) Organization of the outer plexiform layer of the primate retina: electron microscopy of Golgi-impregnated cells. *Phil. Trans. R. Soc. B.* **258,** 261–283.
Kuffler S. W. (1953) Discharge patterns and functional organization of mammalian retina. *J. Neurophysiol., Lond.* **16,** 37–68.
Land E. H. (1964) The retinex. *Am. Scientist* **52,** 247–264.
Land E. H. and McCann J. J. (1971) Lightness and retinex theory. *J. opt. Soc. Am.* **61,** 1–11.
Lennie P. and MacLeod D. I. A. (1973) Background configuration and rod threshold. *J. Physiol., Lond.* **233,** 143–156.
McIlwain J. T. (1964) Receptive fields of optic tract axons and lateral geniculate cells: peripheral extent and barbiturate sensitivity. *J. Neurophysiol.* **27,** 1154–1173.
McIlwain J. T. (1966) Some evidence concerning the physiological basis of the periphery effect in the cat's retina. *Expl Brain Res.* **1,** 265–271.
McKee S. and Westheimer G. (1970) Specificity of cone mechanisms in lateral interactions. *J. Physiol., Lond.* **206,** 117–128.
Maffei L. and Fiorentini A. (1972) Retinogeniculate convergence and analysis of contrast. *J. Neurophysiol.* **35,** 65–72.
Missotten L. (1965) *The Ultrastructure of the Human Retina.* Arscia Uitgaven N.V., Brussels.
Naka K. I. (1969) Computer assisted analysis of *S*-potentials. *Biophys. J.* **9,** 845–859.
Naka K. I. and Rushton W. A. H. (1966) *S*-potentials from colour units in the retina of fish (Cyprinidae). *J. Physiol., Lond.* **185,** 536–555.
Naka K. I. and Rushton W. A. H. (1967) The generation and spread of *S*-potentials in fish (Cyprinidae). *J. Physiol., Lond.* **192,** 437–461.
Naka K. I. and Rushton W. A. H. (1968) *S*-potential and dark adaptation in fish. *J. Physiol., Lond.* **194,** 259–269.
Ratliff F. (1965) *Mach Bands: Quantitative Studies on Neural Networks in the Retina.* Holden-Day, San Francisco.
Rodieck R. W. (1967) Receptive fields in the cat retina: a new type. *Science, N.Y.* **157,** 90–92.
Rushton W. A. H. (1972) Pigments and signals in colour vision (invited lecture to the Physiological Society). *J. Physiol., Lond.* **220,** 1P–31P.
Stone J. and Dreher B. (1973) Projection of *X*- and *Y*-cells of the cat's lateral geniculate nucleus to areas 17 and 18 of visual cortex. *J. Neurophysiol.* **36,** 551–567.
Stone J. and Hoffman K.-P. (1972) Very slow-moving ganglion cells in the cat's retina: a major, new functional type? *Brain Res.* **43,** 610–616.
Tomita T. (1968) Electrical responses of single photoreceptors. *Proc. I. E. E. E.* **56,** 1015–1023.
Toyoda J., Nosaki H. and Tomita T. (1969) Light-induced resistance changes in single photoreceptors of *Necturus* and *Gekko. Vision Res.* **9,** 453–463.
Weber E. H. (1834) *De Pulsu, Resorptione, Audity et Tactu Annotationes Anatomicae et Physiologicae* (cited by Brindley, 1970). C. F. Koehler, Leipzig.
Werblin F. S. and Dowling J. E. (1969) Organization of the retina of the mud-puppy, *Necturus maculosus*: II. Intracellular recording. *J. Neurophysiol.* **32,** 339–355.
Westheimer G. (1970) Rod–cone independence for sensitizing interaction in the human retina. *J. Physiol., Lond.* **206,** 109–116.
Westheimer G. and Wiley R. W. (1970) Distance effects in human scotopic retinal interaction. *J. Physiol., Lond.* **206,** 129–143.
Zeki S. M. (1973) Colour coding in rhesus monkey prestriate cortex. *Brain Res.* **53,** 422–427.

**Résumé**—On propose qu'une des fonctions de la rétine est de calculer la luminosité au moyen d'un algorithme parallèle à deux dimensions. Il y a trois stages: (1) une opération de différence centre-bord; (2) un seuil appliqué au signal de différence; (3) l'inverse de (1), dont le signal de sortie est la luminosité. On analyse on détail l'opération du canal bipolaire naine-ganglionnaire naine, et on donne une interprétation fonctionnelle des diverses structures rétiniennes. On précise les conditions de la théorie au sujet de l'arrangement et des connexions entre cellules, et des signes des synapses, dans la couche plexiforme interne.

**Zusammenfassung**—Es wird angenommen, dass eine der Funktionen der Retina die Berechnung eines Helligkeitssignales unter Verwendung eines zweidimensionalen Parallel-Algorithmus ist. Dabei gibt es drei Stufen: (1) Eine Zentrums-Peripherie Differenz-Operation; (2) eine Schwelle, der das Differenz-Signal unterworfen wird; (3) die inverse Operation von (1), deren Ergebnis das Helligkeitssignal ist. Die Wirkungsweise des Zwergbipolar-Zwergganglion-Kanals wird im einzelnen untersucht. Die Voraussetzungen der Theorie werden formuliert, soweit sie die Anordnung und die Verbindungen von Zellen und die Vorzeichen synaptischer Verbindungen in der inneren plexiformen Schicht der Netzhaut betreffen.

**Резюме**—Предполагается, что одна из функций сетчатки—количественная оценка светлоты, для чего используется двухразмерный алгоритм. Имеются три стадии: (1) центрально-периферическая операция дифференциации; (2) применение порога по отношению к дифференцировочному сигналу; (3) инверсия (1) выходом которой является светлота. Операция: карликовые

биполяры—канал карликовых ганглиозных клеток, детально проанализирована и дается функциональная интерпретация различных структур сетчатки. Установлены теоретвческие требования, касающиеся расположения и связей клеток, знаков синапсов, во внутреннем плексиформном слое.

**Norberto M. Grzywacz**

Commentary on

# *The Computation of Lightness by the Primate Retina*

Neural processes in the brain are marvelously complex devices. They include a wide variety of cell morphologies, intricate rules of synaptic connectivity, and numerous types of neurotransmitters and ionic channels. Thus, it is not surprising that one finds a very large repertoire of physiological behaviors.

The best way to make sense of this neural complexity is to focus the analysis of the neural processes within the scope of behavioral function. In the case of the visual system, one would like to know how its neurons recognize the properties of objects in the visual world. Perhaps one of the most important of such properties is the reflectance of objects. The visual system can somehow perceive objects' reflectances through its ability to compute lightness. This computation requires interneuron interactions, because photoreceptors measure illuminations and not lightness.

In his paper,"The Computation of Lightness in the Primate Retina," David Marr (1974) tried to make sense of several complex features of the retina by proposing that it computes lightness. The retinal circuit he advanced is a rough implementation of the algorithm of Horn (1974) (an elaboration of earlier work by Land and McCann, 1971). This implementation starts with a horizontal-cell mediated lateral inhibition. The resulting difference signal is then passed through a threshold mechanism and transmitted to midget bipolar cells. In turn, these cells make reciprocal excitatory synapses with stratified diffuse amacrine cells (Boycott and Dowling, 1969). This stratification, which takes place in the inner nuclear layer, maintains the segregation of ON and OFF signals coming from two populations of bipolar cells. Finally, the ON and OFF sublamina inhibit each other via the narrow field diffuse amacrine cell (Boycott and Dowling, 1969). The output of the system flows from the bipolar cells to the midget ganglion cells and then to the brain.

As I comment below, the weight of the evidence is against Marr's proposition from the perspectives of retinal output and retinal circuitry. Nevertheless, I will argue that Marr's approach was worthwhile; it generated exciting predictions and raised clear suggestions for retinal function. Marr himself expressed a similar view in a later work (Marr, 1982): "I do not now believe that this is at all a correct analysis of the retina, but it showed the possible style of a correct analysis. Present is a clear understanding of what is to be computed, how it is to be done..."

To start the discussion of why the proposition apparently failed, it is necessary to clarify some nomenclature. Marr called the neurons proposed to relay lightness information "midget-ganglion cells," which is primate terminology

(Polyak, 1941), or "X-ganglion cells," which is cat terminology (Enroth-Cugell and Robson, 1966). But there is an intense debate on the correspondence between the ganglion cells of primates and cats (Shapley and Perry, 1986; Rodieck, 1988). Thus, one wonders which of the two main types of primate ganglion cells, that is, midget or parasol (Polyak, 1941; Watanabe and Rodieck, 1989) Marr referred to. He explicitly wrote that the cells carrying lightness information should be the X-cells, because among all ganglion cells: (a) they have the longest integration time constants, (b) their responses are tonic, (c) they have the highest spatial resolution, and (d) they are the most common ganglion cells near the fovea. Thus, from primate data (Dreher et al., 1976; Schiller and Malpeli, 1978) one should consider the cells referred to by Marr to be the midget ganglion cells. (A needed qualification is that their responses are only tonic for stimuli with narrow wavelength bandwidth [DeMonasterio, 1978].)

After this nomenclature clarification, the question that immediately comes to mind is this: Do midget ganglion cells carry lightness information? (And in particular, does the analysis of color information rely mostly on these cells?) In this case, if the image contrasts remain constant, the responses of midget ganglion cells should be relatively invariant to modulations of background stimulation (as implied by the Weber-Fechner law [Marr, 1974]).

The evidence suggests that midget ganglion cells do not carry lightness information. For instance, these cells are sensitive to the wavelengths of the background stimulus (DeMonasterio, 1978). Furthermore, analysis of the data of Purpura et al. (1988) shows that for mesopic and low-photopic background illuminations, the responses of midget ganglion cells do not obey the Weber-Fechner law. As pointed out by Marr, this law is a necessary condition in any system carrying lightness information.

However, several lines of evidence point out that Marr might have guessed correctly that the midget ganglion cells are important for color analysis. For example, the surround of their receptive field, but not of other ganglion cells, often has a different action spectrum than that of the receptive field center (DeValois, 1960; DeMonasterio and Gouras, 1975). (In primates, double-opponent cells appear for the first time in the cortex [Michael, 1978a,b; Gouras and Kruger, 1979].) Another example is the low contrast gain of midget ganglion cells when compared to other ganglion cells (Purpura et al., 1988; Kaplan and Shapley, 1986) and the low contrast gain of human color perception (Mullen, 1985). (A possible complication is that color perception, but not the responses of midget ganglion cells, has low spatial resolution [Mullen, 1985]. However, color perception might not be highly resolving, because the contrast gain of midget ganglion cells is so low that many of them may have to cooperate to contribute to perception [Shapley and Perry, 1986].)

But the evidence is against Marr not only because of the lack of lightness signals, but also because of the disagreement between the proposed circuitry and experimental findings. In a way, such mistakes are expected, since Marr had to make several assumptions about neural processes that were unknown

at the time.

No fault is found in the assumption that horizontal cells make lateral inhibitory connections to cone pedicles (Kolb, 1970). Also, there is no problem in assuming that bipolar cells make excitatory synapses onto amacrine cells. But perhaps, the first difficulties encountered by Marr's proposition have to do with photoreceptor-bipolar transmission. Because his functions $f(x)$ and $g(x)$ represent voltages in bipolar cells' telodendrons (see his Fig. 3), and because bipolar cells are essentially isopotential, the threshold must occur in the photoreceptor synapse. However, Fain et al. (Fain, 1977; Fain et al., 1977) showed that photoreceptors' voltage signal at visual threshold is small (50-100 mV in human rods and 5-10 mV in turtle cones), implying that their synapses may not have functional thresholds. Another problem with Marr's postulates for the photoreceptor-bipolar transmission is his insistence that $f(x), g(x) > 0$, forcing the ON and OFF bipolar cells to have responses with similar polarity. This is a problem, because on light activation ON bipolar cells depolarize while OFF bipolar cells hyperpolarize (Werblin and Dowling, 1969; Matsumoto and Naka, 1972; Dacheux and Miller, 1981). (A solution for this problem may be to say that OFF bipolar cells depolarize at light offset. Also, I found a way to solve this problem by slightly modifying Marr's diagrams, but this is outside the scope of this commentary. It is significant that Marr guessed correctly that the segregation between the ON and OFF pathways would be maintained in the inner plexiform layer. Direct evidence for this only appeared four years later [Nelson et al., 1978].)

The innovative part of Marr's proposals refer to the inner plexiform layer circuitry, but again, the experimental evidence is against him. However, a cautionary note must be presented here: any criticism of a model of the inner plexiform layer must be tempered with our limited knowledge about this layer. For example, in rabbit, whose retina has been widely studied, only about 30% of all amacrine cells have been characterized with respect to dendritic morphology, neurotransmitter content, and topographic distribution (Vaney, 1990). Thus, if a modeler postulates the existence of a certain cell type, it is possible that this cell exists. Accordingly, I restrain my analysis of Marr's proposition to the cells he mentioned, while keeping in mind that in the future appropriate substitutes might appear.

Marr postulated an inhibitory interconnection between the ON and OFF sublaminas mediated by the narrow field diffuse amacrine cell. This cell, later renamed AII type amacrine cell (Famiglietti and Kolb, 1975), appears to be glycinergic (Marc and Liu, 1985) and thus is presumably inhibitory. However, its inputs and outputs are wrong for Marr's purposes (Famiglietti and Kolb, 1975; Nelson et al., 1976; Nelson, 1982; Kolb and Famiglietti, 1974; Pourcho, 1982; Sterling, 1983): (a) it is mainly driven by rod inputs, (b) it only makes conventional synapses in the OFF sublamina, and (c) it has gap junctions with cone bipolar telodendrons in the ON sublamina. Marr also postulated that the midget bipolar cells make excitatory reciprocal synapses with stratified diffuse amacrine cells. Not a lot has been said about these amacrine cells since

their identification (Boycott and Dowling, 1969), thus little can be said about Marr's postulate. However, of the more extensively studied amacrine cells none appear to fulfill completely Marr's criteria. For example, cells A3 and A4 of Kolb et al. (1981) would be good candidates because they are broadly stratified in the OFF and ON sublaminas, respectively. Unfortunately, they seem to be glycinergic (Pourcho and Goebel, 1985) and, therefore, probably inhibitory. A better example would be the Ca and Cb cholinergic amacrine cells (Masland et al.,1984; Schmidt et al., 1985; Rodieck, 1989), which stratify in the OFF and ON sublaminas respectively, and can be excitatory (they are excitatory to ganglion cells [Ariel and Daw, 1982; Ariel and Adolph, 1985]). The trouble is that they connect to ganglion cells (Famiglietti, 1983; Brandon, 1987) or to each other (Millar and Morgan, 1987; Mariani and Hersh, 1988), but not to bipolar cells.

Marr's paper did not have an impact in retinal research. A search I conducted with the Science Citation Index did not reveal even a single reference to Marr's paper from experimental retinal investigators. This is too bad, because the paper is rich in specific and testable predictions. Also, as it turned out, later experiments corroborated some of the predictions formulated by Marr. For example, his prediction of the ON and OFF segregation through stratification in the inner nuclear layer was later reformulated by retinal anatomists (Famiglietti and Kolb, 1976) and successfully tested (Nelson et al., 1978). (Interestingly, some psychophysicists appreciated this prediction [MacLeod, 1978; Stelmach et al., 1987] seeing virtue in Marr's computational arguments for the early segregation of the ON and OFF pathways.

Despite the negative evidence against his model and its lack of impact in retinal research, I would like to argue that Marr's approach is a worthwhile one. This approach is to formulate putative neural circuits implementing functions that have been previously understood from a computational perspective. (In the present case, the computational studies were those of Land and McCann [1971] and Horn [1974].) The implementation should not be arbitrary, but rather should make maximal use of the available data.

I submit that the incorrect predictions made by the model are not fundamental to Marr's approach in this paper. Rather, they may occur in any theoretical work and are a healthy part of science. As Francis Crick puts it (Crick, 1988):

> The principal error I see in most current theoretical work is that of imagining that a theory is really a good model for . . . nature rather than being merely a demonstration (of possibility) a 'don't worry' theory. . . If elegance and simplicity are . . . dangerous guides, what constraints can be used as guide through the jungle of possible theories? . . . The only useful constraints are contained in the experimental evidence . . . Theorists . . . should realize that it is extremely unlikely that they will produce a useful theory just by having a bright idea distantly related to what they imagine to be the facts . . . The very process of abandoning theories gives them a degree of critical detachment which is almost essential.

Later in his career, Marr shifted to a different approach (Marr, 1982). He

emphasized the computational level of understanding at which the information-processing tasks carried out during perception are analyzed independently of the particular mechanisms that implement them in the brain. Even though this is an important realization, I believe that it might be misleading. Of the several computational approaches available to solve a task, the one that the brain uses may not be same one that is optimal under some mathematical criterion, but rather one that makes good use of the available neural circuitry (Grzywacz and Poggio, 1990). I believe that Marr's early foray into computational visual neuroscience is the way to go, and if coupled with experiments, it will lead to fruitful brain research.

ACKNOWLEDGMENT

I would like to thank Lyle Borg-Graham for critically reading this commentary.

## REFERENCES

Ariel M, Adolph AR (1985): Neurotransmitter inputs to directionally sensitive turtle retinal ganglion cells. *J Neurophysiol* 54:1123-143

Ariel M, Daw NW (1982): Pharmacological analysis of directionally sensitive rabbit retinal ganglion cells. *J Physiol* 324:161-185

Boycott BB, Dowling JE (1969): Organization of the primate retina: light microscopy. *Phil Trans R Soc B* 255:109-184

Brandon C (1987): Cholinergic neurons in the rabbit retina: dendritic branching and ultrastructural connectivity. *Brain Res* 426:119-130

Crick F (1988): *What Mad Pursuit.* New York: Basic Books

Dacheux RF, Miller RF (1981): An intracellular electrophysiological study of the ontogeny of functional synapses in the rabbit retina. 1. Receptors, horizontal and bipolar cells. *J Comp Neurol* 198:307-326

De Monasterio FM (1978): Properties of concentrically organized X and Y ganglion cells of macaque retina. *J Neurophysiol* 41:1394-1417

DeMonasterio FM, Gouras P (1975): Functional properties of ganglion cells of the rhesus monkey retina. *J Physiol* 251:167-195

DeValois RL (1960): Color vision mechanisms in the monkey. *J Gen Physiol* 43 (Suppl.):115-128

Dreher B, Fukuda Y, Rodieck RW (1976): Identification, classification, and anatomical segregation of cells with X-like and Y-like properties in the lateral geniculate nucleus of old world primates. *J Physiol* 258:433-452

Enroth-Cugell C, Robson JG (1966): The contrast sensitivity of retinal ganglion cells of the cat. *J Physiol* 187:517-552

Fain GL (1977): The threshold signal of photoreceptors. In : *Vertebrate Photoreception*, Barlow HB, Fatt P, eds. London: Academic Press, pp 305-323

Fain GL, Granda AM, Maxwell JH (1977): The voltage signal of photoreceptors at the visual threshold. *Nature*, 265:181-183

Famiglietti EV Jr. (1983): ON and OFF pathways through amacrine cells in mammalian retina: the synaptic connection of starburst amacrine cells. *Vision Res* 23:1265-1279

Famiglietti EV Jr., Kolb H. (1975): A bistratified amacrine cell and synaptic circuitry in the inner plexiform layer of the retina. *Brain Res* 84:293-300

Famiglietti EV Jr., Kolb H (1976): Structural basis for ON- and OFF-center responses in retinal ganglion cells. *Science*, 194:193-195

Gouras P, Kruger J (1979): Responses of cells in foveal visual cortex of the monkey to pure color contrast. *J Neurophysiol* 42:850-860

Grzywacz NM, Poggio T (1990): Computation of motion by real neurons. In: *An Introduction to Neural and Electronic Networks*, Zornetzer SF, Davis JL and Lau C, eds. Orlando: Academic Press, pp 379-403

Horn BKP (1974): Determining lightness from an image. *Comput Graph Image Process* 3:277-299

Kaplan E, Shapley R (1986): The primate retina contains two types of ganglion cells, with high and low contrast sensitivity. *Proc Natl Acad Sci USA* 83:2755-2757

Kolb H (1970): Organization of the outer plexiform layer of the primate retina: electron microscopy of Golgi-impregnated cells. *Phil Trans R Soc B* 258:261-283

Kolb H, Famiglietti EV Jr (1974): Rod and cone pathways in the inner plexiform layer of the cat retina. *Science*, 186:47-49

Kolb H, Nelson R, Mariani A (1981): Amacrine cells, bipolar cells and ganglion cells of the cat retina: A Golgi study. *Vision Res* 21:1081-1114

Land EH, McCann JJ (1971): Lightness and retinex theory. *J Opt Soc* Am. 61:1-11

MacLeod DIA (1978): Visual sensitivity. *Ann Rev Psychol* 29:613-645

Marc RE, Liu WS (1985): (3H) glycine accumulating neurons of the human retina. *J Comp Neurol* 232:241-260

Mariani AP, Hersh LB (1988): Synaptic organization of cholinergic amacrine cells in the rhesus monkey retina. *J Comp Neurol* 267:269-280

Marr D (1974): The computation of lightness by the primate retina. *Vision Res* 14:1377-1388

Marr D (1982): *Vision.* San Francisco: WH Freeman

Masland RH, Mills JW, Hayden SA (1984): Acetylcholine synthesizing amacrine cells: identification and selective staining using autoradiography and fluorescent markers. *Proc R Soc Lond B* 223:79-100

Matsumoto N, Naka KI (1972): Identification of the intracellular responses of the frog retina. *Brain Res* 42:59-71

Michael CR (1978a): Color vision mechanisms in monkey striate cortex: dual opponent cells with concentric receptive fields. *J Neurophysiol* 41:572-588

Michael CR (1978b): Color vision mechanisms in monkey striate cortex: simple cells with dual opponent-color receptive fields. *J Neurophysiol* 41:1233-1249

Millar TJ and Morgan IG (1987): Cholinergic amacrine cells in the rabbit retina synapse onto other cholinergic amacrine cells. *Neurosci Lett* 74:281-285

Mullen KT (1985): The contrast sensitivity of human colour vision to red- green and blue-yellow chromatic gratings. *J Physiol* 359:381-400

Nelson R (1982): AII amacrine cells quicken time course of rod signals in the cat retina. *J Neurophysiol* 47:928-947

Nelson R, Famiglietti EV Jr., Kolb H (1978): Intracellular center ganglion cells in cat retina. *J Neurophysiol* 41:472-483

Nelson R, Kolb H, Famiglietti EV Jr, Gouras P (1976): Neural responses in the rod and cone systems of the cat retina: intracellular recordings and procion stains. *Invest Ophthalmol* 41:472-483

Polyak SL (1941): *The Retina.* Chicago: University of Chicago Press

Pourcho RG (1982): Dopaminergic amacrine cells in the cat retina. *Brain Res* 252:101-109

Pourcho RG and Goebel DJ (1985): A combined Golgi and autoradiographic study of [$^3$H]glycine-accumulating amacrine cells in the cat retina. *J Comp Neurol* 233:473-480

Purpura K, Kaplan E, Shapley RM (1988): Background light and the contrast gain of primate P and M retinal ganglion cells. *Proc Nat Acad Sci USA* 85:4534-4537

Rodieck RW (1988): The primate retina. In: *Comparative Primate Biology,* Steklis HD, Erwin J, eds., New York: Alan R. Liss, vol 4 pp 203-278

Rodieck RW (1989): Starburst amacrine cells of the primate retina. *J Comp Neurol* 285:18-37

Schiller PH, Malpeli JG (1978): Functional specificity of lateral geniculate nucleus laminae of the rhesus monkey. *J Neurophysiol* 41:788-797

Schmidt M, Humphrey MF, Wassle H (1985): Action and localization of acetylcholine in the cat retina. *J Neurophysiol* 58:997-1015

Shapley R, Perry VH (1986): Cat and monkey retinal ganglion cells and their visual functional roles. *Trends Neurosci* 9:229-235

Stelmach LB, Bourassa CM, Di Lollo V (1987): ON and OFF systems in human vision. *Vision Res* 27:919-928

Sterling P (1983): Microcircuitry of the cat retina. *Annu Rev Neurosci* 3:149-185

Vaney D (1990): The mosaic of amacrine cells in the mammalian retina. In: *Progress in Retinal Research,* Osborne N, Chader G eds. New York: Pergamon Press, vol. 9, pp 49-100

Watanabe M, Rodieck RW (1989): Parasol and midget ganglion cells of the primate retina. *J Comp Neurol* 289:434-454

Werblin FS, Dowling JE (1969): Organization of the retina of the mudpuppy. Necturus Maculosus. II. Intracellular recordings. *J Neurophysiol* 32:339-355

*Research Scientist*
*Center for Biological Information Processing*
*Department of Brain and Cognitive Sciences*
*Massachusetts Institute of Technology*
*Cambridge, Massachusetts*

# Binocular Depth Perception

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

Memo No.327 December, 1974

# A Note on the Computation of Binocular Disparity in a Symbolic, Low-Level Visual Processor

David Marr

*Abstract*

The goals of the computation that extracts disparity from pairs of pictures of a scene are defined, and the constraints imposed upon that computation by the three-dimensional structure of the world are determined. Expressing the computation as a gray-level correlation is shown to be inadequate. A precise expression of the goals of the computation is possible in a low-level symbolic visual pro- cessor: the constraints translate in this environment to prerequisites on the binding of disparity values to low-level symbols. The outline of a method based on this is given.

## Introduction

Commercial pressures have led to considerable interest in the automatic extraction of disparity information from pairs of pictures of a scene. Since 1968, there has been available a machine, the Wild-Raytheon B8 stereomat automated plotter, which can draw a contour map from two aerial photographs. The machine correlates intensity measurements obtained over local scans made on the two images: the scan paths are the machine's current approximation to the contour lines (i.e., lines of constant disparity), and the failures of correspondence between the scans on the two images are used to improve the approximation. Adequate accuracy, if achieved at all, is reached within about six iterations.

Machines that seek to assign disparity values to an image by performing correlations between intensity arrays are subject to troublesome problems due to local minima: Mori, Kidode & Asada (1973) describe the problems, and have recently implemented some ways of avoiding them. Their principal cures are (i) to correlate the two images using local averages taken over regions that are initially relatively large, and which are subsequently reduced in size as the solution is approached; and (ii) to avoid local minima traps by introducing a small amount of Gaussian noise into the images. These techniques reduce considerably the incidence of false assignments, but they fail to remove them altogether.

There has been less progress in the study of parallel algorithms for making use of disparity information, despite considerable recent interest in the processing of disparity information by the visual system (Barlow, Blakemore & Pettigrew [1967], Julesz [1971]; Julesz [1971, p. 204] and Sperling [1970]) have both suggested possible schemes. Julesz's model is informal in nature, being more phenomenological than computational: it is very interesting because despite its great simplicity, it displays an astonishing number of properties that are exhibited by the human disparity processing system. Sperling's model is

more formal, but it is difficult to tell how well it would work. This is a problem with all complex parallel methods; they are very expensive to simulate, and it is extremely difficult to derive analytically, from a system with complex non-linear components, quantities that could be measured experimentally. About the best one can hope to do at present is to state criteria that distinguish one family of methods from another, and ask whether those criteria are satisfied by the particular method that we use.

This note enquires about the exact nature of the disparity computation. It is in some sense a correlation, and because that is common knowledge and fairly precise, a deeper characterization has not been sought. But in order to formulate a method of carrying it out, one needs to be very precise about the goals of the computation, and about the constraints imposed upon it by the structure of the three-dimensional world. Unless one uses a method that is based on all and only the correct assumptions, there will be situations in which it will fail unnecessarily.

## Measuring disparity

If a scene is photographed from two slightly different positions, the relative positions of the objects in the scene will differ slightly on the two images. The discrepancies of interest arise from the different distances of the objects from the viewing position, and measurements of the discrepancies contain useful information about the relative distances of the objects. The term binocular disparity refers to the difference in the angle from each eye to a point in the scene, measured relative to some suitably chosen angle of convergence. The central difficulty in defining what is meant by the process of extracting binocular disparity from an image is that disparity has to refer to a physical entity—a point on a visible surface—yet it appears that we compute it at a level far below that at which the world is described in terms of surfaces and objects (Julesz, 1971). It was probably this fact that made so surprising Julesz's conclusion that disparity assignment is a low-level computation.

In order to compute disparity correctly, the following steps must be carried out: first, a particular location on a surface in the scene must be located in one image; second, the identical location must be identified in the other image; third, the relative positions of the two images of that location must be measured. The most interesting and most troublesome part of the process concerns the selection of a location on the viewed surface, and the identification of its two images. The difficulty is that the choice of a point on the surface must be made from the images: if it could be chosen in some absolute way —by lighting it up at that point for example—the problem would be simple.

We are now in a position to understand why the disparity computation is not the same thing as a gray-level correlation. The reason is that gray-level measurements correspond to properties of the image, rather than to properties of the objects being viewed. An $(x, y)$ coordinate pair on an image is an artefact of the transducer, since it does not define a point on a physical surface in a way that allows it to be identified on the other image. The most glaring example of the failure is the case where an image point corresponds to two surface points, the nearer of which is transparent or translucent: a goldfish in a pond is one such case, where the water surface and the goldfish are simultaneously visible. Other examples are provided by figures 5.7.1 and

6.3.2 of Julesz (1971). But the argument applies equally well to the case of a single visible surface, and its consequence is that gray-level matching methods are incorrect. On simple images, the method will succeed, because it is close enough to the right idea: but as Mori et al. (1973) have found, it will not succeed on complex images because it is based on incorrect premises. Their technique of introducing local smearing may be viewed as a way of beginning to identify a point in the image with a point on the physical surface (by adding additional constraints on what is matched): insofar as it does so, the method will become more reliable, but it is probably better to attack the underlying issue directly.

**The use of low-level symbols**

In order to formulate the disparity computation in a usable way, we therefore need to be able to identify surface points from the two images, and match them up. It is clearly fruitless to try to label points of a smooth featureless surface, but if that surface contains a scratch, boundary, or other identifying physical mark that produces a local and fairly sharp change in reflectance, that change in reflectance may be used to define the surface point. Provided that the change in reflectance has been identified and described separately in the two images, the resulting descriptions will correspond to an underlying physical reality. The computation of such a low-level description from one image has been dealt with at length elsewhere (Marr 1974a & 1974b), and is called the low-level symbolic representation of an image. Hence we see that provided stereo matching takes place between two low-level symbolic descriptions, it is a well-founded operation.

Finally, we need to ask whether any reasonably complex measurement could be used—or is there something special about a low-level symbolic description? Simple cell-like measurements are for example nearly suitable, because they are sometimes quite near to low-level assertions; but when assigning disparity values to simple cells, one meets all the usual problems associated with measurements—a whole set of simple cells, at neighboring positions and orientations, corresponds to the underlying scratch or whatever in the image, and it is that complex that needs to be matched against the corresponding complex derived from the other image. If the important matching step is carried out on each individual simple cell measurement, the computation becomes very uneconomical. Hence one may expect that when disparity is actually assigned, the process operates on a very low-level symbolic description. This method will fail only when the low-level descriptions obtained from the two images are very different; but this is comparatively rare, and one seems to notice it when it happens. In any case, this circumstance cannot be dealt with at the same very low level.

The problem has now been reduced to the comparison of two low-level symbolic descriptions, and the assignment of disparity values to pairs of symbols, drawn appropriately from each image. We turn now to examine briefly the rules and constraints to which this process is subject.

**The "use once" condition**

We have seen that an element of a low-level symbolic description of an image corresponds to a physically identifiable entity in a way that an image coordinate does not, and in which measurements made on an image only approximate.

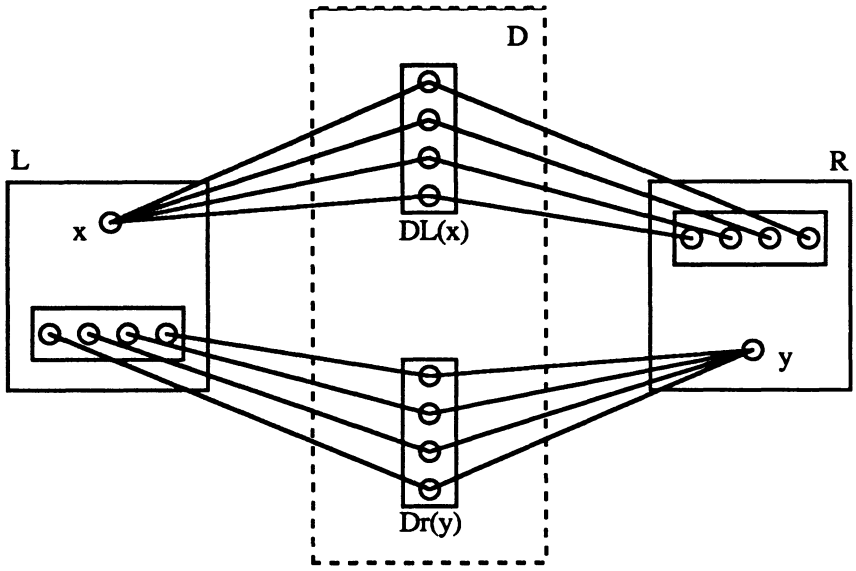



**Figure 1**

Fig. 1. L denotes the collection of left-image symbols, and R the collection of right-image symbols. These are connected through the set D of disparity value symbols. The sets DL and Dr referred to in the text are show in the figure.

This allows one to state the first condition that controls the matching of two low-level symbolic descriptions. It is that each low-level symbol should be assigned exactly one disparity value, which in turn implies that it should be associated with at most one symbol computed from the other image. This is called the "use-once" condition, and it is nonlinear.

The use-once condition may be implemented in the following way. Let L be the set of all left-image low-level symbols, and let R be the set of all right-image low-level symbols. To each element x in L, there correspond several elements in R, one for each of the possible disparity values; and for each element y in R, there is a corresponding set of elements in L. This situation is illustrated in Figure 1. The matching of one element from L with one from R corresponds to the assignment of a single disparity value to both elements, so let us introduce a third set D that consists of collections of elements representing all of the possible disparity values that may be bound to each low-level symbol. In principle, one needs one such collection for each low-level symbol, although members of L and R share elements in D in the appropriate way (see below). The use-once condition translates into the constraint that each element of L and of R may be bound to at most one element of D.

In practice, the set D will be very large unless steps are taken to economize on the number of units that are necessary to represent the disparity values; so let us consider how disparity-representing units may be shared between several elements of L (say). D has to be large enough so that (a) each low-level symbol can find an unused collection in D that can be used for representing its disparity; and (b) the correspondence between L and R through D is well

defined. To accomplish this, the collection of left- (or right-) image symbols that share a given disparity-representing unit should have the property that it is very rare for two to be provoked simultaneously by the image.

When one considers how to construct a parallel network that implements the use-once condition, it is apparent that at least three variables must be accommodated: ascension and declination in the visual field, and disparity. A satisfactory arrangement is shown in Figure 2; in this, the units representing disparity values are arranged in stripes of constant disparity. The collections in D that represent disparity values for left-image symbols (L) lie along the diagonal lines marked Dl and those for right-image symbols lie along the opposing set of lines Dr. Thus D is divided up in two ways into disparity- representing units, which are simultaneously shared in an appropriate fashion by L and R. The connections that implement the use-once condition are clearly marked: they run along Dl and Dr, joining places that contain representations of all possible disparity values that could be bound to a given left- or right-generated symbol. (In a neural implementation of this scheme, such connections would be inhibitory.)

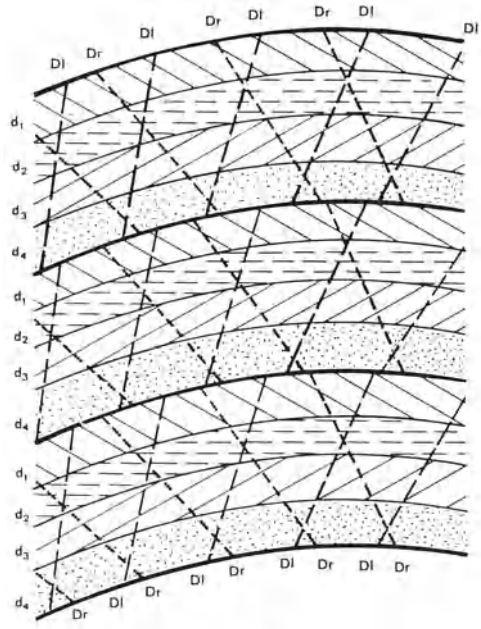



**Figure 2**

There are interesting differences between the implications for neurophysiology of these ideas, and of the model of Julesz. First, the important part of the computation involves constraints on the disparity values that may be bound to low-level symbols. The magnets in Julesz's model seem, however, to correspond to rather unspecific local disparity values, and we saw earlier that the disparity computation cannot be accurately expressed in these terms. The second point rests on the way in which disparity-representing sets in D are assigned to L and to R. The most economical way of forming the collections of low-level symbols, that are to use the same disparity-representing units, is probably to group together all symbols that describe a small region and a small

orientation range. Only very rarely will two such symbols be used simultaneously. If some scheme of this sort were being used, it would account for the existence of cells that behave sensitively to disparity, but are relatively tolerant to position and orientation (Hubel and Wiesel, 1970). Furthermore, it would be within these units that the main disparity computation is being carried out, and between which the governing connections should be made. One would not expect to find other cells, expressing a free-floating disparity value in a "region" of the image, because the essence of the computation requires that it be carried out on bindings to low-level symbols. The model of Julesz would, I think, lead one to expect such cells.

### Disparity is continuous almost everywhere

The use-once condition must be satisfied by the final assignment of disparity values to the low-level symbolic descriptions, but it is not much help in finding it. When applied to a random-dot stereogram, it will ensure that the description of each dot, or group of dots, in one image is mated with not more than one similar description computed from the other image; and a solution that satisfies this and leaves very few dots out will probably be correct (see the next condition). But there is another useful property of the real world that can be introduced with advantage to speed the analysis. It is that except at object boundaries, disparity is a function that varies smoothly over the image. Fine texture, which is the best source of disparity information about a surface, will be represented at the lowest level by assertions about very small features; and except at object boundaries, the disparity values that become bound to neighboring symbols will be about the same. This fact allows the existence of an interaction that "proposes" the disparity used at one point as a strong candidate for the value at neighboring points: it corresponds in Julesz's (1971) model to the lateral coupling provided by the small springs that join adjacent magnets. The implementation of this constraint in Sperling's (1970) model is obscure.

The implementation of a suggestion is one of those questions that it is not profitable to pursue in detail, because of the difficulty in testing, either physiologically or computationally, the conclusions to which one may be led. I shall therefore make only three points about it. The first is that, in principle, one would like a suggestion to influence the route to a solution, without disturbing the values in the solution once they are found. This implies a time-dependence in the interaction. Second, in order that the solution may be stable, one would probably also need to add a small DC component. The third point, and one that may actually be useful, concerns the geometry of the suggested interactions, and this is shown in Figure 2. There are connections between all disparity units that represent similar disparity values, and that refer to symbols in nearby portions of the visual field. In a neural implementation, they would be excitatory.

### Goodness-of-fit

The final important aspect of disparity measurement is the question of how satisfactory a solution is. Julesz emphasized the need for such a measure, and in his model, it corresponds to the total potential energy in the two superimposed assemblies of magnets. Sperling (1970) also used a potential energy measure in his formulation. Julesz showed that in an ambiguous stereogram, we perceive the better correlated solution even if both have quite high corre-

lations. This is good evidence that the matching process is parallel rather than serial, and that the goodness-of-fit measure is computed on a local basis.

In an implementation of the kind that we are discussing, the goodness-of-fit of a solution may be measured by the proportion of left- and right-image symbols that become bound to disparity values. In a perfect solution, the proportion will be 1.0; and an inappropriate disparity binding will have the effect of depressing the proportion of correct bindings in its neighborhood. The local goodness-of-fit function would affect the confidence with which disparity assignments are made locally, i.e., the strength with which they are asserted—which in turn would affect the potency with which they are suggested to nearby regions. The goodness-of-fit function would therefore be implemented by a unit that depressed the local disparity-representing units by an amount that depended upon the proportion of image symbols in the vicinity that have been assigned disparity values. I do not see how to test for the presence of such units, except by trial and error.

### Summary of the disparity interactions

The interactions described above are now drawn together, and the conditions that are necessary to an implementation of this kind are made explicit.

(1) The disparity assignment is made as a result of a matching operation performed on two low-level symbolic descriptions, computed independently from the left and right images.

(2) The matching is implemented by applying conditions to the process of binding symbolic disparity descriptors to the low-level symbols. These constraints are the use-once condition, the suggestion interaction, and maximizing the goodness of fit.

(3) The use-once condition requires interactions whose geometry appears in Figure 2. These interactions inhibit the confidence of those assignments that they connect.

(4) The suggestion interactions have the geometry shown also in Figure 2. They connect disparity descriptors that represent similar disparity values, and that are capable of being bound to low-level symbols referring to neighboring positions in the visual field. Such interactions would probably have a time-dependent component as well as a DC component.

(5) Maximizing the goodness of fit of a solution would have to be implemented by a local goodness-of-fit function, that measures the proportion of low-level symbols that have successfully been bound to a disparity value, and that affects the confidence level of the bindings in that local region.

### Discussion

I shall not attempt in this note to derive any properties of the above system. I have been unable to make much progress with an analytical approach to the problem, and the amount of time required for a computational study is very large. The approach set out here does, however, illustrate (a) how the disparity computation may be well founded; (b) the importance of low-level symbols in the formulation; and (c) how the important constraints may, at least in a general way, be represented by connections with a straightforward geometry. This kind of geometrical arrangement is one that it is becoming possible to detect.

The second large issue concerns the way in which disparity information

may be used. It is one thing to assign disparity values to low-level symbols, and quite another to divide up an image into regions on the basis of disparity information alone, and compute a description of the spatial extent of each region. For example, in any stereogram, the orientation information associated with the small squares or groups of squares that are actually matched bears no relation to the orientation of the edge at which disparity changes. One way of computing the higher, induced edges would, of course, be to treat disparity like intensity, and subject its values to a process like that used to obtain a low-level symbolic description from an intensity array (Marr, 1974b). This method seems somewhat clumsy, however, because disparity is not the only kind of information (excluding intensity) from which directions and boundaries may be computed locally. Texture changes are another example, and so are more abstract outlines, like the envelope of a sparse tree in winter, or the boundary defined by a row of small, separated bushes across a garden. One would like to know whether all of these problems may be dealt with by a single method that can describe configurations of "places" in an image—these places being identified by a rather simple kind of local measurement made on the relevant type of information. There seems to be clear evidence (and a definite computational need) for such a mechanism one of its main functions being to set up an orientation in the image at a point, to describe configurations of places relative to that orientation, and to influence the direction relative to which local shapes in an image are described. It is, however, far from clear whether one such mechanism would suffice to service all of the demands of this kind, or whether the slightly differing computational requirements force the existence of a number of separate, but similar, mechanisms. The question will be raised elsewhere (Marr, 1975).

*Acknowledgment.* Work reported herein was conducted at the Artificial Intelligence Laboratory, a Massachusetts Institute of Technology research program supported in part by the Advanced Research Projects Agency of the Department of Defense and monitored by the Office of Naval Research under Contract number N00014-70-A-0362-0005.

## REFERENCES

Barlow HB, Blakemore C and Pettigrew JD (1967): The neural mechanism of binocular depth discrimination. *J. Physiol. (Lond.)*, 193:327-342

Hubel DH and Wiesel TN (1970): Stereoscopic vision in macaque monkey. *Nature*, 225:41-42

Julesz B (1971): *Foundations of Cyclopean Perception.* Chicago: The University of Chicago Press

Marr D (1975): Configurations, regions, and simple texture vision

Marr D (1974a): The purpose of low-level vision. MIT Artificial Intelligence Laboratory Memo 324

Marr D (1974b): The low-level symbolic representation of intensity changes in an image. MIT Artificial Intelligence Laboratory Memo 325

Mori K, Kidode M and Asada H (1973): An iterative prediction and correction method for automatic stereocomparation. *Computer Graphics and Image Processing* 2:393-401

Sperling G (1970): Binocular fusion: a physical and neural theory. *J Am Physiol* 83:461-534

# Cooperative Computation of Stereo Disparity

A cooperative algorithm is derived for extracting disparity information from stereo image pairs.

D. Marr and T. Poggio

Perhaps one of the most striking differences between a brain and today's computers is the amount of "wiring." In a digital computer the ratio of connections to components is about 3, whereas for the mammalian cortex it lies between 10 and 10,000 (*1*).

Although this fact points to a clear structural difference between the two, this distinction is not fundamental to the nature of the information processing that each accomplishes, merely to the particulars of how each does it. In Chomsky's terms (*2*), this difference affects theories of performance but not theories of competence, because the nature of a computation that is carried out by a machine or a nervous system depends only on a problem to be solved, not on the available hardware (*3*). Nevertheless, one can expect a nervous system and a digital computer to use different types of algorithm, even when performing the same underlying computation. Algorithms with a parallel structure, requiring many simultaneous local operations on large data arrays, are expensive for today's computers but probably well-suited to the highly interactive organization of nervous systems.

The class of parallel algorithms includes an interesting and not precisely definable subclass which we may call cooperative algorithms (*3*). Such algorithms operate on many "input" elements and reach a global organization by way of local, interactive constraints. The term "cooperative" refers to the way in which local operations appear to cooperate in forming global order in a well-regulated manner. Cooperative phenomena are well known in physics (*4, 5*), and it has been proposed that they may play an important role in biological systems as well (*6–10*). One of the earliest suggestions along these lines was made by Julesz (*11*), who maintains that stereoscopic fusion is a cooperative process. His model, which consists of an array of dipole magnets with springs coupling the tips of adjacent dipoles, represents a suggestive metaphor for this idea. Besides its biological relevance, the extraction of stereoscopic information is an important and yet unsolved problem in visual information processing (*12*). For this reason—and also as a case in point—we describe a cooperative algorithm for this computation.

In this article, we (i) analyze the computational structure of the stereo-disparity problem, stating the goal of the computation and characterizing the associated local constraints; (ii) describe a cooperative algorithm that implements this computation; and (iii) exhibit its performance on random-dot stereograms. Although the problem addressed here is not directly related to the question of

D. Marr is at the Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge 02139. T. Poggio is at the Max-Planck Institut für Biologische Kybernetik, 74 Tubingen 1, Spemannstrasse 38, Germany.

how the brain extracts disparity information, we shall briefly mention some questions and implications for psychophysics and neurophysiology.

### Computational Structure of the Stereo-Disparity Problem

Because of the way our eyes are positioned and controlled, our brains usually receive similar images of a scene taken from two nearby points at the same horizontal level. If two objects are separated in depth from the viewer, the relative positions of their images will differ in the two eyes. Our brains are capable of measuring this disparity and of using it to estimate depth.

Three steps (S) are involved in measuring stereo disparity: (S1) a particular location on a surface in the scene must be selected from one image; (S2) that same location must be identified in the other image; and (S3) the disparity in the two corresponding image points must be measured.

If one could identify a location beyond doubt in the two images, for example by illuminating it with a spot of light, steps S1 and S2 could be avoided and the problem would be easy. In practice one cannot do this (Fig. 1), and the difficult part of the computation is solving the correspondence problem. Julesz found that we are able to interpret random-dot stereograms, which are stereo pairs that consist of random dots when viewed monocularly but fuse when viewed stereoscopically to yield patterns separated in depth. This might be thought surprising, because when one tries to set up a correspondence between two arrays of random dots, false targets arise in profusion (Fig. 1). Even so, we are able to determine the correct correspondence. We need no other cues.

In order to formulate the correspondence computation precisely, we have to examine its basis in the physical world. Two constraints (C) of importance may be identified (*13*): (C1) a given point on a physical surface has a unique position in space at any one time; and (C2) matter is cohesive, it is separated into objects, and the surfaces of objects are generally smooth compared with their distance from the viewer.

These constraints apply to locations on a physical surface. Therefore, when we translate them into conditions on a computation we must ensure that the items to which they apply there are in one-to-one correspondence with well-defined locations on a physical surface. To do this, one must use surface markings, normal surface discontinuities, shadows, and so forth, which in turn means using predicates that correspond to changes in intensity. One solution is to obtain a primitive description [like the primal sketch (*14*)] of the intensity changes present in each image, and then to match these descriptions. Line and edge segments, blobs, termination points, and tokens, obtained from these by grouping, usually correspond to items that have a physical existence on a surface.

The stereo problem may thus be reduced to that of matching two primitive descriptions, one from each eye. One can think of the elements of these descriptions as carrying only position information, like the white squares in a random-dot stereogram, although in practice there will exist rules about which matches between descriptive elements are possible and which are not. The two physical constraints C1 and C2 can now be translated into two rules (R) for how the left and right descriptions are combined:

R1) *Uniqueness.* Each item from each image may be assigned at most one disparity value. This condition relies on the assumption that an item corresponds to something that has a unique physical position.

R2) *Continuity.* Disparity varies smoothly almost everywhere. This condition is a consequence of the cohesiveness of matter, and it states that only a small fraction of the area of an image is composed of boundaries that are discontinuous in depth.

In real life, R1 cannot be applied simply to gray-level points in an image. The simplest counterexample is that of a goldfish swimming in a bowl: many points in the image receive contributions from the bowl and from the goldfish. Here, and in general, a gray-level point is in only implicit correspondence with a physical location, and it is therefore impossible to ensure that gray-level points in the two images correspond to exactly the same physical position. Sharp changes in intensity are usually due either to the goldfish, to the bowl, or to a reflection, and therefore define a single physical position precisely.

### A Cooperative Algorithm

By constructing an explicit representation of the two rules, we can derive a cooperative algorithm for the computation. Figure 2a exhibits the geometry of the rules in the simple case of a one-dimensional image. $L_x$ and $R_x$ represent the positions of descriptive elements on the left and right images. The thick vertical and horizontal lines represent lines of sight from the left and right eyes, and their intersection points correspond to possible disparity values. The dotted diagonal lines connect points of constant disparity.

The uniqueness rule R1 states that only one disparity value may be assigned to each descriptive element. If we now think of the lines in Fig. 2a as a network, with a node at each intersection, this means that only one node may be switched on along each horizontal or vertical line.

The continuity rule R2 states that disparity values vary smoothly almost everywhere. That is, solutions tend to spread along the dotted diagonals.

If we now place a "cell" at each node (Fig. 2b) and connect it so that it inhibits cells along the thick lines in the figure and excites cells along the dotted lines, then, provided the parameters are appropriate, the stable states of such a network will be precisely those in which the two rules are obeyed. It remains only to show that such a network will converge to a stable state. We were able to carry out a combinatorial analysis [as in (*9, 15*)] which established its convergence for random-dot stereograms (*16*).

This idea may be extended to two-dimensional images simply by making the local excitatory neighborhood two dimensional. The structure of each node in the network for two-dimensional images is shown in Fig. 2c.

A simple form of the resulting algorithm (*3*) is given by the following set of difference equations:

$$\mathbf{C}^{(n+1)} = \sigma\{\Xi(\mathbf{C}^{(n)}) + \mathbf{C}^{(0)}\} \qquad (1)$$

that is,

$$C_{xyd}^{(n+1)} = \sigma\Big\{ \sum_{x'y'd' \in S(xyd)} C_{x'y'd'}^{(n)} - \epsilon \sum_{x'y'd' \in O(xyd)} C_{x'y'd'}^{(n)} + C_{xyd}^{(0)} \Big\} \qquad (2)$$

where $C_{xyd}^{(n)}$ represents the state of the node or cell at position $(x,y)$ with disparity $d$ at iteration $n$, $\Xi$ is the linear operator that embeds the local constraints ($S$ and $O$ are the circular and thick line neighborhoods of the cell $xyd$ in Fig. 2c), $\epsilon$ is the "inhibition" constant, and $\sigma$ is a sigmoid function with range [0, 1]. The state $C_{xyd}^{(n+1)}$ of the corresponding node at time $(n + 1)$ is thus determined by a nonlinear operator on the output of a linear transformation of the states of neighboring cells at time $n$.

The desired final state of the computation is clearly a fixed point of this al-

gorithm; moreover, any state that is inconsistent with the two rules is not a stable fixed point. Our combinatorial analysis of this algorithm shows that, when $\sigma$ is a simple threshold function, the process converges for a rather wide range of parameter values (*16*). The specific form of the operator is apparently not very critical.

Noniterative local operations cannot solve the stereo problem in a satisfactory way (*11*). Recurrence and nonlinearity are necessary to create a truly cooperative algorithm that cannot be decomposed into the superposition of local operations (*17*). General results concerning such algorithms seem to be rather difficult to obtain, although we believe that one can usually establish convergence in probability for specific forms of them.

### Examples of Applying the Algorithm

Random-dot stereograms offer an ideal input for testing the performance of the algorithm, since they enable one to bypass the costly and delicate process of transforming the intensity array received by each eye into a primitive description (*14*). When we view a random-dot stereogram, we probably compute a description couched in terms of edges rather than squares, whereas the inputs to our algorithm are the positions of the white squares. Figures 3 to 6 show some examples in which the iterative algorithm successfully solves the correspondence problem, thus allowing disparity values to be assigned to items in each image. Presently, its technical applications are limited only by the preprocessing problem.

This algorithm can be realized by various mechanisms, but parallel, recurrent, nonlinear interactions, both excitatory and inhibitory, seem the most natural. The difference equations set out above would then represent an approximation to the differential equations that describe the dynamics of the network.

### Implications for Biology

We have hitherto refrained from discussing the biological problem of how stereopsis is achieved in the mammalian brain. Our analyses of the computation, and of the cooperative algorithm that implements it, raise several precise questions for psychophysics and physiology. An important preliminary point concerns the relative importance of neural fusion and of eye movements for stereopsis. The underlying question is whether there are many disparity "layers" (as our algorithm requires), or whether there are just three "pools" (*18*)—crossed, uncrossed, and zero disparity. Most physiologists and psychologists seem to accept the existence of numerous, sharply tuned binocular "disparity detectors," whose peak sensitivities cover a wide range of disparity values (*19, 20*). We do not believe that the available evidence is decisive (*21*), but an answer is critical to the biological relevance of our analysis. If, for example, there were only three pools or layers with a narrow range of disparity sensitivities, the problem of false targets is virtually removed, but at the expense of having to pass the convergence plane of the eyes across a surface in order to achieve fusion. Psychophysical experiments may provide some insight into this problem, but we believe that only physiology is capable of providing a clear-cut answer.

If this preliminary question is settled in favor of a "multilayer" cooperative algorithm, there are several obvious implications of the network (Fig. 2) at the physiological level: (i) the existence of many sharply tuned disparity units that are rather insensitive to the nature of the descriptive element to which they may refer; (ii) organization of these units into disparity layers (or stripes or columns); (iii) the presence of reciprocal excitation within each layer; and (iv) the presence of reciprocal inhibition between layers along the two lines of sight. Ideally, the inhibition should exhibit the characteristic "orthogonal" geometry of the thick

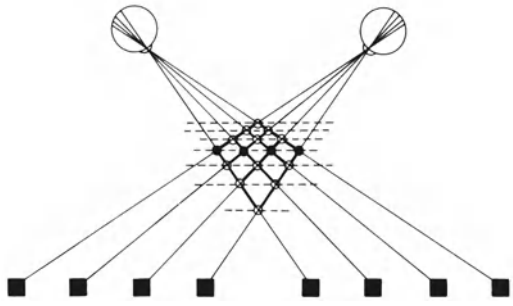
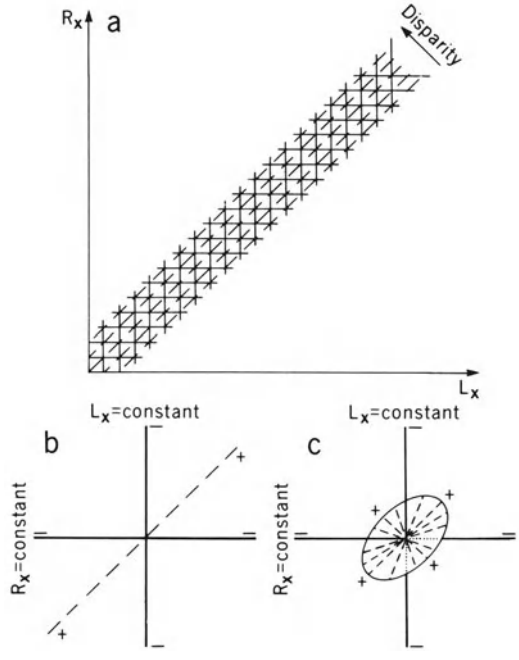


Fig. 1 (left). Ambiguity in the correspondence between the two retinal projections. In this figure, each of the four points in one eye's view could match any of the four projections in the other eye's view. Of the 16 possible matchings only four are correct (closed circles), while the remaining 12 are "false targets" (open circles). It is assumed here that the targets (closed squares) correspond to "matchable" descriptive elements obtained from the left and right images. Without further constraints based on global considerations, such ambiguities cannot be resolved. Redrawn from Julesz (*11*, figure 4.5-1). Fig. 2 (right). The explicit structure of the two rules R1 and R2 for the case of a one-dimensional image is represented in (a), which also shows the structure of a network for implementing the algorithm described by Eq. 2. Solid lines represent "inhibitory" interactions, and dotted lines represent "excitatory" ones. The local structure at each node of the network in (a) is given in (b). This algorithm may be extended to two-dimensional images, in which case each node in the corresponding network has the local structure shown in (c). Such a network was used to solve the stereograms exhibited in Figs. 3 to 6.

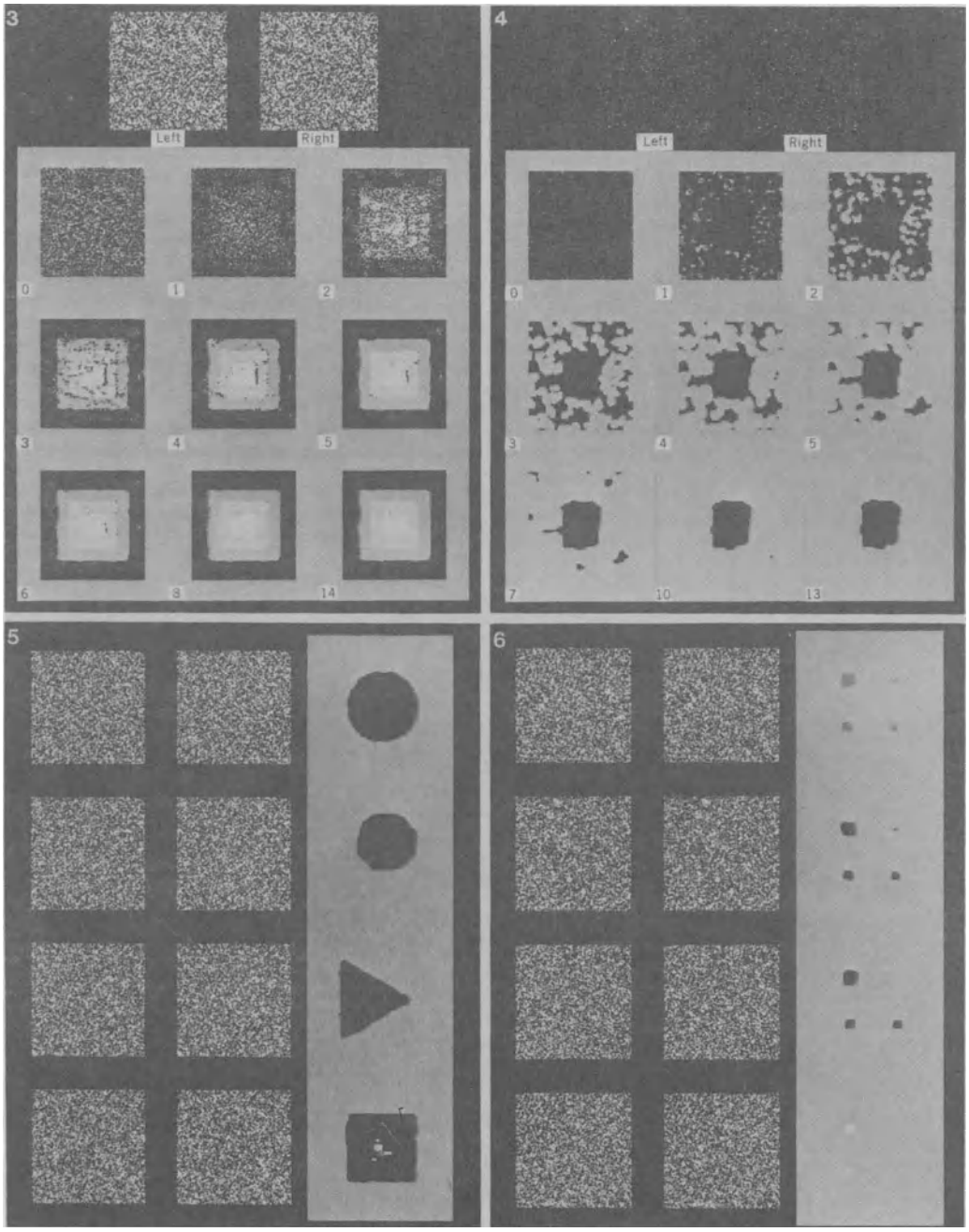


lines in Fig. 2, but slight deviations may be permissible (*16*).

At the psychophysical level, several experiments (under stabilized image conditions) could provide critical evidence for or against the network: (i) results about the size of Panum's area and the number of disparity "layers"; (ii) results about "pulling" effects in stereopsis (*20*); and (iii) results about the relationship between disparity and the minimum fusable pattern size (Fig. 6).

**Discussion**

Our algorithm performs a computation that finds a correspondence function between two descriptions, subject to the two constraints of uniqueness and conti-



nuity. More generally, if one has a situation where allowable solutions are those that satisfy certain local constraints, a cooperative algorithm can often be constructed so as to find the nearest allowable state to an initial one. Provided that the constraints are local, use of a cooperative algorithm allows the representation of global order, to which the algorithm converges, to remain implicit in the network's structure.

The interesting difference between this stereo algorithm and standard correlation techniques is that one is not required to specify minimum or maximum correlation areas to which the analysis is subsequently restricted. Previous attempts at implementing automatic stereocomparison through local correlation measurement have failed in part because no single neighborhood size is always correct (*12*). The absence of a "characteristic scale" is one of the most interesting properties of this algorithm, and it is a central feature of several cooperative phenomena (*22*). We conjecture that the matching operation implemented by the algorithm represents in some sense a generalized form of correlation, subject to the a priori requirements imposed by the constraints. The idea can easily be generalized to different constraints and to other forms of equations 1 or 2, and it is technically quite appealing.

Cooperative algorithms may have many useful applications [for example, to make best matches for associative retrieval problems (*15*)], but their relevance to early processing of information by the brain remains an open question (*23*). Although a range of early visual processing problems might yield to a cooperative approach ["filling-in" phenomena, subjective contours (*24*), grouping, figural reinforcement, texture "fields," and the correspondence problem for motion], the first important and difficult task in problems of biological information processing is to formulate the underlying computation precisely (*3*). After that, one can study good algorithms for it. In any case, we believe that an experimental answer to the question of whether depth perception is actually a cooperative process is a critical prerequisite to further attempts at analyzing other perceptual processes in terms of similar algorithms.

## Summary

The extraction of stereo-disparity information from two images depends upon establishing a correspondence between them. In this article we analyze the nature of the correspondence computation and derive a cooperative algorithm that implements it. We show that this algorithm successfully extracts information from random-dot stereograms, and its implications for the psychophysics and neurophysiology of the visual system are briefly discussed.

Figs. 3 to 6. The results of applying the algorithm defined by Eq. 2 to two random-dot stereograms. Fig. 3. The initial state of the network $C^{(0)}$ is defined by the input such that a node takes the value 1 if it occurs at the intersection of a 1 in the left and right eyes (Fig. 2), and it has the value 0 otherwise. The network iterates on this initial state, and the parameters used here, as suggested by the combinatorial analysis, were $\theta = 3.0$, $\epsilon = 2.0$, and $M = 5$, where $\theta$ is the threshold and $M$ is the diameter of the "excitatory" neighborhood illustrated in Fig. 2c. The stereograms themselves are labeled *Left* and *Right*, the initial state of the network as 0, and the state after $n$ iterations is marked as such. To understand how the figures represent states of the network, imagine looking at it from above. The different disparity layers in the network lie in parallel planes spread out horizontally, so that the viewer is looking down through them. In each plane, some nodes are on and some are off. Each of the seven layers in the network has been assigned a different gray level, so that a node that is switched on in the top layer (corresponding to a disparity of +3 pixels) contributes a dark point to the image, and one that is switched on in the lowest layer (disparity of −3) contributes a lighter point. Initially (iteration 0) the network is disorganized, but in the final state stable order has been achieved (iteration 14), and the inverted wedding-cake structure has been found. The density of this stereogram is 50 percent. Fig. 4. The algorithm of Eq. 2, with parameter values given in the legend to Fig. 3, is capable of solving random-dot stereograms with densities from 50 percent to less than 10 percent. For this and smaller densities, the algorithm converges increasingly slowly. If a simple homeostatic mechanism is allowed to control the threshold $\theta$ as a function of the average activity (number of "on" cells) at each iteration [compare (*15*)], the algorithm can solve stereograms whose density is very low. In this example, the density is 5 percent and the central square has a disparity of +2 relative to the background. The algorithm "fills in" those areas where no dots are present, but it takes several more iterations to arrive near the solution than in cases where the density is 50 percent. When we look at a sparse stereogram, we perceive the shapes in it as cleaner than those found by the algorithm. This seems to be due to subjective contours that arise between dots that lie on shape boundaries. Fig. 5. The disparity boundaries found by the algorithm do not depend on their shapes. Examples are given of a circle, an octagon (notice how well the difference between them is preserved), and a triangle. The fourth example shows a square in which the correlation is 100 percent at the boundary but diminishes to 0 percent in the center. When one views this stereogram, the center appears to shimmer in a peculiar way. In the network, the center is unstable. Fig. 6. The width of the minimal resolvable area increases with disparity. In all four stereograms the pattern is the same and consists of five circles with diameters of 3, 5, 7, 9, and 13 dots. The disparity values exhibited here are +1, +2, +3, and +6, and for each pattern we show the state of the network after ten iterations. As far as the network is concerned, the last pair (disparity of +6) is uncorrelated, since only disparities from −3 to +3 are present in our implementation. After ten iterations, information about the lack of correlation is preserved in the two largest areas.

### References and Notes

1. D. A. Sholl, *The Organisation of the Cerebral Cortex* (Methuen, London, 1956). The comparison depends on what is meant by a component. We refer here to the level of a gate and of a neuron, respectively.
2. A. N. Chomsky, *Aspects of the Theory of Syntax* (MIT Press, Cambridge, Mass., 1965).
3. D. Marr and T. Poggio, *Neurosci. Res. Prog. Bull.*, in press (also available as *Mass. Inst. Technol. Artif. Intell. Lab. Memo 357*).
4. H. Haken, Ed. *Synergetics-Cooperative Phenomena in Multicomponent Systems* (Teuber, Stuttgart, 1973).
5. H. Haken, *Rev. Mod. Phys.* **47**, 67 (1975).
6. J. D. Cowan, *Prog. Brain Res.* **17**, 9 (1965).
7. H. R. Wilson and J. D. Cowan, *Kybernetik* **13**, 55 (1973).
8. M. Eigen, *Naturwissenschaften* **58**, 465 (1971).
9. P. H. Richter, paper contributed to a competition of the Bavarian Academy of Science, Max-Planck-Institut für Biophysikalische Chemie, 1974.
10. A. Gierer and H. Meinhardt, *Kybernetik* **12**, 30 (1972).
11. B. Julesz, *Foundations of Cyclopean Perception* (Univ. of Chicago Press, Chicago, 1971).
12. K. Mori, M. Kidode, H. Asada, *Comp. Graphics Image Process.* **2**, 393 (1973).
13. D. Marr, *Mass. Inst. Technol. Artif. Intell. Lab. Memo 327* (1974).
14. ———, *Philos. Trans. R. Soc. London Ser. B* **275**, 483 (1976).
15. ———, *ibid.* **252**, 23 (1971). See especially section 3.1.2.
16. ——— and T. Poggio, in preparation.
17. T. Poggio and W. Reichardt, *Q. Rev. Biophys.*, in press.
18. W. Richards, *J. Opt. Soc. Am.* **62**, 410 (1971).
19. H. B. Barlow, C. Blakemore, J. D. Pettigrew, *J. Physiol. (London)* **193**, 327 (1967); J. D. Pettigrew, T. Nikara, P. O. Bishop, *Exp. Brain Res.* **6**, 391 (1968); C. Blakemore, *J. Physiol. (London)* **209**, 155 (1970).
20. B. Julesz and J.-J. Chang, *Biocybernetics* **22**, 107 (1976).
21. D. H. Hubel and T. N. Wiesel, *Nature (London)* **225**, 41 (1970).
22. K. G. Wilson, *Rev. Mod. Phys.* **47**, 773 (1975).
23. Julesz (*11*), Cowan (*6*), and Wilson and Cowan (*7*) were the first to discuss explicitly the cooperative aspect of visual information processing. Much has been published recently on possible cooperative processes in nervous systems, ranging from the "catastrophe" literature [E. C. Zeeman, *Sci. Am.* **234**, 65 (April 1976)] to various attempts of more doubtful credibility. There has hitherto been no careful study of a cooperative algorithm in the context of a carefully defined computational problem [but see (*15*)], although algorithms that may be interpreted as cooperative were discussed, for instance, by P. Dev [*Int. J. Man-Mach. Stud.* **7**, 511 (1975)] and by A. Rosenfeld, R. A. Hummel, and S. W. Zucker [*Syst. Man. Cybern.* **6**, 420 (1976)]. In particular neither Dev nor J. I. Nelson [*J. Theor. Biol.* **49**, 1 (1975)] formulated the computational structure of the stereo-disparity problem. As a consequence, the resulting geometry of the inhibition between their disparity detectors does not correspond to ours (Fig. 2c) and apparently fails to provide a satisfactory algorithm.
24. S. Ullmann, *Mass. Inst. Technol. Artif. Intell. Lab. Memo 367* (1967); also in *Biol. Cybern.*, in press.
25. We thank W. Richards for valuable discussions, H. Lieberman for making it easy to create stereograms, and K. Prendergast for preparing the figures. The research described was done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the laboratory's artificial intelligence research is provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contract N00014-75-C-0643; T.P. acknowledges the support of the Max-Planck-Gesellschaft during his visit to the Massachusetts Institute of Technology.

Biol. Cybernetics 28, 223—239 (1978)

Biological Cybernetics


# Analysis of a Cooperative Stereo Algorithm

D. Marr*, G. Palm** and T. Poggio**

**Abstract.** Marr and Poggio (1976) recently described a cooperative algorithm that solves the correspondence problem for stereopsis. This article uses a probabilistic technique to analyze the convergence of that algorithm, and derives the conditions governing the stability of the solution state. The actual results of applying the algorithm to random-dot stereograms are compared with the probabilistic analysis. A satisfactory mathematical analysis of the asymptotic behaviour of the algorithm is possible for a suitable choice of the parameter values and loading rules, and again the actual performance of the algorithm under these conditions is compared with the theoretical predictions. Finally, some problems raised by the analysis of this type of "cooperative" algorithm are briefly discussed.

## 1. Introduction

The extraction of stereo-disparity information from two images depends upon establishing a correspondence between them. In a recent article, Marr and Poggio (1976) analyzed the nature of the correspondence computation and derived a cooperative algorithm that implements it. Although several examples were given of the performance of the algorithm on random-dot stereograms (Marr and Poggio 1976, Figs. 3–6), space did not permit a thorough analysis of the fixed points of the algorithm, or of its convergence. In this article, we shall examine these issues in detail.

### *1.1. Computational Structure of the Correspondence Problem*

Marr and Poggio (1976) argued that the stereo problem may be reduced to that of matching two primitive descriptions, one from each eye. They showed that the central problem is to find a correspondence between the left and right descriptions, that satisfies the two rules (p. 284 and Marr, 1974):

*(R1) Uniqueness.* Each item from each image may be assigned at most one disparity.

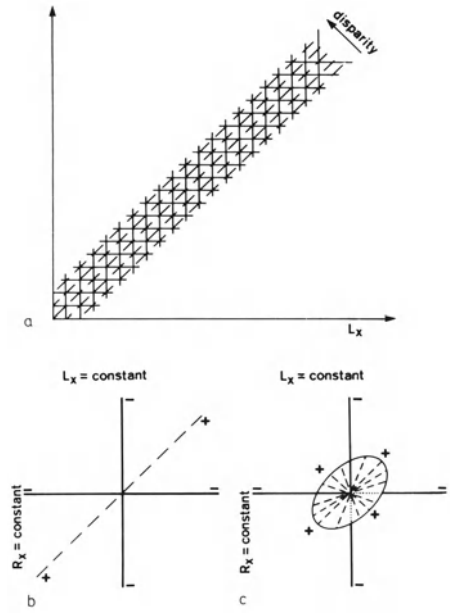



**Fig. 1a—c.** **a** shows the explicit structure of the two rules *R1* and *R2* for the case of a one-dimensional image, and it also represents the structure of a network for implementing the algorithm described by (1). Solid lines represent "inhibitory" interactions, and dotted lines represent "excitatory" ones. **b** Gives the local structure at each node of the network **a**. This algorithm may be extended to two-dimensional images, in which case each node in the corresponding network has the local structure shown in **c**. (Marr and Poggio, 1976, Fig. 2)

* Massachusetts Institute of Technology, Department of Psychology, Cambridge, MA, USA
** Max-Planck-Institut für Biologische Kybernetik, Tübingen, FRG

Reprinted with permission of Springer Verlag Heidelberg from *Biological Cybernetics*, Volume 28, pp 223-239 (1978).

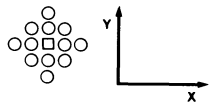



**Fig. 2.** The excitatory neighborhood (Fig. 1c) used in our implementation has a diameter of 5, and contains 13 cells. The central cell, marked by a square, receives at most 12 excitatory inputs from its neighbours

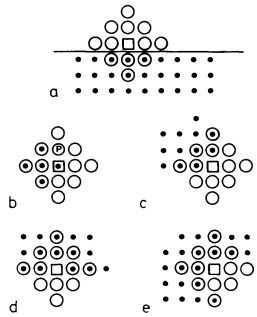



**Fig. 3a—e.** The total excitatory contribution for various configurations of "on" cells. The excitatory neighborhood (Fig. 2) is shown with open circles, except for the central cell which is indicated by a square because it makes no contribution to the total excitation. With a threshold of 4.0: **a** shows that a flat border will grow in the absence of inhibition, **b** exhibits the smallest stable configuration, **c** the sharpest stable convexity, and **d** and **e** show concavities that fill in

*(R2) Continuity.* Disparity varies smoothly almost everywhere.

By constructing an explicit geometrical representation of these two rules (Fig. 1c), they were able to derive a cooperative algorithm that implements them. If one thinks of Figure 1a as a network, with a cell at each node, the uniqueness rule $R1$ means that only one cell is "on" along each vertical or horizontal line (the line of sight from the left and right eyes); and the continuity rule $R2$ implies that solutions (its asymptotic states) tend to spread along the dotted diagonals (lines of constant disparity).

In order to implement these rules, each cell sends "inhibitory" connections to all other cells along the same vertical and horizontal lines, and excitatory connections along its diagonal. This gives the local network geometry shown in Figure 1b. For a two-dimensional image, the only change needed is to make the excitatory neighborhood two-dimensional, which gives the local geometry shown in Figure 1c.

Let $C^t_{x,y;d}$ denote the state at time $t$ of the cell corresponding to coordinate $(x, y)$ on the left retina, matching position $(x+d, y)$ on the right retina. Let $S(xyd)$ denote its excitatory neighborhood (the disc in Fig. 1c), and $O(xyd)$ its inhibitory neighborhood (the horizontal and vertical lines in Figure 1c). The algorithm implemented by the network may be written (Marr and Poggio 1976, Equation (2))

$$C^{t+1}_{x,y;d} = \sigma\left\{\sum_{x',y',d' \in S(x,y,d)} C^t_{x',y';d'} - \varepsilon \sum_{x',y',d' \in O(x,y,d)} C^t_{x',y';d'} + C^0_{x,y;d}\right\} \quad (1)$$

where $\sigma$ is a threshold function that takes values 0 or 1, and $\varepsilon$ is an "inhibition" constant.

This article is concerned with the properties of the algorithms defined by (1) or, equivalently, with the behavior of the corresponding network (Fig. 1). The two inputs to the algorithm or network, from which the initial state of the network is determined, are usually two matrices whose entries consist of 0's and 1's. The second matrix is constructed from the first by $x$-translations of regions of it. As we shall discuss later the algorithm defined by (1) has some analogies with games like "life".

The plan of the paper is as follows: Section 2 describes the loading rules, which determine the initial state from the input stereograms, and also defines the algorithm precisely. The relations between the fixed points of the algorithm and the states that satisfy the two conditions $R1$ and $R2$ are then discussed (Section 3). A probabilistic approach to the convergence of the algorithm is outlined in Section 4. Actual computer simulations of the algorithm are compared with the probabilistic analysis, and the range of parameter values that yield a "nice" convergence is discussed. Some special situations are also analyzed (Section 5). A suitable (and restrictive) choice of the parameter values in (1) allows a satisfactory mathematical analysis of the algorithm: Section 6 is devoted to such an approach. Finally, we briefly discuss the mathematical problems raised by the analysis of this type of "cooperative" algorithm.

## 2. The Algorithm

### 2.1. Loading Conditions

Let the positions on the left and right retinas be denoted by $L_{x,y}$ and $R_{x,y}$ respectively. These arrays take the values 0, indicating the absence of a feature, or 1, indicating the presence. The initial condition of the network, for stereogram $L$, $R$ is given by

$$C^0_{x,y;d} = L_{x,y} \cdot R_{x+d,y} \quad (2)$$

within the appropriate range $d$ of disparity.

### 2.2. The Algorithm

The relation between states at times $t$ and $t+1$, is given by the recurrence relation (1), where $\sigma$ is a sigmoid

function in general, and here is taken to be the threshold function

$$\sigma(u)=1 \quad \text{if} \quad u \geqq \theta,$$
$$=0 \quad \text{otherwise}. \qquad (3)$$

$\varepsilon$ is a constant, known as the "inhibition constant". The number of disparity layers $d$ we shall denote by $D$, and we shall let $M$ be the diameter of the excitatory neighborhood $S(x, y, d)$. In the example shown in Figure 2, $M=5$, and the total number of cells in an excitatory neighborhood is 13. The number less the cell itself is 12, which we shall denote by $E$. The number of cells in an inhibitory neighborhood of a given cell is $2D-2$, excluding the cell itself.

### 2.3. *Parameter Values and Some Facts*

The parameter values chosen for our original algorithm[1] (Marr and Poggio, 1976) were $E=12$, $D=7$, $\varepsilon=2$, $\theta=4$, with the excitatory neighborhood shown in Figure 2. Among other constraints, these parameter values were chosen to satisfy the following conditions:

*2.3.1.* in the absence of inhibition and of a contribution from the term $C^0$, straight line borders should fill in as shown in Figure 3a. This is true when $\theta \leqq 4$.

*2.3.2.* Straight line borders between two "filled-in" planes at different disparities should not grow. This requires that $4-2\varepsilon<\theta$.

*2.3.3.* With the particular values chosen: — A pattern of 5 connected points is the smallest configuration that can survive (see Fig. 3b). It will not grow unless one other point is added (e.g. at $P$ in Fig. 3b).
— The sharpest convexity capable of surviving against one inhibition, with the help of a contribution from $C^0$ is a right-angle. Figure 3c shows that the condition is $6-\varepsilon \geqq \theta$.
— A convex or flat border cannot grow against one inhibition; it can grow only into scattered active cells.
— The least concave patterns capable of growing under two inhibitions are shown in Figures 3d and e. They fill in by one or two cells and then are no longer concave enough to grow under two inhibitions.

## 3. Invariant States and the Matching Rules

The matching rules for stereopsis that were given in the introduction take the following form for the algorithm discussed here:

*(1) Uniqueness.* Each item from each image may be assigned at most one disparity value.

[1] In Marr and Poggio (1976), the value of $\theta$ was given as 3.0, whereas here it is 4.0. The reason for the discrepancy is that the algorithm used to produce the stereograms for that article essentially used the condition $>\theta$, whereas here, we use the condition $\geqq\theta$

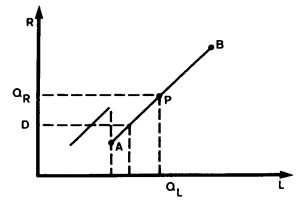



**Fig. 4.** The solid lines indicate solution planes (cf. Fig. 1a). Lines of sight $PQ_R$, $PQ_L$ intersect solution planes at only one point $P$, except possibly near the (rare) disparity boundaries like A. Thus configurations that obey rule *R1* are invariant

*(2) Continuity.* Disparity does not change almost everywhere.

*Comment.* R2 has now taken a slightly different form. This is because disparity takes only discrete value in this algorithm. Images containing smoothly varying disparities may be handled by a modified version of the algorithm, which will be discussed in Section 5.

We now show that the states in which these two rules are obeyed are for all practical purposes *invariant*, i.e. they are fixed points of (1), and once achieved, do not change in subsequent iterations.

### 3.1. *Configurations that Satisfy the Matching Rules are Invariant*

The continuity and uniqueness conditions mean that, for each value of $y$, a cross-section of the network has the appearance shown in Figure 4 (the continuity condition also requires that the active segment has some extension in the $y$ direction). That is, the "on" cells in the network form extended segments like that shown as $AB$ (continuity), and most lines of sight (e.g. $PQ_L$, $PQ_R$) intersect only one of these extended segments (uniqueness). Some lines of sight (e.g. to $D$) may intersect two planes: this occurs only at the (rare) boundaries at which disparity changes. The physical situation is that one surface is obscuring the other.

We show now that these configurations are invariant if the parameter values are appropriate.

(i) Interior points like $P$ are certainly invariant if

$$\sum_{x',y',d' \in S(x,y,d)} C^t_{x',y';d} \geqq \theta \qquad (4)$$

if $P$ is interior in both $x$ and $y$.

(ii) Equation (4) implies that boundary points like $A$ (Fig. 4) on a straight boundary (in the $x-y$ plane) will not grow into the interior of an existing segment at another disparity provided that

$$E/2+1-2\varepsilon<\theta. \qquad (5)$$

Concave pieces of boundaries can in principle grow, but not much for two reasons. Firstly, boundaries

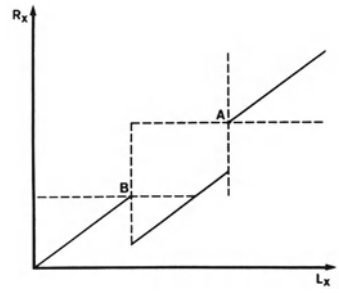
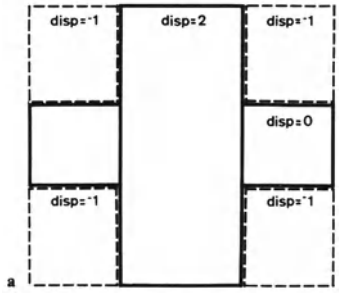
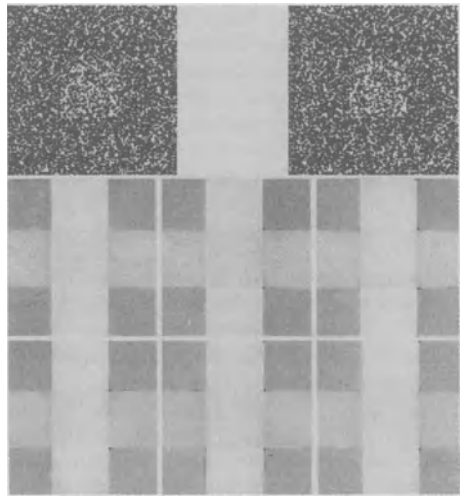



**Fig. 5.** The two possible stable edges for flat boundaries. Depending on the initial conditions, edges can occur that are defined by the line where cells begin receiving one (A) or two (B) inhibitions from the other surface

cannot be everywhere concave, and secondly, with our particular excitatory neighborhood and parameter values (see Figs. 3d and e) the amount a concave border can fill in is limited to at most two elements. Figure 5 shows the two possible stable edges for flat boundaries.




a



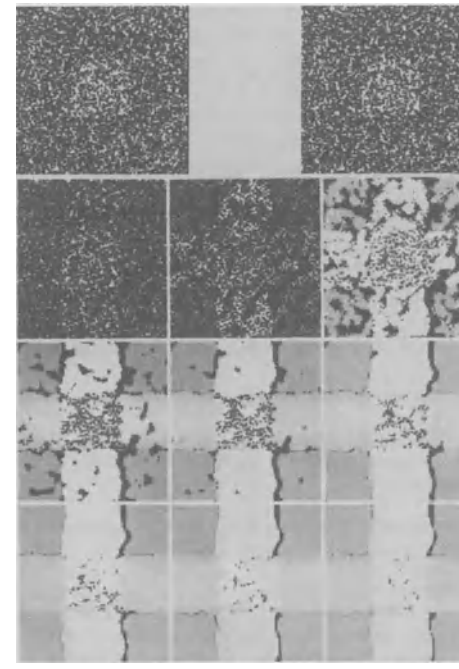
b

### *3.2. Not All Invariant Configurations Satisfy the Matching Rules*

Strictly speaking, the converse result to that of the last section is not true. A counter-example to the uniqueness condition that is stable with our parameters appears in Figure 6. Interior points of a plane, wholly surrounded by other points in the same sheet, can survive inhibition from two other cells and so can boundary points where the boundary is straight. In Figure 6b, points of these two types are the only ones that occur. A counter-example to the continuity condition appears in Figure 7, and it is left as an exercise to show that this pattern is invariant. In practice, neither of these configurations can actually develop from a random-dot stereogram.

When the input consists of two stereograms portraying a single surface, the probabilistic analysis of the next section shows that with high probability, the solutions will in fact obey the uniqueness condition.

**Fig. 6a—c.** A stable geometrical configuration that violates the uniqueness condition **a**. The central square consists of two planes, one at disparity 2 and one at disparity 0. This configuration is a stable state of the algorithm, in the sense that if it is loaded directly into the network, an invariant configuration is quickly reached in which both planes are represented; **b** demonstrates this. The stereogram is marked Left and Right, and 5 iterations of the algorithm are shown. If the network is loaded in the usual way, however, the algorithm develops a solution that is a mosaic of patches from the two levels **c**



c

If the input stereograms portray a transparent surface in front of another surface, the algorithm with our parameter values will usually fail to represent the input accurately, tending instead to develop a solution that obeys the two conditions and consists of a mosaic of patches from the two levels (Fig. 6c). With the parameters we chose, there seems to be no convenient and precise definition of the stability of configurations that forces the uniqueness and continuity of solutions. For instance, even if one requires in addition to invariance some kind of *spatial stability*[2], the counter-example of Figure 6 cannot be avoided, although a reasonable "spatial stability" condition would exclude the counter-example of Figure 7.

If one could exclude significant overlaps between surfaces lying at different disparities, it appears that one can derive the continuity conditions for invariant configurations. The argument is based here on the notion of a *hole*[3], and shows by straightforward geometry that holes are not invariant.

In one dimension (in which the network consists *only* of the part shown in Fig. 1a) the problem of this section becomes easier. Apparently, the only way of reducing the 2-dimensional problem to a satisfactory state is by changing the parameter values (see Section 6).

## 4. Probabilistic Analysis of the Algorithm

We have been unable to obtain general results about the convergence of this type of algorithm. Standard approaches — e.g. Liapunov-type methods and the usual fixed point theorems — apparently fail in this situation for reasons that we shall mention in the discussion.

The probabilistic analysis given here, although not completely satisfactory, nevertheless provides useful information about the algorithm's convergence for random-dot stereograms. Strictly speaking its application is restricted to inputs with a random structure.

The idea behind our analysis is that the cells in the network can be divided into populations on which the excitatory and inhibitory inferences are statistically homogeneous (cf. Marr, 1971). Our analysis is very specific to the algorithm of (1) because the way in which the cells are divided into populations depends critically on the geometry of the algorithm and on our a priori knowledge of its invariant state.

[2] A configuration is "spatially stable" if it is in some sense invariant under small perturbations (for instance each active point can be required to belong to a $3 \times 3$ neighborhood of points with the same disparity)

[3] There is a hole in the network for a given $y$ if there exist two intersecting lines of sight neither of which contains an "on" cell

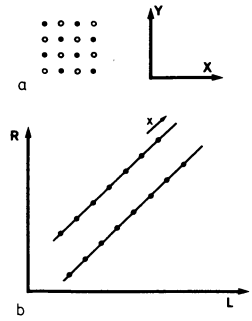



**Fig. 7a and b.** A stable geometrical configuration that violates the continuity condition. At each of two disparity values, the "on" cells form a checkerboard pattern, but they are arranged in such a way that neither level can fill in, because of inhibition from the other

### *4.1. Assumptions and Notation*

The algorithm has the structure shown in Figure 1 and the network is loaded from the input as specified by (2). We shall assume that the inputs have the following properties.

*4.1.1.* The 1's in each image occur randomly with probability $v$, and the autocorrelation of each input sequence (for any given $y$) is a Kronecker $\delta$.

*4.1.2.* The input admits a unique solution surface that is large enough to neglect boundary effects.

Condition 4.1.2 means that the left input is equal to the right one, modulo $x$-translation. Condition 4.1.1 implies that in the initial state of the network $C$, the density of 1's on the solution layer equals $v$, and elsewhere it is $v^2$. We subdivide the cells into five populations, by classifying them in two ways:

(i) according to whether or not they are a "on" in the initial state $C^0$, and

(ii) according to the number of active inputs from the images.

We draw both the populations 0 and 1 from cells that lie on the solution layer; population 0 is defined to receive no active inputs from the image, and population 1 receives two. Notice that there are no cells in the solution layer that receive exactly one active input.

The other three populations that we define refer to cells that lie off the solution layer; population 11 receives two active inputs from the image, population 10 receives one, population 00 receives none. The five populations {0, 1, 11, 10, 00} are exclusive and exhaustive.

We denote by $p_0(t)$, $p_1(t)$, etc. the probability that a cell in the respective population is "on" at time $t$. The goal of our analysis is to express the values of the $p_\pi(t)$ in terms of $p_\pi(t-1)$ for the various populations $\pi$. This allows us to examine the convergence numerically, and

**Table 1a—d.** The behavior of the algorithm compared with the probabilistic theory of the algorithm, for the stereograms having four different densities that are exhibited in Figure 8

**a.** $v=0.5$, $E=12$, $D=7$, $\varepsilon=2$, $\theta=4.0$ (SQ 50%)

| Iteration | | $p_r$ | $p_w$ | $p_0$ | $p_1$ | $p_{00}$ | $p_{01}$ | $p_{11}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | Algorithm | 0.46 | 0.07 | 0.93 | 0.01 | 0.29 | 0 | 0 |
| | Theory | 0.47 | 0.087 | 0.92 | 0.01 | 0.35 | 0 | 0 |
| 2 | Algorithm | 0.70 | 0.04 | 0.43 | 0.97 | 0 | 0 | 0.16 |
| | Theory | 0.62 | 0.02 | 0.26 | 0.97 | 0.04 | 0 | 0.08 |
| 3 | Algorithm | 0.90 | 0.01 | 0.99 | 0.81 | 0.04 | 0.001 | 0.001 |
| | Theory | 0.97 | 0 | 0.99 | 0.96 | 0 | 0 | 0 |
| 4 | Algorithm | 0.98 | 0.002 | 0.96 | 1.0 | 0 | 0 | 0.01 |
| | Theory | 1.0 | 0 | 1.0 | 1.0 | 0 | 0 | 0 |
| 5 | Algorithm | 0.99 | 0 | 1.0 | 0.99 | 0 | 0 | 0 |
| | Theory | 1.0 | 0 | 1.0 | 1.0 | 0 | 0 | 0 |

**b.** $v=0.25$, $E=12$, $D=7$, $\varepsilon=2$, $\theta=4.0$ (SQ 25%)

| Iteration | | $p_r$ | $p_w$ | $p_0$ | $p_1$ | $p_{00}$ | $p_{01}$ | $p_{11}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | Algorithm | 0.24 | 0.001 | 0.31 | 0.02 | 0.002 | 0 | 0 |
| | Theory | 0.27 | 0.003 | 0.35 | 0.04 | 0.005 | 0 | 0 |
| 2 | Algorithm | 0.39 | 0 | 0.26 | 0.81 | 0 | 0 | 0 |
| | Theory | 0.475 | 0 | 0.41 | 0.68 | 0 | 0 | 0 |
| 3 | Algorithm | 0.58 | 0 | 0.52 | 0.78 | 0 | 0 | 0 |
| | Theory | 0.92 | 0 | 0.90 | 0.97 | 0 | 0 | 0 |
| 4 | Algorithm | 0.74 | 0 | 0.70 | 0.87 | 0 | 0 | 0 |
| | Theory | 1.0 | 0 | 1.0 | 1.0 | 0 | 0 | 0 |

**c.** $v=0.1$, $E=12$, $D=7$, $\varepsilon=2$, $\theta=3.0$ (SQ 10%)

| Iteration | | $p_r$ | $p_w$ | $p_0$ | $p_1$ | $p_{00}$ | $p_{01}$ | $p_{11}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | Algorithm | 0.10 | 0 | 0.11 | 0.05 | 0 | 0 | 0 |
| | Theory | 0.11 | 0 | 0.11 | 0.106 | 0 | 0 | 0 |
| 2 | Algorithm | 0.20 | 0 | 0.15 | 0.72 | 0 | 0 | 0 |
| | Theory | 0.17 | 0 | 0.14 | 0.39 | 0 | 0 | 0 |
| 3 | Algorithm | 0.36 | 0 | 0.32 | 0.68 | 0 | 0 | 0 |
| | Theory | 0.35 | 0 | 0.32 | 0.61 | 0 | 0 | 0 |
| 4 | Algorithm | 0.53 | 0 | 0.51 | 0.77 | 0 | 0 | 0 |
| | Theory | 0.86 | 0 | 0.85 | 0.96 | 0 | 0 | 0 |
| 5 | Algorithm | 0.71 | 0 | 0.69 | 0.83 | 0 | 0 | 0 |
| | Theory | 1.0 | 0 | 1.0 | 1.0 | 0 | 0 | 0 |

**d.** $v=0.05$, $E=12$, $D=7$, $\varepsilon=2$, $\theta=2.0$ (SQ 05%)

| Iteration | | $p_r$ | $p_w$ | $p_0$ | $p_1$ | $p_{00}$ | $p_{01}$ | $p_{11}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | Algorithm | 0.13 | 0 | 0.12 | 0.22 | 0 | 0 | 0 |
| | Theory | 0.13 | 0 | 0.12 | 0.26 | 0 | 0 | 0 |
| 2 | Algorithm | 0.38 | 0 | 0.35 | 0.88 | 0 | 0 | 0 |
| | Theory | 0.48 | 0 | 0.46 | 0.81 | 0 | 0 | 0 |
| 3 | Algorithm | 0.67 | 0 | 0.65 | 0.88 | 0 | 0 | 0 |
| | Theory | 1.0 | 0 | 1.0 | 1.0 | 0 | 0 | 0 |
| 4 | Algorithm | 0.90 | 0 | 0.89 | 0.96 | 0 | 0 | 0 |
| | Theory | 1.0 | 0 | 1.0 | 1.0 | 0 | 0 | 0 |
| 5 | Algorithm | 0.98 | 0 | 0.98 | 0.98 | 0 | 0 | 0 |
| | Theory | 1.0 | 0 | 1.0 | 1.0 | 0 | 0 | 0 |

we say that a solution is achieved at time $T$ when

$$p_0(t)=p_1(t)=1, \quad \text{and}$$

$$p_{00}(t)=p_{10}(t)=p_{11}(t)=0, \quad \text{for every } t \geqq T.$$

The critical assumption here is that the quantity $p_\pi(t)$ completely describes the structure of active cells in the respective population $\pi$. This assumption is true for the initial iteration and only approximate thereafter. We shall discuss this point at the end of the section.

### 4.2. *Formulae*

The state of a cell $(x, y, d)$ at time $(t+1)$ depends upon the number of active cells in its excitatory $S(x, y, d)$ and inhibitory $0(x, y, d)$ neighborhoods at time $t$.

If we denote the populations to which the cell belongs by $\pi$, ($\pi$ running through the five populations 0, 1, 00, 11, 01), let us define:

$e_\pi(r)$ to be the probability that exactly $r$ cells are "on" in the excitatory neighborhood $S(x, y, d)$ at time $t$ and

$i_\pi(r)$ to be the probability that exactly $r$ cells are "on" in the inhibitory neighborhood $0(x, y, d)$ at time $t$.

It is convenient to introduce some further quantities:

$q_s(t)$ is the probability that a given cell on the "solution" plane is active at time $t$.

$q_w(t)$ is the probability that a given cell elsewhere in the network is active.

$q_-(t)$ is the probability that a given cell is active in the inhibitory neighborhood of a cell in the population 0.

$q_+(t)$ is the probability that a given cell is active in the inhibitory neighborhood of a cell in the population 1.

Then

$$\begin{aligned} q_s(t) &= p_0(t)\cdot(1-v)+p_1(t)\cdot v \\ q_w(t) &= p_{00}(t)\cdot(1-v)^2+p_{01}(t)\cdot 2v(1-v) \\ &\quad +p_{11}(t)\cdot v^2 \\ q_-(t) &= p_{00}(t)\cdot(1-v)+p_{01}(t)\cdot v \\ q_+(t) &= p_{11}(t)\cdot v+p_{10}(t)\cdot(1-v). \end{aligned} \tag{6}$$

Writing $B(n, f; m)={}_mC_n\cdot f^n(1-f)^{m-n}$. where ${}_mC_n$ is the binomial coeficient, we have immediately

$$\begin{aligned} e_1(r) &= e_0(r)=B(r, q_1(t); E) \\ i_1(r) &= B(r, q_+(t); 20-2) \\ i_0(r) &= B(r, q_-(t); 20-2) \\ e_{11}(r) &= e_{00}(r)=e_{10}(r)=B(r, q_w(t); E). \end{aligned} \tag{7}$$

The remaining $i_\pi$ are more difficult to obtain, since the inhibitory contributions to cells lying off the solution plane come from cells lying on the solution plane and from cells lying off the solution plane, and these two

populations obey different statistics. In fact

$$i_{11}(r)=[p_1(t)]^2\cdot B(r-2,q_+(t);2D-4)$$
$$+2p_1(t)\cdot(1-p_1(t))\cdot B(r-1,q_+(t);2D-4)$$
$$+[1-p_1(t)]^2\cdot B(r,q_+(t);2D-4) \quad (8)$$

$$i_{00}(r)=[p_0(t)]^2\cdot B(r-2,q_-(t);2D-4)$$
$$+2p_0(t)\cdot(1-p_0(t))\cdot B(r-1,q_-(t);2D-4)$$
$$+[1-p_0(t)]^2\cdot B(r,q_-(t);2D-4).$$

The final case $i_{10}$ is especially awkward, because along one of the inhibitory lines the probability of a cell being "on" is $q_+$ and along the other diagonal it is $q_-$.

$$i_{10}(r)=\Sigma\{p_1(t)B(k-1,q_+(t);D-2)$$
$$+(1-p_1(t))B(k,q_+(t);D-2)\}$$
$$\cdot\{p_0(t)\cdot B(r-k+1,q_-(t);D-2)$$
$$+(1-p_0(t))B(r-k,q_-(t);D-2)\}. \quad (9)$$

We now need to relate the $p_\pi(t+1)$ and the $p_\pi(t)$ in terms of the $e_\pi$ and $i_\pi$. For each cell population we know the distributions of incoming excitation and inhibition, and we know that a cell will be on whenever the excitations exceed the inhibitions by at least $\theta$. Hence:

$$p_\pi^{t+1}=\sum_{\substack{n=\theta \text{ to } E\\ M=\theta \text{ to } 2D-2\\ n-\varepsilon m\geq\theta_\pi}} e_\pi^t(n)\cdot i_\pi^t(m) \quad (10)$$

where

$\theta_\pi=\theta-1$ for $\pi=1,\ \pi=11$

$\theta_\pi=\theta$ otherwise.

If the input term $C^\theta_{x,y;d}$ of (1) is neglected, $\theta_\pi=\theta$ for all $\pi$.

The Equations (10) are too complex to be solved analytically. Numerical solutions were however obtained for various values of the parameters and some of the results are given in Table 1 and Figure 8.

### *4.3. Range of Parameter Values and Comparison with Actual Runs*

Figure 8 exhibits the performance of the algorithm for stereograms having densities of from 0.5 to 0.05. Table 1 gives the statistics that were measured from these runs, and also the parameters predicted by the probabilistic theory. The values obtained from the theory match those from the algorithm quite well for the first iteration, but except for the case $\nu=0.05$, they diverge quite rapidly thereafter, and even this case diverges by the third iteration.

We have already noted the main reason for the discrepancy. The assumption that the statistical structure of various populations is purely random (inside each population and between populations) holds exactly for the first iteration but only approximately thereafter, because the operator of Figure 1c has a local structure which can preserve local clusters of active cells. There are two ways in which this affects our probabilistic description for the second and subsequent iterations. The first is that clusters are more stable than the assumption of randomness would predict. Thus clusters forming on the solution layer will in certain circumstances change the rate of convergence predicted by the randomness assumption.

The difficulties arise where clusters form off the solution layer. These will again tend to be more stable than our analysis assumes, but their effect acts against convergence. However, we shall argue that the probability of large "wrong" clusters is small for most patterns. In fact, the typical value of the probability that a wrong cell is "on" after the first iteration lies around 0.1. The probability (after the first iteration) of a self-supporting $3\times3$ cluster at a given position in a wrong layer (asuming that the cluster was absent in the initial state and accepting the oversimplified assumption of randomness after the first iteration) is about $10^{-9}$, and hence less than $10^{-4}$ that one exists off the solution plane somewhere in the network.

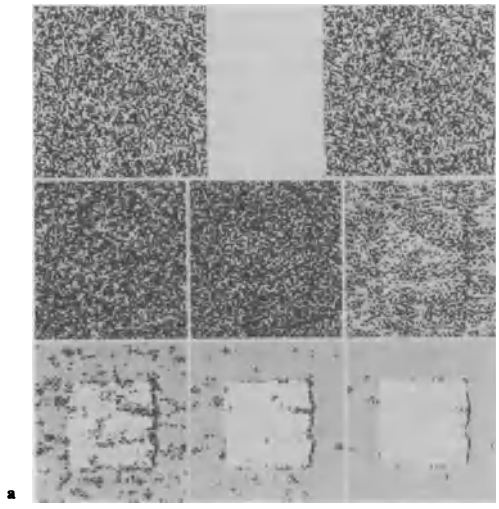
A cluster of this size may survive permanently, because every element in it has at least 6 cells in its excitatory neighborhood, and this is enough to resist 1 inhibition. The probability of this or something larger arising by chance is so small that if it occurs it is likely to be a consequence of the particular image. In fact, some small "wrong" patches do sometimes occur (inspect Marr and Poggio, 1976, Fig. 5d) but such instances can usually be traced to an accidental correlation in the image. In this sense, extended patches are "correct" solution regions.

The second effect that leads to discrepancies between the theory and the behavior of the algorithm is also a side-effect of clustering, since as well as being stable, the clusters tend to concentrate "on" cells more than the randomness assumption would predict. For example, at iteration 2 of the case $\nu=0.25$ (Fig. 8b), although the overall density of ones on the solution plane is about 0.39, it is far from true that each cell can expect to find $0.39E$ "on" cells in its excitatory neighborhood. Cells in the filled in regions have almost all their neighbors on, whereas those in the interstices have none. Convergence is achieved by a growth outwards that fills in the blank regions, but although it is steady, it is necessarily slower than the theory predicts.

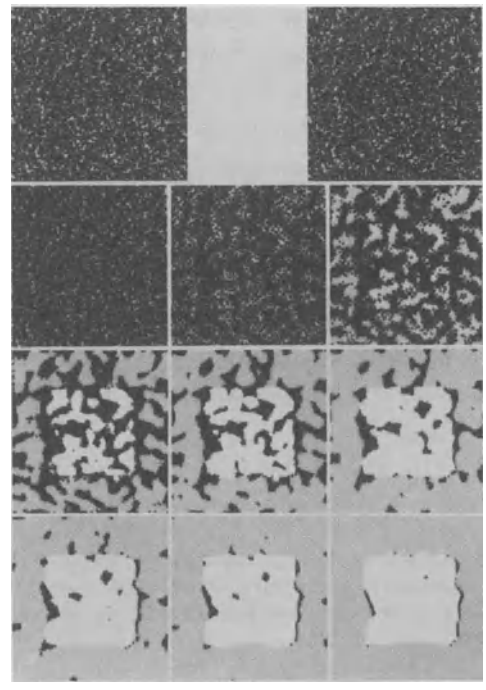
## 5. Observations

### *5.1.*

There is a wide latitude in the range of parameters for which the network converges. Table 2 shows firstly the

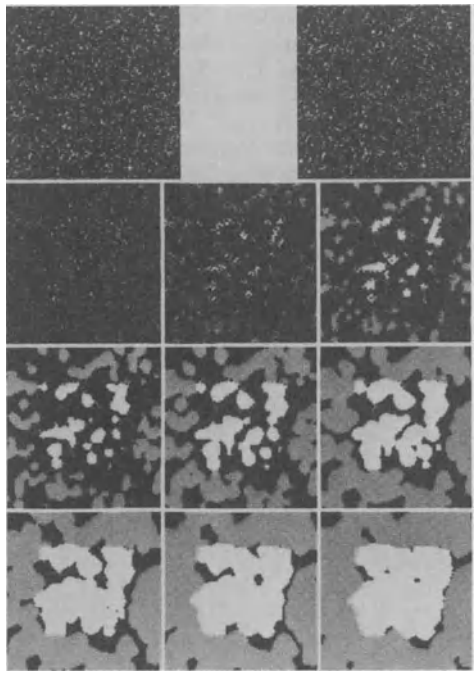


a

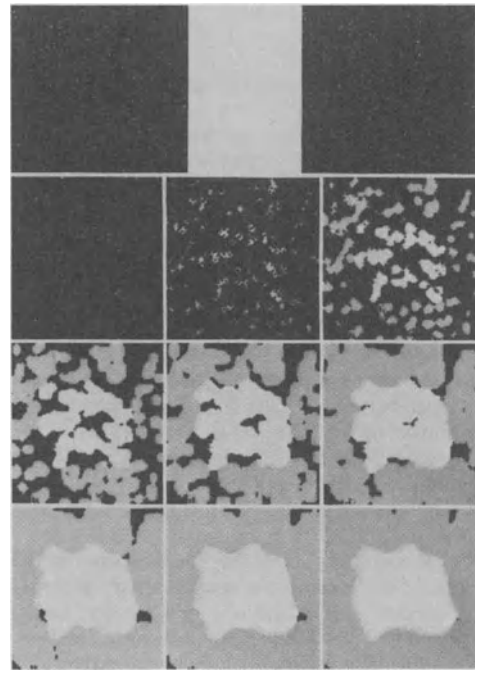


b

**Fig. 8a—d.** The stereograms (Left, Right) and iterations tabulated in Table 1. Stereogram densities are 50% **a**, 25% **b**, 10% **c** and 5% **d**. Parameters are as shown in Table 1



c



d

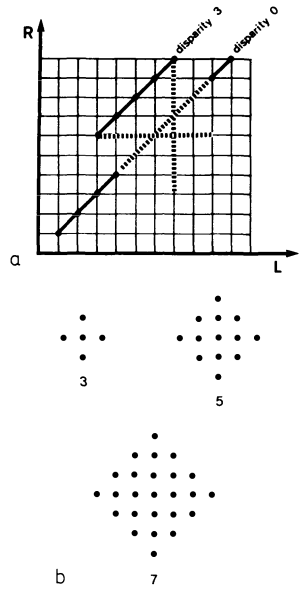



**Fig. 9a and b.** The minimum resolvable area of a small pattern against a background increases with disparity. To prevent the background from filling in completely, the length of the patch in the $x$-direction must be at least $|d|+2$ **a. b** Shows the circles of diameters 3, 5 and 7 used in Figure 6 of Marr and Poggio, 1976

wide range in stereogram density $v$ that is tolerated by our parameters (with fixed $\theta$), and secondly, for a fixed value of $v$ ($v=0.5$) gives some idea of the range of the other parameter values for which the network will converge. Note that in the implementation described by Marr and Poggio (1976), the threshold was not fixed, but was determined by the density of "on" cells in the network. This allowed solution to the matching problem over a very wide range of dot densities.

*5.2.*

Let us define the probability that a cell on the solution layer is "on" at time $t$ to be

$$p_r(t)=v\cdot p_1(t)+(1-v)p_0(t)$$

and the probability that a cell off the solution layer is on at time $t$ as

$$p_w(t)=v^2\cdot p_{11}(t)+2v(1-v)p_{10}(t)+(1-v^2)p_{00}\,.$$

In a successful run, $p_r$ converges to 1 and $p_w$ to 0. With our particular parameters, convergence is monotonic if it occurs. This is not true, however, for the individual quantities $p_1$, $p_0$, $p_{11}$, $p_{10}$, $p_{00}$, neither is it true of $p_r$ and $p_w$ for all values of the parameters (see Table 2).

*5.3.*

We have already seen that the sharpest local corner capable of resisting 1 inhibitory input is about 90° or more, hence thin, sharp regions will tend to be rounded off locally (see Marr and Poggio, 1976, Fig. 5c). The exact shape of the input pattern is preserved only up to this limit.

**Table 2a—c.** The algorithm of (1) converges for a wide range of control parameters. Tables 2a,b show convergence for $v=0.5$ and $v=0.1$ with the same parameters. Table 2c shows convergence for an entirely different set of parameters

**a.** $v=0.5$, $E=12$, $D=7$, $\varepsilon=2$, $\theta=3.0$

| Iteration | $p_r$ | $p_w$ | $p_0$ | $p_1$ | $p_{00}$ | $p_{10}$ | $p_{11}$ |
|---|---|---|---|---|---|---|---|
| 1 | 0.50 | 0.15 | 0.98 | 0.026 | 0.61 | 0 | 0 |
| 2 | 0.57 | 0.13 | 0.15 | 0.997 | 0 | 0 | 0.54 |
| 3 | 0.69 | 0.039 | 0.995 | 0.39 | 0.16 | 0 | 0 |
| 4 | 0.97 | 0.007 | 0.935 | 1.0 | 0 | 0 | 0.029 |
| 5 | 1.0 | 0 | 1.0 | 1.0 | 0 | 0 | 0 |

**b.** $v=0.1$, $E=12$, $D=7$, $\varepsilon=2$, $\theta=3.0$

| Iteration | $p_r$ | $p_w$ | $p_0$ | $p_1$ | $p_{00}$ | $p_{10}$ | $p_{11}$ |
|---|---|---|---|---|---|---|---|
| 1 | 0.11 | 0 | 0.11 | 0.106 | 0 | 0 | 0 |
| 2 | 0.17 | 0 | 0.14 | 0.39 | 0 | 0 | 0 |
| 3 | 0.35 | 0 | 0.32 | 0.62 | 0 | 0 | 0 |
| 4 | 0.86 | 0 | 0.85 | 0.96 | 0 | 0 | 0 |
| 5 | 1.0 | 0 | 1.0 | 1.0 | 0 | 0 | 0 |

**c.** $v=0.5$, $E=2$, $D=7$, $\varepsilon=0.5$, $\theta=1.0$

| Iteration | $p_r$ | $p_w$ | $p_0$ | $p_1$ | $p_{00}$ | $p_{10}$ | $p_{11}$ |
|---|---|---|---|---|---|---|---|
| 1 | 0.40 | 0.11 | 0.75 | 0.058 | 0.43 | 0 | 0.010 |
| 2 | 0.55 | 0.23 | 0.11 | 0.99 | 0.004 | 0 | 0.90 |
| 3 | 0.45 | 0.083 | 0.78 | 0.11 | 0.32 | 0 | 0.006 |
| 4 | 0.59 | 0.20 | 0.20 | 0.99 | 0.003 | 0 | 0.80 |
| 5 | 0.51 | 0.063 | 0.82 | 0.20 | 0.24 | 0 | 0.009 |
| 6 | 0.65 | 0.17 | 0.32 | 0.99 | 0.003 | 0 | 0.66 |
| 7 | 0.62 | 0.042 | 0.87 | 0.36 | 0.15 | 0.001 | 0.014 |
| 8 | 0.76 | 0.11 | 0.54 | 0.97 | 0.002 | 0 | 0.43 |
| 9 | 0.82 | 0.021 | 0.94 | 0.71 | 0.053 | 0.002 | 0.026 |
| 10 | 0.94 | 0.027 | 0.88 | 0.995 | 0 | 0 | 0.11 |
| 11 | 0.995 | 0.009 | 0.996 | 0.995 | 0.001 | 0 | 0.031 |
| 12 | 1.0 | 0.004 | 1.0 | 1.0 | 0 | 0 | 0.014 |
| 13 | 1.0 | 0.002 | 1.0 | 1.0 | 0 | 0 | 0.007 |
| 14 | 1.0 | 0 | 1.0 | 1.0 | 0 | 0 | 0.003 |

*5.4. Minimum Size vs. Disparity*

A natural consequence of the structure of the algorithm is that the minimum resolvable area of a small pattern against a background increases with disparity (see Marr and Poggio, 1976, Fig. 6). We give an estimate of the dependence of minimum patch size on disparity difference. Consider a section for some fixed $y$ of the network (Fig. 9). Assume that the patch and background regions are filled in. The condition for growth at a point $(x, y, d)$ under 1 inhibition is that the number of "on" cells in an excitatory neighborhood should be not less than $\theta+\varepsilon-C^0=5$ or 6, depending

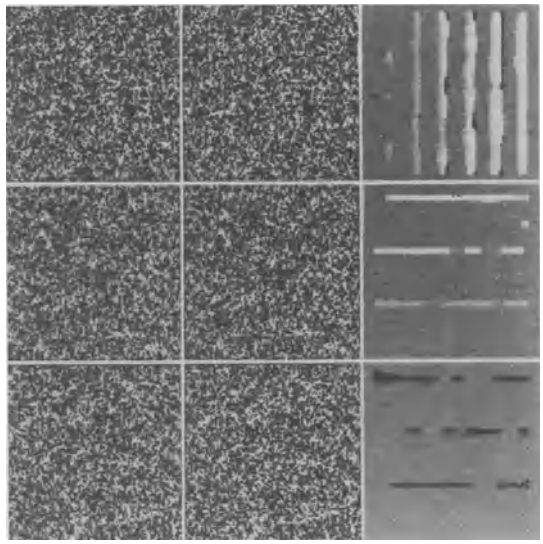


**Fig. 10a—c.** Thin vertical and horizontal stripes of various disparities. The left and right stereograms are shown with the stable network solution to them. The stereograms are 100 by 100, and consist of stripes with the following coordinates ($x$ or $y$), thicknesses $s$ and disparities $d$:

| a | | | b | | | c | | |
|---|---|---|---|---|---|---|---|---|
| $x$ | $s$ | $d$ | $y$ | $s$ | $d$ | $y$ | $s$ | $d$ |
| 15 | 2 | −1 | 15 | 2 | −1 | 15 | 2 | +1 |
| 30 | 3 | −1 | 30 | 3 | −1 | 30 | 3 | +1 |
| 45 | 3 | −2 | 45 | 2 | −2 | 45 | 2 | +2 |
| 60 | 4 | −2 | 60 | 3 | −2 | 60 | 3 | +2 |
| 75 | 4 | −3 | 75 | 2 | −3 | 75 | 2 | +3 |
| 90 | 5 | −3 | 90 | 3 | −3 | 90 | 3 | +3 |

on the initial conditions. From Figure 3 we see that flat or convex regions will not grow whereas concave regions will. Hence our small patch will not tend to grow, whereas the background will spread until stopped by two inhibitions. We see from Figure 9a that to prevent the background from filling in completely (which would subsequently destroy the patch because convex borders cannot survive two inhibitions), the length of the patch in the $x$ direction must be at least $|d|+2$. This condition must hold for at least three adjacent lines aligned in the $y$ direction. Figure 6 of Marr and Poggio (1976) illustrates the approximate validity of this relation. Figure 9b shows the sizes of circles of diameter 3, 5, and 7 used in the input for that figure. These precise patterns do not necessarily emerge in the appropriate layer of the network because of the random nature of the borders. The circle of diameter 3 contains no $3 \times 3$ subset and therefore does not survive at any disparities. The circle of diameter 5 contains one $3 \times 3$ square and survives as expected at disparity 1; it also survives, apparently accidentally, at disparity 2, but not at disparity 3. The circle of diameter 7 contains one $5 \times 5$ square and thus survives at disparity 3.

A trivial consequence of this analysis is that horizontal stripes (parallel to the $x$ axis) are in general more stable than vertical ones (parallel to the $y$ axis). The minimum thickness for horizontal stripes is about 3 and is independent of disparity whereas the minimum thickness for vertical stripes is about $|d|+2$ (see Fig. 10).

### 5.5. *Uncorrelated Areas*

If there exists a sufficiently large area in the input where there is no correlation between the two images, the network will detect it (see Figs. 5 and 6 of Science). After the first iteration (with our parameter values and $v=0.5$) only a few cells remain "on" in the uncorrelated region, but provided the region is sufficiently large they will receive no inhibition from the surrounding more organized layers. Hence those cells that are on may act as germs for small regions that have become stable by the time the surround encroaches upon them, e.g. Figure 5d of Marr and Poggio (1976). Relatively small ($\ll d$) uncorrelated areas probably have to develop stable platelets to survive (see Fig. 6d of Marr and Poggio, 1976), and large uncorrelated areas decompose into a random mosaic of stable platelets (see Fig. 11).

Uncorrelated areas can be recognized as such during the read-out from the network, when the 1's that appear in the solution found by the network are used to establish an explicit correspondence between the two images.

### 5.6. *Extension to Images in which Disparity Varies Continuously*

The algorithm of (1) with the loading rules of (2) can deal only with images having discrete disparity values. This disparity in natural images commonly varies continuously. There are two approaches to this problem. One is to incorporate the representation of continuous values directly into the algorithm, and the other is to use the same algorithm, but with special rules for loading it and for interpreting its final state.

The first approach would clearly lead to a considerably different algorithm, perhaps more along the lines of the networks studied by Wilson and Cowan (1973), (see also Wilson, 1977). Such an algorithm could not be treated within the framework of this article.

The second approach does not require any changes in the analysis of the algorithm itself. One could, for example, define the loading conditions as follows:

Let $\Delta$ be the disparity attached to a possible correspondence between items in the left and right images. For integral $d$,

*5.6.1.* If $d-\eta \leqq \Delta < d+\eta$, load the cell corresponding to disparity level $d$ in the network.

For surfaces whose disparity does not oscillate too much or too densely, the value $\eta=0.5$ will lead to satisfactory results. The final state of the network establishes a correspondence between items in the left and right images, but their associated disparity is read not from the network (i.e. $d$) but directly from the input (i.e. $\Delta$). Confusions may of course arise in the correspondence established by the network if the value of $d$ spans the disparity range too coarsely.

In order to deal with surfaces that are less well-behaved, one can incorporate some hysteresis into the loading rules. The loading process then consists of the following steps:

*5.6.2.* Load cells according to 5.6.1 with $\eta=0.3$ (say).

*5.6.3.* Moving across the image $(x, y)$ in a spatially ordered way, if a possible match $(x,y,\Delta)$ was not loaded by 5.6.2, adopt the following procedure:

Let $d^-=$ Integral part of $\Delta$, $d^+=1+d^-$. Examine $(x, y)$ neighborhoods of $(x, y, d^-)$ and of $(x, y, d^+)$ in the network as it is loaded so far. Assign the current match to that $d$ whose neighborhood contains more loaded cells, if one of them does. Else load this point according to 5.6.1 with $\eta=0.5$.

This process will load most images in satisfactory way, and the read-out procedure is similar to that of the previous case.

## 6. A Mathematically Tractable Version of the Algorithm

A suitable choice of the parameter values and of the loading rules of the algorithm allows a complete mathematical analysis of its asymptotic behavior. In this section we introduce this "strict" version of the algorithm and we characterize rigorously its properties. The actual performance of this version of the algorithm for various random dot stereograms will be then compared with the original algorithm.

### *6.1. Loading Conditions*

The initial state $C^0$ of the network is loaded from the stereograms $L$, $R$ in a way similar to the previous case but according to (11) (instead of (2)).

$$C^0_{x,y;d}=L_{x,y}\cdot R_{x+d,y} \tag{11}$$

where $1\cdot 1=0\cdot 0=1, \quad 1\cdot 0=0\cdot 1=0.$

This loading rule can be easily extended to cases in which more than two features are present. It is enough to define

$$f_i\cdot f_j=\delta_{ij}, \tag{12}$$

where $f_i$ and $f_j$, $(i\neq j)$ are two different features. The case when only two features are present clearly poses the hardest matching problem. We shall later compare this loading rule with the original (2) and discuss their relative merits for real images.

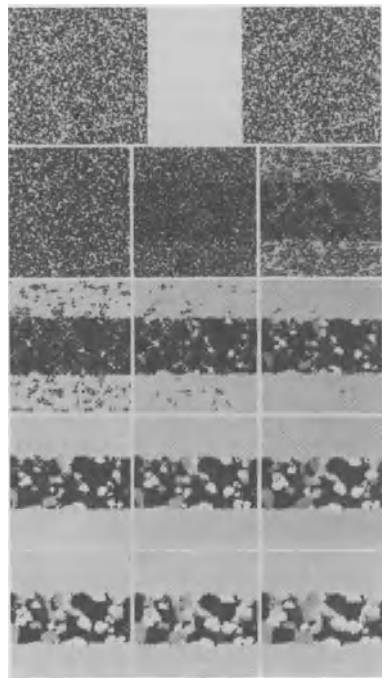


**Fig. 11.** The central band is uncorrelated. It decomposes into a random mosaic of patches, each of which is eventually stable

### *6.2.* The Algorithm

The relation between states at times $t$ and $t+1$ is given by (compare (1))

$$C^{t+1}_{x,y;d}=\sigma\left\{\inf\left[\sum_{x',y',d'\in S(x,y,d)} C^t_{x',y';d'}, H\right]\right.$$
$$\left.-\varepsilon\sum_{x',y',d'\in 0(x,y,d)} C^t_{x',y';d'}\right\} \tag{13}$$

where $H$ is a number that represents the "saturation" value for the excitation.

### *6.3. Choice of Parameter Values*

In this case the loading rules lead, for random dot stereograms with two features, to a density of 1 for the

**Table 3a and b.** The behavior of the mathematically tractable version of the algorithm, together with the probabilistic theory of the first iteration, for the two stereograms exhibited in Figures 14a and b

**a.** $v=0.5$, $E=13$, $D=7$

| $\varepsilon$ | $\theta$ | $H$ | Iteration | $p_r$ | $p_w$ | $p_0$ | $p_1$ | $p_{00}$ | $p_{10}$ | $p_{11}$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.2 | 10.75 | 13.0 | 1 | 0.9998 | 0.0017 | 0.9998 | 0.9998 | 0.0023 | 0.0011 | 0.0023 | Theory |
| | | | 1 | 0.99 | 0.0003 | 0.99 | 0.98 | 0.0008 | 0 | 0.0004 | Algorithm |
| 4.0 | 3.75 | 7.0 | 2 | 0.99 | 0 | 0.99 | 0.99 | 0 | 0 | 0 | |
| | | | 3 | 1.0 | 0 | 1.0 | 1.0 | 0 | 0 | 0 | |

**b.** $v=0.25$, $E=13$, $D=7$

| $\varepsilon$ | $\theta$ | $H$ | Iteration | $p_r$ | $p_w$ | $p_0$ | $p_1$ | $p_{00}$ | $p_{10}$ | $p_{11}$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.2 | 10.75 | 13.0 | 1 | 0.976 | 0.0078 | 0.968 | 1.0 | 0.0039 | 0.0015 | 0.082 | Theory |
| | | | 1 | 0.96 | 0.0002 | 0.95 | 1.0 | 0.002 | 0.0003 | 0.002 | Algorithm |
| 4.0 | 3.5 | 7.0 | 2 | 0.98 | 0 | 0.98 | 0.97 | 0 | 0 | 0 | |
| | | | 3 | 1.0 | 0 | 1.0 | 1.0 | 0 | 0 | 0 | |

"on" cells on the "correct" diagonal segments and, correspondingly, to a density of $v^2+(1-v)^2$ for the "on" cells on the "wrong" diagonal segments ($v$ is the density of 1's in the input images). When $v=0.5$, the density of the wrong cells is also 0.5; for smaller or larger $v$ the density is higher. The idea behind this approach is to choose parameter values for the first iteration that "kill" most of the "wrong" cells (and of course some of the "right" ones); from the second iteration on, the parameter values are such to ensure "filling-in" of the right diagonal segments, allowing, at the same time, a satisfactory mathematical analysis of the evolution of the network's state. This approach, which is carried out in the next two sections, leads to the following parameter values:

*6.3.1.* $S$ and 0 are as in Figure 2, $D=7$ and $M=5$ as before. Selfexcitation is now included but the $C^\theta$ term is omitted. We therefore write $E=13$ instead of 12.

*6.3.2. Iteration 1:*

$H=13$ (so that the inf operation can be neglected)
$\varepsilon=0.2$.
$\theta=10.75$.

*6.3.3. Second and Subsequent Iterations:*

$H=7$
$\varepsilon=4.0$
$\theta=3.5$.

*6.4. Probabilistic Analysis of the First Iteration*

We shall assume that the inputs have the properties 4.1.1 and 4.1.2. As in Section 4.1, we distinguish several populations of cells which are homogeneous with respect to the interaction structure: the populations are again denoted by 0, 1, 11, 10, 00 according to their respective inputs from the two images (see Section 2), and $p_0$, $p_1$, etc. denote the probability that a cell in the respective population is "on" after the first iteration. In this case the formulae for the solution layer are:

$$p_1 = \sum_{i=0}^{\frac{13-\theta}{\varepsilon}} {}_{12}C_i \cdot v^i (1-v)^{12-i}$$

$$p_0 = \sum_{i=0}^{\frac{13-\theta}{\varepsilon}} {}_{12}C_i \cdot (1-v)^i \cdot v^{12-i}.$$

For the "wrong" layers (writing $\mu=v^2+(1-v)^2$), the formulae are

$$p_{11} = \sum_{k=0}^{12} {}_{12}C_k \mu^k (1-\mu)^{12-k} \sum_{i=0}^{\frac{k+1-\theta}{\varepsilon}-2} {}_{10}C_k v^i (1-v)^{10-i}$$

$$p_{00} = \sum_{k=0}^{12} {}_{12}C_k \mu^k (1-\mu)^{12-k} \sum_{i=\theta}^{\frac{k+1-\theta}{\varepsilon}-2} {}_{10}C_k (1-v)^i v^{10-i}$$

$$p_{10} = \sum_{k=0}^{12} {}_{12}C_k \mu^k (1-\mu)^{12-k} \sum_{i=0}^{\inf\left\{\left(\frac{k-\theta}{\varepsilon}-2\right),\,5\right\}} \cdot {}_5C_i v^i (1-v)^{5-i} \sum_{j=0}^{\inf\left\{\left(\frac{k-\theta}{\varepsilon}-2\right)-i;\,5\right\}} {}_5C_j (1-v)^j v^{5-j}.$$

Therefore the probability that a cell in the solution layer is "on" after the first step is

$$p_r = p_1 \cdot v + p_0 \cdot (1-v)$$

and the probability that a cell off the solution layer is "on" after the first step is

$$p_w = v^2 p_{11} + 2v(1-v)p_{10} + (1-v)^2 p_{00}.$$

These equations can be used to find suitable parameter values. The parameters given in the previous section yield the values for $p_r$ and $p_w$ shown in Table 3.

### 6.5. Equivalent Rules

The parameter values from the second iteration on imply the following main "rules" for the algorithm:

*6.5.1.* One "on" cell in the inhibitory neighborhood always suffices to kill an "on" cell.

*6.5.2.* Without inhibition, at least three excitatory "on" cells are needed for "survival" of an "on" cell and four for its "birth".

### 6.6. Analysis of the Second Iteration

Table 3 gives the densities $p_r$ and $p_w$ after the first iteration. Only for the first iteration can a probabilistic analysis provide a reliable estimate of the density of "on" cells on the solution surface. As in our earlier analysis (Table 1), it becomes unreliable for the second iteration, because clusters of "on" cells can be expected to form off the solution layer (see Fig. 14 below). Rule 6.5.1 implies, however, that "wrong" clusters will disappear after the second iteration, unless they consist of at least four elements. Moreover, these elements must in practise be very close together for each to support the other three. In addition, according to rule 6.5.2, none of them can lie in the inhibitory neighborhood of other "on" cells (for instance on the solution layer where the density of on cells is relatively high, see Table 3). We argue that the probability of such situations is very small (actually much smaller than in the case considered in Section 4.4). If this occurs it can be attributed to an accidental correlation in the images. In this sense extended clusters are in fact "right" solution regions.

### 6.7. Asymptotic Analysis

The probabilistic analysis of the first iteration (Table 3) shows that one can assume that, from the second iteration onwards, there are no wrong "on" cells. It remains now to show that the density of "on" cells on the solution layer is high enough to allow asymptotic filling-in of the "right" surfaces. We prove the following:

*6.7.1. Filling-in Proposition.* Assume that (at some iteration $n$) there are no "on" cells off a given layer (diagonal), and that the density of "on" cells on this layer exceeds $0.4375 = 7/16$. Then, in the asymptotic configuration, there are no "off" cells on this layer.

*Proof.* Divide the solution plane into squares of 4 by 4 cells (we neglect boundaries). At least one of these squares must contain 8 "on" cells, for, otherwise, every square would contain at most 7 "on" cells yielding a density of at most 7/16, in contradiction with the hypothesis. This square will fill up with "on" cells.

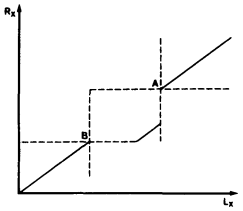


**Fig. 12.** With the modified parameters, cells cannot survive against one inhibition. Hence stable states satisfy the uniqueness condition, because no overlap is possible (compare Fig. 5)

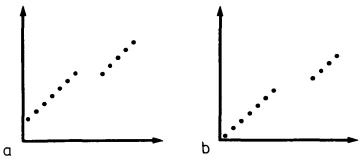



**Fig. 13a and b.** An oscillating solution with the modified parameters. The state **a** occurs at iterations $i, i+2, i+4, \ldots$, whereas state **b** occurs at iterations $i+1, i+3, i+5, \ldots$

(This can be seen by examining the various possible ways in which the 8 cells can be distributed, and we leave it as an exercise for the reader). Starting from this square, the whole plane will asymptotically be filled by "on" cells (since, by hypothesis, no inhibitory cells need be considered).

### 6.8. Invariant States and Matching Rules

The matching rules were defined in Section 3. States that satisfy the matching rules with the present parameter values are shown in Figure 12. In view of the rules 6.5.1 and 6.5.2, the following clearly hold:

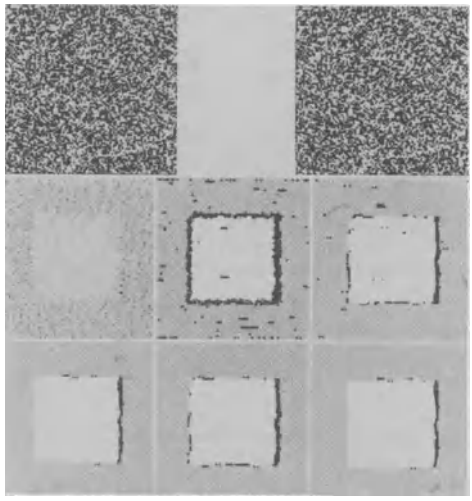
i) Configurations that satisfy the matching rules (Fig. 12) are invariant.

ii) Conversely, invariant configurations clearly have to obey the uniqueness condition (because of 6.5.1). The probabilistic analysis of the second step, together with the "filling-in" proposition 6.7.1, ensures in practise that there will be no holes[2] in the asymptotic invariant configurations.

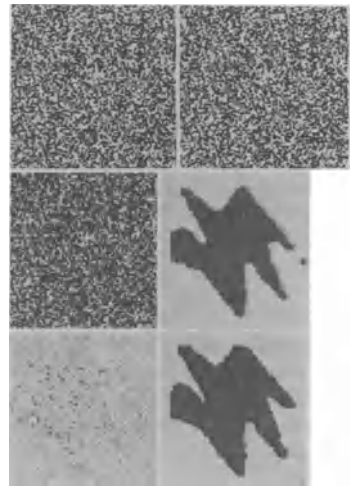
### 6.9. Asymptotic Liapunov Description

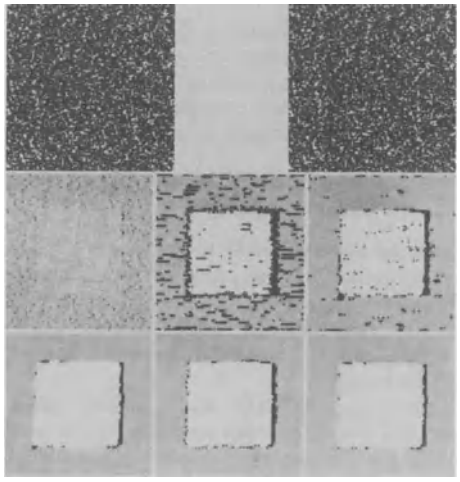
Besides the invariant asymptotic configuration, limit cycles of the type described in Figure 13 may also occur. Thus the previous description of asymptotic invariant states is not complete. We provide here an asymptotic analysis in terms of a Liapunov-like function which also encompasses such non-invariant states.

For a given state $C^i$, we define $F(C^i)$ to be the number of "on" cells having no "on" cells in their inhibitory neighborhood. We call an "on" cell that has less than three "on" cells in its excitatory neigh-



a





b

**Fig. 14a—c.** The behavior of the algorithm with modified parameters. The densities are 50% **a** and 25% **b**. The parameters are as stated in Section 6.3, and in Table 3. **c** Compares the two sets of parameters on a stereogram of a star that contains arms of various angles. The original parameters tend to give a more accurate final configuration

borhood a "solitary cell". Observe that solitary cells can never be "born" and that, after a finite number of iterations, all solitary cells will have disappeared.

*6.9.1. Growth Proposition.* After a finite number of ierations, the function $i \rightarrow F(C^i)$ is non-decreasing.
*Proof.* After a finite number of iterations $i$, all solitary cells have died out. Let us consider the transition from $C^i$ to $C^{i+1}$. If a new cell is born, rule 6.5.1 implies that it cannot lie in the inhibitory neighborhood of an already present "on" cell. Thus $F$ will not decrease (from $C^i$ to $C^{i+1}$). If a cell dies out, it cannot be a solitary cell. Therefore it must have had an "on" cell in its inhibitory neighborhood at iteration $i$. Thus $F$ will not decrease.

The growth of $F$ adequately describes the filling-in process, respecting at the same time the "uniqueness" matching rule.

The growth proposition implies that:

*6.9.2.* For any initial configuration $C^0$, the limit $\mathrm{Lim} F(C^i) = F(C)$ exists (since $F$ is bounded above by the number of cells in one layer).

*6.9.3.* After a finite number of iterations, $F(C^i)$ remains constant.

Thus the asymptotic behavior of the system is characterized by the following: Apart from invariant solutions, only those cycles can (asymptotically) occur for which $F$ remains constant. This is a strong restriction on the possible asymptotic oscillatory states, for it means that they have to be of the type shown in Figure 13. The growth proposition by itself does not exclude for instance the "zero" invariant state. The probabilistic analysis of the second step of the algorithm together with the "filling-in" proposition ensures, however, that invariant states as well as limit cycles will in practice have no "holes".

### *6.10. Observations*

*6.10.1.* Figure 14 shows the performance of the algorithm in this form for a few different patterns and pattern densities. A comparison with Figure 8 reveals that the type of "strategy" for achieving a successful

matching is different: Firstly, wrong cells are drastically eliminated at the expense of losing many right cells, and then filling-in of the surviving surfaces takes place. This contrasts with the more complicated "strategy" revealed in Figures 3–6 of Marr and Poggio, 1976. It is remarkable how, while the basic structure of the algorithm remains the same, a change of parameter values and loading conditions can bring about so deep a change in the algorithms behavior.

*6.10.2.* Because of the rules of the present algorithm, especially rule 6.5.1, the sharpest corner capable of surviving (of course under no inhibitions) is limited only by the need for an "on" cell to have at least three excitatory neighboring cells. This allows a 45 degree corner.

*6.10.3. Minimum Size vs. Disparity.* Again because of rule 6.5.1, the minimum resolvable area of a small pattern against a background does not depend on disparity. It is given by the minimum self-supporting configuration (four adjacent cells, from rule 6.5.2).

This contrasts sharply with the property discussed in Section 5.4, where the minimum size has a characteristic dependence on disparity.

*6.10.4. Loading Conditions.* While the present loading rule is characterized by

$$f_i f_j = \delta_{ij}, \tag{14}$$

where $f_i$ is a feature and $\delta_{ij}$ is the Kronecker $\delta$,

the previous loading rule (Section 2) can also be characterized by (12) with the convention that $\delta_{ij}$ is also zero when either $i$ or $j$ are zero. In other words the "null" feature has a special status ($f_0 f_j = f_j f_0 = 0$, all $j$).

In case of two-valued (0, 1) random dot stereograms, the choice of either one of the two loading rules is somewhat arbitrary. For densities around 0.5, the straight Equation (14) seems to make more sense, since the black and white dots play equivalent roles. This is not clear, however, at very low densities (nor at very high ones).

In the case of natural images, more than two feature types have to be used (for instance, lines and edges at various orientations). In this case, however, not every point is labelled with a corresponding feature; the absence of any feature at a given point is a common event. The null feature seems to have a basically different role from the other features. These arguments clearly support the loading conditions used in the first part of the paper (see Marr and Poggio, 1976). It is clear, on the other hand, that both loading conditions may work. For both, an increasing number of "feature types" implies of course an increasingly better algorithm convergence. The choice between them depends in the end on the typical feature densities that one wants to deal with. For natural images, quantitative estimates have only recently become possible (Marr, 1976, 1977).

## 7. Discussion

### *7.1. Alternative Algorithms*

The algorithm (1) can be modified in various ways. One can adopt alternative loading rules for the network as in Section 6, and one can vary the parameters over a substantial range. Such apparently minor changes can cause considerable changes in the network's behavior, but often without changing the end result (see for instance Section 6), because they still implement the same computational constraints.

If the geometry of the local interactions (i.e. the shape of the excitatory and inhibitory neighborhoods) is changed, the network will in general implement a different computation, because the local constraints will have changed. If only the parameter values are changed, our analysis (Section 3) may still apply. If the geometry is changed, our analysis will in general become irrelevant.

Interestingly, for a specific stereogram density, a non-iterative version of our algorithm can recover disparity satisfactorily (see Fig. 14a iteration 1). John Fairfield (personal communication) suggested an algorithm in which 1) excitation is summed independently within each disparity layer, and 2) for each position, one selects only the most excited of the cells in the different disparity layers. This algorithm performs well for the case $v = 0.5$.

### *7.2. Comments on Analyzing Such Operations*

We find the style of analysis that we were forced to adopt to be unsatisfactory for a number of reasons. Firstly, although our arguments appear to provide a qualitatively accurate description of the algorithms's behavior, the arguments are not completely rigorous. The main reasons for this lie in the difficulty of assessing the validity of the randomness assumptions that are necessary for the probabilistic analysis; and, to a lesser extent, in the need to examine a number of special cases in order to establish the stability of various solutions.

Secondly, our analysis is very specific to the particular algorithm and the particular parameters. This style of proof cannot lead to any general results about the convergence of such operators.

In order to overcome the first of these problems one can follow the approach of Section 6. The price one pays is that the analysis is valid for a narrower parameter range, which happens not to include the

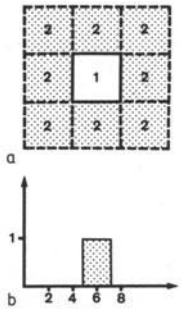



**Fig. 15.** Conway's game "life", which is played on an infinite plane square lattice, may be represented in a manner very similar to that of our stereo operator. The excitatory neighborhood, together with appropriate weights, is shown in 15a, and the threshold function appears in 15b. This combination reproduces the rules of life exactly, and these are; 1. A cell will die at generation $n+1$ if $<2$ or $>3$ of its 8 neighbors are alive at generation $n$ (death by starvation or overfeeding). 2. A cell with exactly 2 living neighbors at generation $n$ will be alive at generation $n+1$ if and only if it is alive at generation $n$. 3. A cell with exactly 3 living neighbors at generation $n$ will be alive at generation $n+1$

original parameter values (see Marr and Poggio, 1976). The difficulty with the assumption of randomness arises because of the constant spatial structure of the operator $\Xi$ (Eq. (1) and Fig. 2c of Marr and Poggio, 1976). It should perhaps be noted that this objection does not apply to the similar analysis given by Marr (1971) of a cooperative associative memory algorithm, because there the local operator had a variable and essentially random structure.

The second of these difficulties seems to be inherent in the nature of this type of cooperative algorithm. No general approach is at present available. Standard approaches[4] that we have tried have failed up to now. The flavor of the difficulties is the following. A configuration that is stable may be perturbed by changing a large number of cells without affecting its asymptotic state, provided that the perturbed cells are well scattered and interior. On the other hand, one fixed point of the algorithm can be shifted into another by perturbing only a few cells, provided that they have a suitable configuration. Thus, the usual distance between two configurations, namely the number of cells having different states, does not reflect the behavior of the algorithm. Therefore, the problem seems to be how to incorporate the geometry of the interactions into the metric distance between configurations.

It seems unlikely that one can construct a useful general theory of algorithms of the form

$$C^{n+1}=\sigma\{L(C^n)\}, \tag{15}$$

where $L$ is a linear operator on the vector $C$, and $\sigma$ is a nonlinear (coordinate-wise) function. J. H. Conway's game "Life" can, for example, be written this way (see Fig. 15) and with an appropriate input pattern is Turing universal (unpublished result discovered independently by J. H. Conway and R. W. Gosper).

This suggests that theories of this type of algorithm must take due account of the structure of the input data and will probably be restricted to very specific forms of (15).

A mathematical understanding of the behavior of (15) would represent a breakthrough of rather general importance. Cooperative phenomena similar to those which can be described by (15) are important in physics (Haken, 1977, Kawasaki, 1972; K. G. Wilson, 1975), in development (Mostow, 1975), and in biology (Eigen, 1971; Marr, 1971; Richter, 1976).

Furthermore, such a theory might also allow one to synthesize in a standard way cooperative algorithms of the form of (15) from an analysis of the constraints on a computation.

*Acknowledgements*: We thank K. P. Hadeler and W. Reichardt for interesting discussions, Karen Prendergast for preparing the illustrations, and the Mathlab Group for allowing us to run the probabilistic theory in Macsyma, the M.I.T. symbolic algebraic manipulation system. Pirooz Vatan helped with Macsyma programming. Science kindly gave permission for the reproduction of Figure 1. This work was conducted at the Artificial Intelligence Laboratory, a Massachusetts Institute of Technology research program supported in part by the Advanced Research Projects Agency of the Department of Defense, and monitored by the Office of Naval Research under contract number N00014-75-C-0643. D. M. thanks the Max-Planck-Institut für biologische Kybernetik in Tübingen for its kind hospitality during his visit there.

[4] The continuous version of the algorithm (1) cannot be described in terms of a potential dynamics. In fact the dynamical system

$$C=\sigma\{\Xi C\}-C=f(C) \quad (\sigma \text{ a "smooth" threshold})$$

does not admit a scalar potential function $V(C)$ such that

$$[f(C)]_{x,y;d}=\partial V(C)/\partial C_{x,y;d}.$$

A necessary condition for this to be true is that

$$\frac{\partial}{\partial C_{x,y;d}} f_{x'y'd'}(C)=\frac{\partial}{\partial C_{x',y';d'}} f_{xyd}(C), \quad \text{all} \quad x,y,d,x',y',d'.$$

This is not true in general, because of the nonlinearity $\sigma$ (consider the case in which $xyd$ and $x'y'd'$ are on the same disparity layer and are reciprocally excitatory)

## References

Eigen, M.: Self-organization of matter and the evolution of biological macromolecules. Naturwissenschaften **10**, 465—523 (1971)

Haken, H.: Synergetics. Berlin-Heidelberg-New York: Springer 1977

Kawasaki, K.: Kinetics of Ising models. In: Phase transitions and critical phenomena, Vol. 2, pp. 443—501. Domb and Green eds., New York: Academic Press 1972

Marr, D.: Simple memory: *s* theory for archicortex. Phil. Trans. Roy. Soc. B**262**, 23—81 (1971)

Marr, D.: A note on the computation of binocular disparity in a symbolic, low-level visual processor. M.I.T. A.I. Lab. Memo 327, (1974)
Marr, D.: Early processing of visual information. Phil. Trans. Roy. Soc. B**275**, 483—524 (1976)
Marr, D.: Representing visual information. AAAS 143rd Annual Meeting, Symposium on Some Mathematical Questions in Biology, February, (in press). Also available as M.I.T. A.I. Lab. Memo 415 (1977)
Marr, D., Poggio, T.: Cooperative computation of stereo disparity. Science **194**, 283—287 (1976)
Mostow: Mathemathical models for cell rearrangement. Newhaven, Yale University Press 1975
Richter, P.H.: The network idea and the immune response. In: Theoretical Immunology, Bell, G.I., Perelson, A.S., Pimbley, G.H., eds. New York: M. Dekker 1976
Wilson, H.R.: Hysteresis in binocular grating perception: contrast effects. Vision Res. (1977) (in press)
Wilson, H.R., Cowan, J.D.: A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. Kybernetik **13**, 55—80 (1973)
Wilson, K.G.: The renormalization group: Critical phenomena and the Kondo problem. Reviews Mod. Phys. **47**, 773—840 (1975)

Received: November 28, 1977

Dr. T. Poggio
MPI für biolog. Kybernetik
Spemannstr. 38
D-7400 Tübingen
Federal Republic of Germany

*Proc. R. Soc. Lond.* B. **204**, 301–328 (1979)

# A computational theory of human stereo vision†

BY D. MARR‡ AND T. POGGIO§

‡ *M.I.T. Psychology Department, 79 Amherst Street, Cambridge Ma 02139, U.S.A.*

§ *Max-Planck-Institut für Biologische Kybernetik, 7400 Tübingen, Spemannstrasse 38, Germany*

(*Communicated by S. Brenner, F.R.S. – Received 26 January* 1978)

An algorithm is proposed for solving the stereoscopic matching problem. The algorithm consists of five steps: (1) Each image is filtered at different orientations with bar masks of four sizes that increase with eccentricity; the equivalent filters are one or two octaves wide. (2) Zero-crossings in the filtered images, which roughly correspond to edges, are localized. Positions of the ends of lines and edges are also found. (3) For each mask orientation and size, matching takes place between pairs of zero-crossings or terminations of the same sign in the two images, for a range of disparities up to about the width of the mask's central region. (4) Wide masks can control vergence movements, thus causing small masks to come into correspondence. (5) When a correspondence is achieved, it is stored in a dynamic buffer, called the $2\frac{1}{2}$-D sketch.

It is shown that this proposal provides a theoretical framework for most existing psychophysical and neurophysiological data about stereopsis. Several critical experimental predictions are also made, for instance about the size of Panum's area under various conditions. The results of such experiments would tell us whether, for example, cooperativity is necessary for the matching process.

## COMPUTATIONAL STRUCTURE OF THE STEREO-DISPARITY PROBLEM

Because of the way our eyes are positioned and controlled, our brains usually receive similar images of a scene taken from two nearby points at the same horizontal level. If two objects are separated in depth from the viewer, the relative positions of their images will differ in the two eyes. Our brains are capable of measuring this disparity and of using it to estimate depth.

Three steps (S) are involved in measuring stereo disparity: (S1) a particular location on a surface in the scene must be selected from one image; (S2) that same location must be identified in the other image; and (S3) the disparity in the two corresponding image points must be measured.

If one could identify a location beyond doubt in the two images, for example by illuminating it with a spot of light, steps S1 and S2 could be avoided and the

† A preliminary and lengthier version of this theory is available from the M.I.T. A.I. Laboratory as Memo 451 (1977).

problem would be easy. In practice one cannot do this (figure 1), and the difficult part of the computation is solving the correspondence problem. Julesz (1960) found that we are able to interpret random dot stereograms, which are stereo pairs that consist of random dots when viewed monocularly but fuse when viewed stereoscopically to yield patterns separated in depth. This might be thought surprising, because when one tries to set up a correspondence between two arrays of random dots, false targets arise in profusion (figure 1). Even so and in the absence of any monocular or high level cues, we are able to determine the correct correspondence.

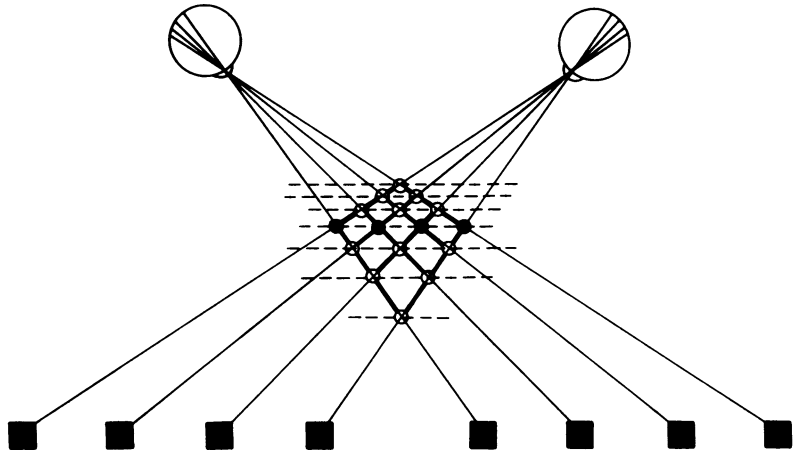


FIGURE 1. Ambiguity in the correspondence between the two retinal projections. In this figure, each of the four points in one eye's view could match any of the four projections in the other eye's view. Of the 16 possible matchings only four are correct (filled circles), while the remaining 12 are 'false targets' (open circles). It is assumed here that the targets (filled squares) correspond to 'matchable' descriptive elements obtained from the left and right images. Without further constraints based on global considerations, such ambiguities cannot be resolved. Redrawn from Julesz (1971, fig. 4.5–1).

In order to formulate the correspondence computation precisely, we have to examine its basis in the physical world. Two constraints (C) of importance may be identified (Marr 1974): (C1) a given point on a physical surface has a unique position in space at any one time; and (C2) matter is cohesive, it is separated into objects, and the surfaces of objects are generally smooth compared with their distance from the viewer.

These constraints apply to locations on a physical surface. Therefore, when we translate them into conditions on a computation we must ensure that the items to which they apply in the image are in one-to-one correspondence with well-defined locations on a physical surface. To do this, one must use image predicates that correspond to surface markings, discontinuities in the visible surfaces, shadows, and so forth, which in turn means using predicates that correspond to changes in intensity. One solution is to obtain a primitive description of the intensity changes present in each image, like the primal sketch (Marr 1976), and then to match these descriptions. Line and edge segments, blobs, termination

points, and tokens, obtained from these by grouping, usually correspond to items that have a physical existence on a surface.

The stereo problem may thus be reduced to that of matching two primitive symbolic descriptions, one from each eye. One can think of the elements of these descriptions as carrying only position information, like the black dots in a random dot stereogram, although for a full image there will exist rules that specify which matches between descriptive elements are possible and which are not. The two physical constraints C1 and C2 can now be translated into two rules (R) for how the left and right descriptions are combined:

(*R*1) *Uniqueness.* Each item from each image may be assigned at most one disparity value. This condition relies on the assumption that an item corresponds to something that has unique physical position.

(*R*2) *Continuity.* Disparity varies smoothly almost everywhere. This condition is a consequence of the cohesiveness of matter, and it states that only a small fraction of the area of an image is composed of boundaries that are discontinuous in depth.

In practice, R1 cannot be applied simply to grey level points in an image, because a grey level point is in only implicit correspondence with a physical location. It is in fact impossible to ensure that a grey level point in one image corresponds to exactly the same physical position as a grey level point in the other. A sharp change in intensity, however, usually corresponds to a surface marking, and therefore defines a single physical position precisely. The positions of such changes may be detected by finding peaks in the first derivative of intensity, or zero-crossings in the second derivative.

In a recent article, Marr & Poggio (1976) derived a cooperative algorithm which implements these rules (see figure 2), showing that it successfully solves the false targets problem and extracts disparity information from random dot stereograms (see also Marr, Palm & Poggio 1978).

## The biological evidence

Apart from AUTOMAP (Julesz 1963) and Sperling (1970), all of the current stereo algorithms proposed as models for human stereopsis are based on Julesz's (1971) proposal that stereo matching is a cooperative process (Julesz 1971, p. 203 ff.; Julesz & Chang 1976; Nelson 1975; Dev 1975; Hirai & Fukushima 1976; Sugie & Suwa 1977; Marr & Poggio 1976). None of them has been shown to work on natural images.

An essential feature of these algorithms is that they are designed to select correct matches in a situation where false targets occur in profusion. They require many 'disparity detecting' neurons, whose peak sensitivities cover a range of disparity values that is much wider than the tuning curves of the individual neurons. That is, apart possibly from early versions of Julesz's dipole model, they do not critically rely on eye movements, since in principle, they have the ability to interpret a random dot stereogram without them.

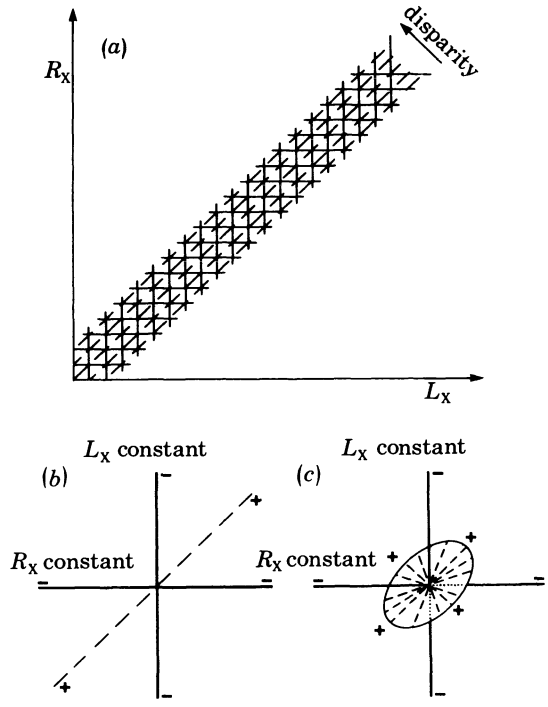



FIGURE 2. The explicit structure of the two rules R1 and R2 for the case of a one dimensional image is represented in (*a*). $L_X$ and $R_X$ represent the positions of descriptive elements in the left and right images. The continuous vertical and horizontal lines represent lines of sight from the left and the right eyes. Their intersection points correspond to possible disparity values. R1 states that only one match is allowed along any given horizontal or vertical line; R2 states that solution planes tend to spread along the dotted diagonal lines, which are lines of constant disparity.

In a network implementation of these rules, one can place a 'cell' at each node; then solid lines represent 'inhibitory' interactions, and dotted lines represent 'excitatory' ones. The local structure at each node of the network in (*a*) is given in (*b*). This algorithm may be extended to two dimensional images, in which case each node in the corresponding network has the local structure shown in (*c*). The ovals in this figure represent circular two dimensional disks rising out of the plane of the page. Formally, the algorithm represented by this network may be written as the iterative algorithm

$$C^{t+1}_{x,y;d} = \sigma\{\sum_{x',y',d' \in S(x,y,d)} C^{t}_{x',y';d'} - \epsilon \sum_{x',y',d' \in O(x,y,d)} C^{t}_{x',y';d'} + C^{0}_{x,y;d}\},$$

where $C^{t}_{x,y;d}$ denotes the state of the cell (0 for inactive, 1 for active) corresponding to position $(x, y)$, disparity $d$ and time $t$ in the network of (*a*); $S(x, y, d)$ is a local excitatory neighbourhood confined to the same disparity layer, and $(x, y, d)$ the inhibitory neighbourhood, consists of cells lying on the two lines of sight (*c*). $\epsilon$ is an inhibition constant, and $\sigma$ is a threshold function. The initial state $C^0$ contains all possible matches, including false targets, within the prescribed disparity range. The rules R1 and R2 are implemented through the geometry of the inhibitory and excitatory neighbourhoods $O$ and $S$ (*c*). (From Marr & Poggio 1976, fig. 2; copyright by the American Association for the Advancement of Science.)

Eye movements seem, however, to be important for human stereo vision (Richards 1977; Frisby & Clatworthy 1975; Saye & Frisby 1975). Other findings these algorithms fail to explain include (*a*) the ability of some subjects to tolerate a 15% expansion of one image (Julesz 1971, fig. 2.8–8), (*b*) the findings about independent spatial-frequency-tuned channels in binocular fusion, of which our tolerance to severe defocusing of one image is a striking demonstration (Julesz 1971, fig. 3.10–3), (*c*) the physiological, clinical, and psychophysical evidence about Richards' two pools hypothesis (Richards 1970, 1971; Richards & Regan 1973); and (*d*) the size of Panum's fusional area (6′–18′, Fender & Julesz 1967; Julesz & Chang 1976) which seems surprisingly small to have to resort to cooperative mechanisms for the elimination of false targets.

Taken together, these findings indicate that a rather different approach is necessary. In this article, we formulate an algorithm designed specifically as a theory of the matching process in human stereopsis, and present a theoretical framework for the overall computational problem of stereopsis. We show that our theory accounts for most of the available evidence and formulate the predictions to which it leads.

For a more comprehensive review of the relevant psychophysics and neurophysiology see Marr & Poggio (1977*a*).

## An outline of the theory

The basic computational problem in binocular fusion is the elimination of false targets, and for any given monocular features the difficulty of this problem is in direct proportion to the range and resolution of the disparities that are considered. The problem can therefore be simplified by reducing either the range, or the resolution, or both, of the disparity measurements that are taken from two images. An extreme example of the first strategy would lead to a diagram like figure 2 in which only three adjacent disparity planes were present (e.g. $+1$, $0$, $-1$) each specifying their degree of disparity rather precisely. The second strategy, on the other hand, would amount to maintaining the range of disparities shown in figure 2, but reducing the resolution with which they are represented. In the extreme case, only three disparity values would be represented, crossed, roughly zero, and uncrossed.

These schemes, based on just three pools of disparity values, substantially eliminate the false targets problem at the cost on the one hand of a very small disparity range, and on the other, of poor disparity resolution. Thus the price of computational simplicity is a trade-off between range and resolution.

One would, however, expect the human visual system to possess both range and resolution in its disparity processing. In this connection, the existence of independent spatial frequency tuned channels in binocular fusion (Kaufman 1964; Julesz 1971, §§ 3.9 and 3.10; Julesz & Miller 1975; Mayhew & Frisby 1976) is of especial interest, because it suggests that several copies of the image, obtained

by successively finer filtering, are used during fusion, providing increasing and, in the limit, very fine disparity resolution at the cost of decreasing disparity range.

A notable feature of a system organized along these lines is its reliance on eye movements for building up a comprehensive and accurate disparity map from two viewpoints. The reason for this is that the most precise disparity values are obtainable from the high resolution channels, and eye movements are therefore essential so that each part of a scene can ultimately be brought into the small disparity range within which high resolution channels operate. The importance of vergence eye movements is also attractive in view of the extremely high degree of precision with which they may be controlled (Riggs & Niehl 1960; Rashbass & Westheimer 1961*a*).

These observations suggest a scheme for solving the fusion problem in the following way (Marr & Poggio 1977*a*, *b*): (1) Each image is analysed through channels of various coarsenesses, and matching takes place between corresponding channels from the two eyes for disparity values of the order of the channel resolution. (2) Coarse channels control vergence movements, thus causing finer channels to come into correspondence.

This scheme contains no hysteresis, and therefore does not account for the hysteresis observed by Fender & Julesz (1967). Recent work in the theory of intermediate visual information processing argues on computational grounds that a key goal of early visual processing is the construction of something like an 'orientation and depth map' of the visible surfaces round a viewer (Marr & Nishihara 1978, fig. 2; Marr 1977, § 3). In this map, information is combined from a number of different and probably independent processes that interpret disparity, motion, shading, texture, and contour information. These ideas are illustrated by the representation shown in figure 3, which Marr & Nishihara called the $2\frac{1}{2}$-D sketch.

Suppose now that the hysteresis Fender & Julesz observed is not due to a co-operative process during matching, but is in fact the result of using a memory buffer, like the $2\frac{1}{2}$-D sketch, in which to store the depth map of the image as it is discovered. Then, the matching process itself need not be cooperative (even if it still could be), and in fact it would not even be necessary for the whole image ever to be matched simultaneously, provided that a depth map of the viewed surface were built and maintained in this intermediate memory.

Our scheme can now be completed by adding to it the following two steps: (3) when a correspondence is achieved, it is held and written down in the $2\frac{1}{2}$-D sketch; (4) there is a backwards relation between the memory and the masks, acting through the control of eye movements, that allows one to fuse any piece of a surface easily once its depth map has been established in the memory.

## The nature of the channels

The articles by Julesz & Miller (1975) and Mayhew & Frisby (1976) establish that spatial-frequency-tuned channels are used in stereopsis and are independent. Julesz & Miller's findings imply that two octaves is an upper bound for the bandwidth of these channels, and suggest that they are the same channels as those previously found in monocular studies (Campbell & Robson 1968; Blakemore & Campbell 1969). Although strictly speaking it has not been demonstrated that these two kinds of channel are the same, we shall make the assumption that they are. This will allow us to use the numerical information available from monocular studies to derive quantitative estimates of some of the parameters involved in our theory.

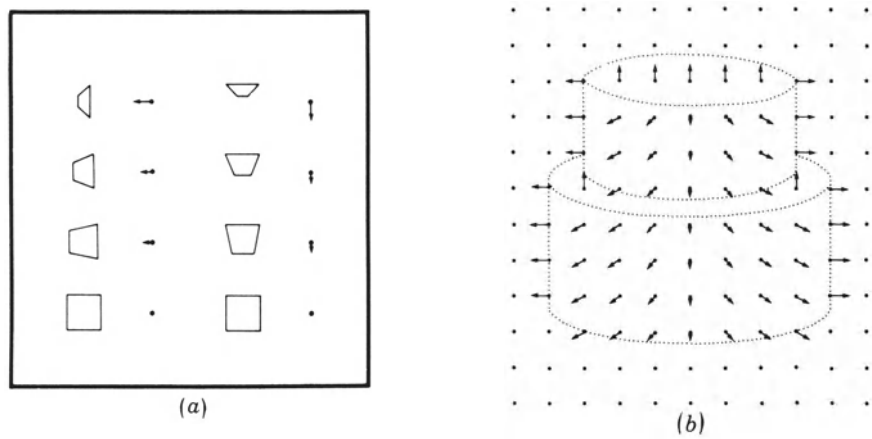



FIGURE 3. Illustration of the $2\frac{1}{2}$-D sketch. In (*a*) the perspective views of small squares placed at various orientations to the viewer are shown. The dots with arrows show a way of representing the orientations of such surfaces symbolically. In (*b*), this representation is used to show the surface orientations of two cylindrical surfaces in front of a background orthogonal to the viewer. The full $2\frac{1}{2}$-D sketch would include rough distances to the surfaces as well as their orientations, contours where surface orientation changes sharply, and contours where depth is discontinuous (subjective contours). A considerable amount of computation is required to maintain these quantities in states that are consistent with one another and with the structure of the outside world (see Marr 1977, § 3). (From Marr & Nishihara 1978, fig. 2.)

The idea that there may be a range of different sized or spatial-frequency-tuned mechanisms was originally introduced on the basis of psychophysical evidence by Campbell & Robson (1968). This led to a virtual explosion of papers dealing with spatial frequency analysis in the visual system. Recently, Wilson & Giese (1977) and Cowan (1977) integrated these and other anatomical and physiological data into a coherent logical framework. The key to their framework is (*a*) the partitioning of the range of sizes associated with the channels into two components, one due to spatial inhomogeneity of the retina, and one due to local scatter of

receptive field sizes; and (*b*) the correlation of these two components with anatomical and physiological data about the scatter of receptive field sizes and their dependence on eccentricity.

On the basis of detection studies, they formulated an initial model embodying the following conclusions: (1) at each position in the visual field, there exist 'bar-like' masks (see figure 4*a*), whose tuning curves have the form of figure 4*b*, and which have a half power bandwidth of between one and two octaves. (2) The half power bandwidth of the local sensitivity function at each eccentricity

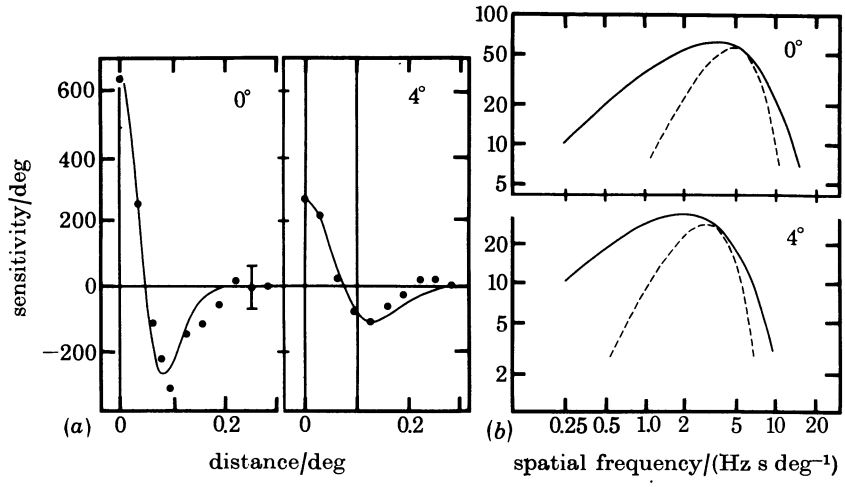



FIGURE 4. (*a*) Line spread functions measured psychophysically at threshold at two different eccentricities. The points are fitted using the difference of two Gaussian functions with space constants in the ratio 1.5 : 1.0. The inhibitory surround exactly balances the excitatory centre so that the area under the curve is zero. (*b*) Predictions of local spatial frequency sensitivity from frequency gradient data and from line spread function data. The local frequency sensitivity functions are plotted as solid lines. The dashed lines are the local frequency response predicted by Fourier transforming the line spread functions in (*a*), which were measured at the appropriate eccentricities. (Redrawn from Wilson & Giese 1977, fig. 9 and 10.)

is about three octaves. Hence the range of receptive field sizes present at each eccentricity is about 4 : 1. In other words, at least three and probably four receptive field sizes are required at each point of the visual field. (3) Average receptive field size increases linearly with eccentricity. In humans at 0° the mean width $w$ of the central excitatory region of the mask is about 6′ (range 3′–12′); and at 4° eccentricity, $w = 12'$ (range 6′–24′) (Wilson & Giese 1977, fig. 9; Hines 1976, figs 2 and 3). If one assumes that this receptive field is described by the difference of two Gaussian functions with space constants in the ratio 1 : 1.5, the corresponding peak frequency sensitivity of the channel is given by $1/f = \lambda = 2.2w$. These figures agree quite well with physiological studies in the Macaque. Hubel & Wiesel (1974, fig. 6*a*) reported that the mean width of the receptive field ($s$) increases linearly with eccentricity $e$ (approximately, $s = 0.05e + 0.25°$, so that at

$e = 4°$, $s = 27'$ which gives a value for $w = \frac{1}{3}s$ of about $9'$ as opposed to the figure of $12'$ assumed here for humans). The data of Schiller, Finlay & Volman (1977, p. 1347, figs. 12 and 14) are in rough agreement with Hubel & Wiesel's. (4) Essentially all of the psychophysical data on the detection of spatial patterns at contrast threshold can be explained by (1), (2) and (3) together with the hypothesis that the detection process is based on a form of spatial probability summation in the channels.

With the characteristic perverseness of the natural world, this happy and concise state of affairs does not provide a precise account of suprathreshold conditions. The known discrepancies can however be explained by introducing two extra hypotheses: (5) contrast sensitivities of the various channels are adjusted appropriately to the stimulus contrast (Georgeson & Sullivan 1975). The point of this is merely to ensure that bars of the same contrast but different widths actually appear to have the same contrast; (6) receptive field properties change slightly with contrast, the inhibition being somewhat decreased when contrast is low (Cowan 1977, p. 511).

In a more recent article, Wilson & Bergen (1979) have found that the situation at threshold may also be more complicated. They proposed a model consisting of four size-tuned mechanisms centred at each point, the smaller two showing relatively sustained temporal responses, and the larger two being relatively transient. As far as is known, this model accurately accounts for all published threshold sensitivity studies.

The two sustained channels, which Wilson & Bergen call N and S, have $w$ values $3.1'$ and $6.2'$; the transient channels, called T and U, have $w$ equal to $11.7'$ and $21'$. The sizes of these channels increase with eccentricity in the same way as described above.

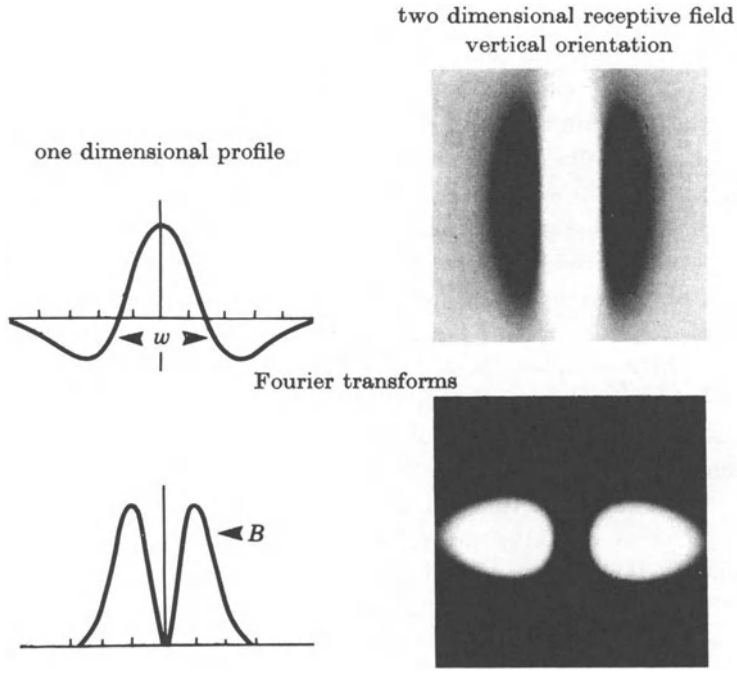
The S channel is the most sensitive under both transient and sustained stimulation, and the U channel is the least, having only $\frac{1}{11}$ to $\frac{1}{4}$ the sensitivity of the S channel. The extent to which the U channel, for example, plays a role in stereopsis is of course unknown.

In what follows, we shall assume that the figures given by Wilson & Giese for the numbers and dimensions of receptive field centres and their scatter hold roughly for suprathreshold conditions. If future experiments confirm that Wilson & Bergen's more recent numbers are relevant for stereopsis, some modification of our quantitative estimates may be necessary.

Wilson & Giese's figures allow us to estimate the minimum sampling density required by each channel, i.e. the minimum spatial density of the corresponding receptive fields. From fig. 10 of Wilson & Giese (1977), a channel with peak sensitivity at wavelength $\lambda$ is band-limited on the high frequency side by wavelengths of about $\frac{2}{3}\lambda$, and $\lambda = 2.2w$. This figure is for a threshold criterion of 15–30 %, but is rather insensitive to the exact value chosen. Hence by the sampling theorem (Papoulis 1968, p. 119), the minimum distance between samples (i.e. receptive fields), in a direction perpendicular to their preferred orientation, is at

TABLE 1. SPATIAL FILTERING: SUMMARY OF PSYCHOPHYSICAL EVIDENCE

(*a*) At each point in the visual field the image is filtered through receptive fields having these characteristics (the half-power bandwidth $B$ is about 1 octave):




(*b*) For each position and orientation there are four receptive field sizes, the smallest being $\frac{1}{4}$ of the largest. The profile $R(x)$ and Fourier transform $\hat{R}(\omega)$ of each receptive field are given by:

$$R(x) = (2\pi)^{-\frac{1}{2}}\{\sigma_e^{-1}\exp[-x^2/2\sigma_e^2] - \sigma_i^{-1}\exp[-x^2/2\sigma_i^2]\},$$
$$\hat{R}(\omega) = \exp[-\tfrac{1}{2}\omega^2\sigma_e^2] - \exp[-\tfrac{1}{2}\omega^2\sigma_i^2],$$

where $\sigma_e$, $\sigma_i$ are the excitatory and inhibitory space constants, and are in the ratio 1 : 1.5. The half-power bandwidth spanned by the four receptive field cells at each point is two octaves.

(*c*) $w$ increases with eccentricity: $w = 3' - 12'$ (possibly $20'$) at 0°, and $w = 6' - 34'$ at 4°. *Note.* receptive field sizes and corresponding spectral sensitivity curves in the suprathreshold condition may be different from the values given here, which were measured at threshold.

(*d*) Formally the output of step 1 is given by the convolution $F_{w,\theta}(x, y) = I * B_{w,\theta}$, where $I(x, y)$ denotes the light intensity in suitable units at the image point $(x, y)$, and $B_{w,\theta}(x, y)$ describes the receptive field of a bar-shaped mask at orientation $\theta$, with central region of width $w$. $\theta$ covers the range 0–180° with 12 values about 15° apart and $w$ takes four values in the range defined by (*c*) above.

(*e*) In practice, cells with on-centre receptive fields will signal positive values of the filtered signal, and cells with off-centre receptive fields will signal negative values.

most $\frac{1}{3}\lambda$. Assuming the overall width of the receptive field is about $\frac{3}{2}\lambda$, the minimum number of samples per receptive field width is about 4.5.

An estimate of the minimum longitudinal sampling distance may be obtained as follows. Assume that the receptive field's longitudinal weighting function (see table 1) is Gaussian with space-constant $\sigma$, thus extending over an effective distance of say $4\sigma$–$6\sigma$. (A value of $\sigma = w$ will give an approximately square receptive field.) Its Fourier transform is also Gaussian with space constant in the frequency domain ($\omega$) of $1/\sigma$, and for practical purposes can be assumed to be band-limited with $f_{\max} = 3/(2\pi\sigma)$ to $2/(2\pi\sigma)$. By the sampling theorem, the corresponding minimum sampling intervals are $\sigma$ to $1.5\sigma$, that is about four samples per longitudinal receptive field distance. Hence the minimum number of measurements (i.e. cells or receptive fields) per receptive field area is about 18. It follows that the number of multiplications required to process the image through a given channel is roughly independent of the receptive field size associated with that channel. Not too much weight should be attached to the estimate of 18, although we feel that the sampling density cannot be significantly lower. In the biological situation, total sampling density will decrease as eccentricity increases.

This model of the preliminary processing of the image is summarized in table 1. There are in fact more efficient ways of implementing it (see Marr & Hildreth 1979).

## The domain of the matching function

In view of this information, the first step in our theory can be thought of as filtering the left and right spatial images through bar masks of four sizes and about twelve orientations at each point in the images. We assume that this operation is roughly linear, for a given intensity and contrast. When matching the left and right images, one cannot simply use the raw values measured by this first stage, because they do not correspond directly to physical features on visible surfaces on which matching may be based. One first has to obtain from these measurements some symbol that corresponds with high probability to a physical item with a well defined spatial position. This observation, which has been verified through computer experiments in the case of stereo vision (Grimson & Marr 1979) formed the starting point for a recent approach to the early processing of visual information (Marr 1974, 1976).

Perhaps the simplest way of obtaining suitable symbols from an image is to find signed peaks in the first (directional) derivative of the intensity array, or alternatively, zero-crossings in the second derivative. The bar masks of table 1 measure an approximation to the second directional derivative at roughly the resolution of the mask size, provided that the image does not vary locally along the orientation of the mask (Marr & Hildreth 1979). If this is so, clear signed zero-crossings in the convolution values obtained along a line lying perpendicular to the receptive field's longitudinal axis (cf. Marr 1976, fig. 2) would specify

accurately the position of an edge in the image.† Edges whose orientations lie near the vertical will of course play a dominant role in stereopsis.

In practice, however, it is not enough to use just oriented edges to obtain horizontal disparity information. Julesz (1971, p. 80) showed that minute breaks in horizontal lines can lead to fusion of two stereograms even when the breaks lie close to the limit of visual acuity, and such breaks cannot be obtained by simple operations on the measurements from even the smallest vertical masks. These breaks probably have to be localized by a specialized process for finding terminations by examining the values and positions of rows of zero-crossings (cf. Marr 1976, p. 496).

Thus zero-crossings and (less importantly) terminations have both to be made explicit (cf. Marr 1976, p. 485). The matching process will then operate on descriptions, of the left and right images, that are built of these two kinds of symbolic primitives, and which specify their precise positions, the mask size and orientation from which they were obtained, and their signs.

## Matching

At the heart of the matching problem lies the problem of false targets. If each channel were very narrowly tuned to a wavelength $\lambda$, the minimum distance between zero-crossings of the same sign in each image would be about $\lambda$. In this case, matching would be unambiguous in a disparity range up to $\lambda$. The same argument holds qualitatively for the actual channels, but because they are not so narrowly tuned, the disparity range for unambiguous matching will be smaller and must be estimated. We have done this only for zero-crossings, since terminations are sparser and pose less of a false-target problem.

Let us consider a two dimensional image filtered through a vertically oriented mask. Matching will take place between zero-crossings of the same sign along corresponding horizontal lines in the two images. If two such zero-crossings lie very close together in one image, the danger of false targets will arise. Hence a critical parameter in our analysis will be the distance between adjacent zero-crossings of the same sign along each of these lines.

This problem is now one dimensional, and we approach it by estimating the probability distribution of the interval between adjacent zero-crossings of the same sign. This depends on (*a*) the image characteristics, and (*b*) the filter (or mask) characteristics. For (*a*) we take the worst case, that in which the power spectrum of the input to the filter is white (within the filter's spectral range). We also assume, for computational convenience, that the filtered output is a

† It is perhaps worth noting that this rather direct way of locating sharp intensity changes in the image is not the only nor necessarily the best method from the point of view of an actual implementation. It is shown elsewhere (Marr & Hildreth 1979) that under certain conditions, the zeroes in an image filtered through a Laplacian operator (like an X-type retinal ganglion cell) provide an equivalent way of locating edges, whose orientation must then be determined.

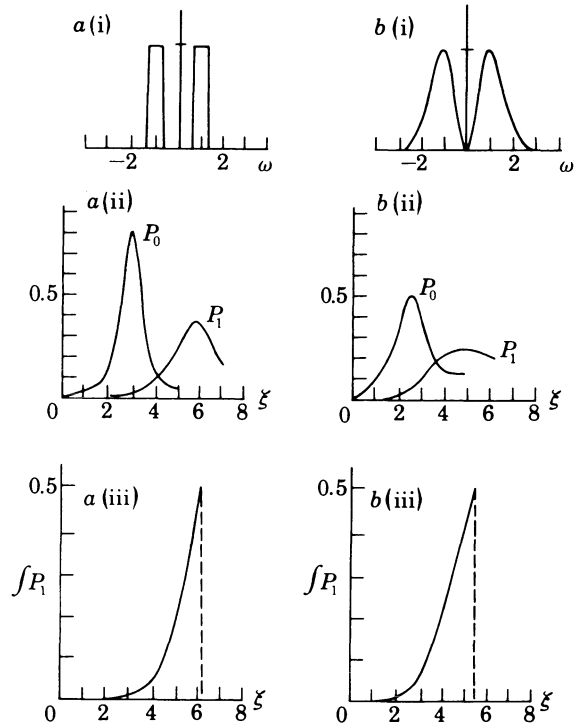



FIGURE 5. Interval distributions for zero-crossings. A 'white' Gaussian random process is passed through a filter with the frequency characteristic (transfer function) shown in (i). The approximate interval distribution for the first ($P_0$) and second ($P_1$) zero-crossings of the resulting zero-mean Gaussian process is shown in (ii). Given a positive zero-crossing at the origin, the probability of having another within a distance $\xi$ is approximated by the integral of $P_1$ and shown in (iii). In (*a*), these quantities are given for an ideal band pass filter one octave wide and with centre frequency $\omega = 2\pi/\lambda$; (*b*) represents the case of the receptive field described by Cowan (1977) and Wilson & Giese (1977). The corresponding spatial distribution of excitation and inhibition, i.e. the inverse Fourier transform of (*b*i) appears, in the same units, in table 1. The ratio of space constants for excitation and inhibition is 1 : 1.5. The width $w$ of the central excitatory portion of the receptive field is 2.8 in the units in which $\xi$ is plotted.

For case (*a*) a probability level of $\int P_1 = 0.001$ occurs at $\xi = 2.3$, and a probability level of 0.5 occurs at $\xi = 6.1$. The corresponding figures for case (*b*) are $\xi = 1.5$ and $\xi = 5.4$. If the space constant ratio is 1 : 1.75 (Wilson 1978*b*) the values of $\int P_1$ change by not more than 5 %.

Gaussian (zero-mean) process. This hypothesis is quite realistic (E. Hildreth, personal communication).

For (*b*), we examine two cases. Since the actual filters have a half-power bandwidth of around one octave, the first case we consider is that of an ideal linear bandpass filter of width one octave, as illustrated in figure 5*a*(i). The second case (figure 5*b*(i)) is the receptive field suggested by the threshold experiments of Wilson & Giese (1977), consisting of excitatory and inhibitory Gaussian distributions, with space constants in the ratio 1 : 1.5 (see figure 4).

Our problem is now transformed into one that many authors have considered, dating from the pioneering work of Rice (1945), and the appendix sets out the formulae in detail. The results we need are all contained in figure 5. $P_0$ is the probability distribution of the interval between adjacent zero-crossings (which perforce have opposite signs), and $P_1$ the distribution of the interval to the second zero-crossing. Since alternate zero-crossings have the same sign, $P_1$ is the quantity of interest, and its integral $\int P_1$ is also given in figure 5.

$\int P_1$ can be understood in the following way. Suppose a positive zero-crossing occurs at the point O. Then $\int P_1$ represents the probability that at least one other positive zero-crossing will occur within a distance $\xi$ of O. (In figure 5*b*(iii), the width $w$ of the central part of the receptive field associated with the filter is equal to 2.8 on the $\xi$ scale.)

From the graphs in figure 5, we see for example that the 0.05 probability level for $\int P_1$ occurs at $\xi = 4.1$ (approximately $\lambda/1.52$) for the ideal band pass filter one octave wide, centred on wavelength $\lambda$ (figure 5*a*(i)), and at $\xi = 3.1$ for the receptive field of fig. 5*b*(i). In the second case, $\xi$ is approximately $\frac{1}{2}\lambda$, where $\lambda$ is the principal wavelength associated with the channel, and $\lambda = 2.2w$, where $w$ is the measured width of the central excitatory area of the receptive field. Thus in this case, the 95% confidence limit occurs at approximately $w$ ($\xi = 3.1$, $w = 2.8$).

At the 0.001 probability level, the ideal bandpass filter is 50% better (the corresponding $\xi$ is 50% larger) than the receptive field filter with the same centre frequency; at the 0.05 probability level it is 30% better; and at the 0.5 probability level, it is 13% better. The legend to figure 5 provides more details about these results.

We have made a similar comparison between the sustained and transient channels of Wilson (1978*a*) and of Wilson & Bergen (1979). If the sustained channels correspond to the case of figure 5*b*, the transient channels have a larger ratio of the space constants for inhibition and excitation, a somewhat larger excitatory space constant, and an excitatory area larger than the inhibitory. Even under these conditions, the values change only slightly.

### *The matching process*

We now apply the results of these calculations to the matching process. Our analysis applies directly to channels with vertical orientation, and is roughly valid for channels with orientation near the vertical.

Within a channel of given size, there are in practice two possible ways of dealing with false targets. If one wishes essentially to avoid them altogether, the disparity range over which a match is sought must be restricted to $\pm\frac{1}{2}w$ (see figure 6*a*). For suppose zero-crossing L in the left image matches zero-crossing R in the right image. The above calculations assure us that the probability of another zero-crossing of the same sign within $w$ of R in the right image is less than 0.05. Hence if the disparity between the images is less than $\frac{1}{2}w$, a search for matches in the range $\pm\frac{1}{2}w$ will yield only the correct match R (with probability

0.95). Such a low error rate can be accommodated without resorting to sophisticated algorithms. For example, two reasonable ways of increasing the matching reliability are (*a*) to demand rough agreement between the slopes of the matched zero-crossings, and (*b*) to fail to accept an isolated match all of whose neighbours give different disparity values. Of course if the disparity between the images exceeds $\frac{1}{2}w$, this procedure will fail, a circumstance that we discuss later.

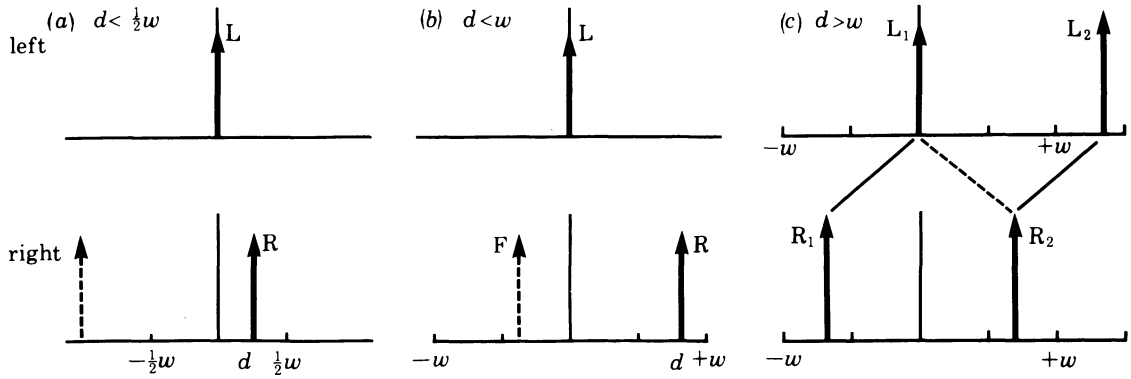



FIGURE 6. The matching process driven from the left image. A zero-crossing L in the left image matches one R displaced by disparity $d$ in the right image. The probability of a false target within $w$ of R is small, so provided that $d < \frac{1}{2}w$ (case *a*), almost no false targets will arise in the disparity range $\pm\frac{1}{2}w$. This gives the first possible algorithm. Alternatively (case *b*), all matches within the range $\pm w$ may be considered. Here, false targets (F) can arise in about 50 % of the cases, but the correct solution is also present. If the correct match is convergent, the false target will with high probability be divergent. Therefore in the second algorithm, unique matches from either image are accepted as correct, and the remainder as ambiguous and subject to the 'pulling effect', illustrated in case (*c*). Here, $L_1$ could match $R_1$ or $R_2$, but $L_2$ can match only $R_2$. Because of this, and because the two matches have the same disparity, $L_1$ is assigned to $R_1$.

There is, however, an alternative strategy, that allows one to deal with the matching problem over a larger disparity range. Let us consider the possible situations if the disparity between the images is $d$, where $|d| < w$ (figure 6*b*). Observe firstly that if $d > 0$, the correct match is almost certainly ($p < 0.05$) the only convergent candidate in the range $(0, w)$. Secondly, the probability of a (divergent) false target is at most 0.5. Therefore, 50 % of all possible matches will be unambiguous and correct, and the remainder will be ambiguous, mostly consisting of two alternatives, one convergent and one divergent, one of which is always the correct one. In the ambiguous cases, selection of the correct alternative can be based simply on the sign of neighbouring unambiguous matches. This algorithm will fail for image disparities that significantly exceed $\pm w$, since the percentage of unambiguous matches will be too low (roughly 0.2 for $\pm 1.5w$). Notice that if there is a match near zero disparity, it is likely ($p > 0.9$) to be the only candidate.

Sparse images like an isolated line or bar, that yield few or no false targets, pose a different problem. They often give rise to unique matches, and may there-

fore be relied upon over quite a large disparity range. Hence if the above strategy fails to disclose candidate matches in its disparity range, the search for possible matches may proceed outwards, ceasing as soon as one is found.

In summary then there are two immediate candidates for matching algorithms. The simpler is restricted to a disparity range of $\pm \frac{1}{2}w$ and in its most straightforward form will fail to assign 5% of the matches. The second involves some straightforward comparisons between neighbouring matches, but even before these comparisons, the 50% unambiguous matches could be used to drive eye movements, and provide a rough sensation of depth.

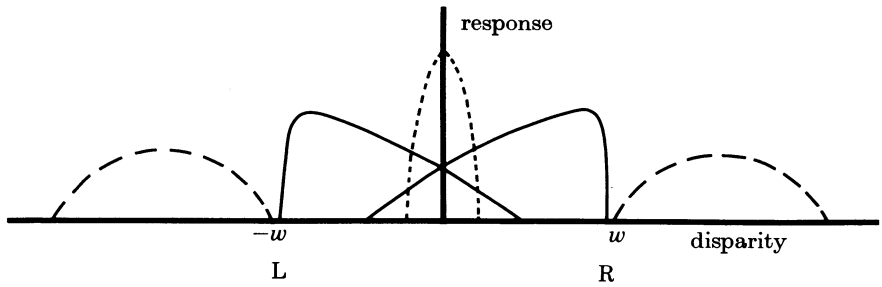



FIGURE 7. An implementation of the second matching algorithm. For each mask size of central width $w$, there are two pools of disparity detectors, signalling crossed or uncrossed disparities and spanning a range of $\pm w$. There may be additional detectors finely tuned to near-zero disparities. Additional diplopic disparities probably exist beyond this range. They are vetoed by detectors of smaller absolute disparity.

The implementation of the first of these algorithms is straightforward. The second one can be implemented most economically using two 'pools', one sensitive in a graded way to convergent and the other to divergent disparities (see figure 7). (In this sense, the first algorithm requires only one 'pool', that is, a single unit sensitive in a graded way to the disparity range $\pm \frac{1}{2}w$.) Candidate matches near zero disparity are likely to be correct, and this fact can be used to improve performance. One way is to add, to the two basic pools, high resolution units tuned to near-zero disparities.

In the second algorithm, matches that are unambiguous or already assigned can 'pull' neighbouring ambiguous matches to whichever alternative has the same sign. This is a form of cooperativity, and may be related to the 'pulling effect' described in psychophysical experiments by Julesz & Chang (1976). Notice however that this algorithm requires the existence of pulling only across pools and not within pools (in the terminology of Julesz & Chang 1976, p. 119).

Disparities larger than $w$ can be examined in very sparse images. If, for example, both primary pools (covering a disparity range of $\pm w$) are silent, detectors operating outside this range, possibly with a broad tuning curve, may be consulted. In a biologically plausible implementation, these detectors should be inhibited by activity in the primary pools (see figure 7). It is tempting to suggest

that detectors for these outlying disparities (i.e. exceeding about $\pm w$) may give rise to depth sensations and eye movement control in diplopic conditions.

If the image is not sparse, and the disparity exceeds the operating range, both algorithms will fail. Can the failure be recognized simply at this low level?

For the first algorithm, no correct match will be possible in the range $\pm \frac{1}{2}w$. The probability of a random match in this range is about 0.4, i.e. significantly less than 1.0. When the disparity between the two images lies in the range $\pm \frac{1}{2}w$, there will *always* be at least one match. It is therefore relatively easy to discriminate between these two situations.

For the second algorithm, an analogous argument applies; in this case the probability of no candidate match is about 0.3 for image disparities lying outside the range $\pm w$, and zero for disparities lying within it. Again, it is relatively easy to discriminate between the situations.

Finally, W. E. L. Grimson (personal communication) has pointed out that matching can be carried out from either image or from both. Observe for example in figure 6*c*, that if matching is initiated from the left image, the match for $L_1$ is ambiguous, but for $L_2$ it is unambiguous. Similarly from the right image.

It seems most sensible to initiate matching simultaneously from both images. Then, before any 'pulling', there are three possible outcomes. (1) The matching of an element starting from both images is unambiguous, in which case the two must agree. (2) Matching from one image is ambiguous, but from the other it is not. In this case, the unambiguous match should be chosen. (3) Matching from both images is ambiguous, in which case they must be resolved by pulling from unambiguous neighbours.

*Implications for psychophysical measurements of Panum's fusional area*

Using the second of the above algorithms, matches may be assigned correctly for a disparity range $\pm w$. The precision of the disparity values thus obtained should be quite high, and a roughly constant proportion of $w$ (which one can estimate from stereoacuity results at about $\frac{1}{20}w$). For foveal channels, this means $\pm 3'$ disparity with resolution $10''$ for the smallest, and $\pm 12'$ (perhaps up to $\pm 20'$ if Wilson & Bergen (1979) holds for stereopsis) with resolution $40''$ for the largest ones. At 4° eccentricity, the range is $\pm 5.3'$ to about $\pm 34'$. We assume that this range corresponds to stereoscopic fusion, and that outside it one enters diplopic conditions, in which disparity can be estimated only for relatively sparse images.

Under these assumptions, our predicted values apparently correspond quite well to available measures of the fusional limits without eye movements (see Mitchell 1966; Fender & Julesz 1967; Julesz & Chang 1976; and predictions 3–6 below).

## Dynamic memory storage: the $2\frac{1}{2}$-D sketch

According to our theory, once matches have been obtained using masks of a given size, they are represented in a temporary buffer. These matches also control vergence movements of the two eyes, thus allowing information from large masks to bring small masks into their range of correspondence.

The reasons for postulating the existence of a memory are of two kinds, those arising from general considerations about early visual processing, and those concerning the specific problem of stereopsis. A memory like the $2\frac{1}{2}$-D sketch (see figure 3) is computationally desirable on general grounds, because it provides a representation in which information obtained from several early visual processes can be combined (Marr 1977; § 3.6 and table 1). The more particular reason associated specifically with stereopsis is the computational simplicity of the matching process, which requires a buffer in which to preserve its results as (1) disjunctive eye movements change the plane of fixation, and (2) objects move in the visual field. In this way, the $2\frac{1}{2}$-D sketch becomes the place where 'global' stereopsis is actually achieved, combining the matches provided independently by the different channels and making the resulting disparity map available to other visual processes.

### *The nature of the memory*

The $2\frac{1}{2}$-D sketch is a dynamic memory with considerable intrinsic computing power. It belongs to early visual processing, and cannot be influenced directly from higher levels, for example via verbal instructions, *a priori* knowledge or even previous visual experience.

One would however expect a number of constraints derived from the physical world to be embedded in its internal structure. For example, the rule R2 stated early in this article, that disparity changes smoothly almost everywhere, might be implemented in the $2\frac{1}{2}$-D sketch by connections similar to those that implement it in Marr & Poggio's (1976) cooperative algorithm (figure 2*c*). This active rule in the memory may be responsible for the sensation of a continuous surface to which even a sparse stereogram can give rise (Julesz 1971; fig. 4.4–5).

Another constraint is, for example, the continuity of discontinuities in the visible surfaces, which we believe underlies the phenomenon of subjective contours (Marr 1977, § 3.6). It is possible that even more complicated consistency relations, concerning the possible arrangements of surfaces in three dimensional space, are realized by computations in the memory (e.g. constraints in the spirit of those made explicit by Waltz 1975). Such constraints may eventually form the basis for an understanding of phenomena like the Necker-cube reversal.

From this point of view, it is natural that many illusions concerning the interpretation of three dimensional structure (the Necker cube, subjective contours, the Muller–Lyer figure, the Poggendorff figure, etc., Julesz 1971, Blomfield 1973) should take place after stereoscopic fusion.

According to this theory, the memory roughly preserves depth (or disparity) information during the scanning of a scene with disjunctive eye movements, and during movement of viewed objects. Information management will have limitations both in depth and in time, and the main questions here are over what range of disparities can the $2\frac{1}{2}$-D sketch maintain a record of a match in the presence of incoming information, and how long can it do this in its absence? The temporal question is less interesting because the purpose of the buffer is to organize incoming perceptual information, not to preserve it when there is none.

The spatial aspects of the $2\frac{1}{2}$-D sketch raise a number of interesting questions. First, are the maximal disparities that are preserved by the memory in stabilized image conditions the same as the maximum range of disparities that are simultaneously visible in a random dot stereogram under normal viewing conditions? Secondly, does the distribution of the disparities that are present in a scene affect the range that the memory can store? For example, is the range greater for a stereogram of a spiral, in which disparity changes smoothly, than in a simple square-and-surround stereogram of similar overall disparity?

For the first question, the available evidence seems to indicate that the range is the same in the two cases. According to Fender & Julesz (1967), the range is about 2° for a random dot stereogram. When the complex stereograms given by Julesz (1971, e.g. 4.5–3) are viewed from about 20 cm, they give rise to disparities of about the same order. If this were true, it would imply that the maximal range of simultaneously perceivable disparities is a property of the $2\frac{1}{2}$-D sketch alone, and is independent of eye movements.

With regard to the second question, it seems at present unlikely that the maximum range of simultaneously perceivable disparities is much affected by their distribution. It can be shown that the figure of about 2°, which holds for stabilized image conditions and for freely viewed stereograms with continuously varying disparities, also applies to stereograms with a single disparity.

Perception times do however depend on the distribution of disparities in a scene (Frisby & Clatworthy 1975; Saye & Frisby 1975). A stereogram of a spiral staircase ascending towards the viewer did not produce the long perception times associated with a two planar stereogram of similar disparity range. This is to be expected, within the framework of our theory, because of the way in which we propose vergence movements are controlled. We now turn to this topic.

## Vergence movements

Disjunctive eye movements, which change the plane of fixation of the two eyes, are independent of conjunctive eye movements (Rashbass & Westheimer 1961*b*), are smooth rather than saccadic, have a reaction time of about 160 ms, and follow a rather simple control strategy. The (asymptotic) velocity of eye vergence depends linearly on the amplitude of the disparity, the constant of proportionality being about 8°/s per degree of disparity (Rashbass & Westheimer 1961*a*). Vergence

movements are accurate to within about 2′ (Riggs & Niehl 1960), and voluntary binocular saccades preserve vergence nearly exactly (Williams & Fender 1977). Furthermore, Westheimer & Mitchell (1969) found that tachistoscopic presentation of disparate images led to the initiation of an appropriate vergence movement, but not to its completion. These data strongly suggest that the control of vergence movements is continuous rather than ballistic.

Our hypothesis is, that vergence movements are accurately controlled by matches obtained through the various channels, acting either directly or indirectly through the $2\frac{1}{2}$-D sketch. This hypothesis is consistent with the observed strategy and precision of vergence control, and also accounts for the findings of Saye & Frisby (1975). Scenes like the spiral staircase, in which disparity changes smoothly, allow vergence movements to scan a large disparity range under the continuous control of the outputs of even the smallest masks. On the other hand, two-planar stereograms with the same disparity range require a large vergence shift, but provide no accurate information for its continuous control. The long perception times for such stereograms may therefore be explained in terms of a random-walk-like search strategy by the vergence control system. In other words, guidance of vergence movements is a simple continuous closed loop process (cf. Richards 1975) which is usually inaccessible from higher levels.

There may exist some simple learning ability in the vergence control system. There is some evidence that an observer can learn to make an efficient series of vergence movements (Frisby & Clatworthy 1975). This learning effect seems however to be confined to the type of information used by the closed loop vergence control system. *A priori*, verbal or high level cues about the stereogram are ineffective.

## Experiments

In this section, we summarize the experiments that are important for the theory. We separate psychophysical experiments from neurophysiological ones, and divide the experiments themselves into two categories according to whether their results are critical and are already available (A), or are critical and not available and therefore amount to predictions (P). In the case of experimental predictions, we make explicit their importance to the theory by a system of stars; three stars indicates a prediction which, if falsified, would disprove the theory. One star indicates a prediction whose disproof remnants of the theory could survive.

### *Computation*

The algorithm we have described has been implemented, and is apparently reliable at solving the matching problem for stereo pairs of natural images (Grimson & Marr 1979). It depends on the uniqueness and continuity conditions formulated at the beginning of this article, and it is perhaps of some interest to see exactly how.

The continuity assumption is used in two ways. First, vergence movements

driven by the larger masks are assumed to bring the smaller masks into register over a *neighbourhood* of the match obtained through the larger masks. Secondly, local matching ambiguities are resolved by consulting the sign of *nearby* unambiguous matches.

The uniqueness assumption is used in quite a strong way. If a match found from one image is unique, it is assigned without further checking. This is permissible only because the uniqueness assumption is based on true properties of the physical world. If the algorithm is presented with a stereo pair in which the uniqueness assumption is violated, as it is in Panum's limiting case, the algorithm will assign a match that is unique from one image but not from the other (O. J. Braddick, in preparation).

### *Psychophysics*

1 (A, P**). Independent spatial-frequency-tuned channels are known to exist in binocular fusion and rivalry. The theory identifies these with the channels described from monocular experiments (Julesz & Miller 1975; Mayhew & Frisby 1976; Frisby & Mayhew 1979; Wilson & Giese 1977; Cowan 1977; Wilson 1978*a*, *b*; Wilson, Phillips, Rentschler & Hilz 1979; Wilson & Bergen 1979; and Felton, Richards & Smith 1972).

2 (P***). Terminations, and signed, roughly oriented zero-crossings in the filtered image are used as the input to the matching process.

3 (P**). In the absence of eye movements, discrimination between two disparities in a random dot stereogram is only possible within the range $\pm w$, where $w$ is the width associated with the largest active channel. Stereo acuity should scale with the width $w$ of the smallest active matched channels (i.e. about 10″ for the smallest and 40″ for the largest foveal channels).

4 (P***). In the absence of eye movements, the magnitude of perceived depth in non-diplopic conditions is limited by the lowest spatial-frequency channel stimulated.

5 (P***). In the absence of eye movements, the minimum fusable disparity range (Panum's fusional area) is $\pm 3.1'$ in the fovea, and $\pm 5.3'$ at 4° eccentricity. This requires that only the smallest channels be active.

6 (P***). In the absence of eye movements, the maximum fusable disparity range is $\pm 12'$ (possibly up to $\pm 20'$) in the fovea, and about $\pm 34'$ at 4° eccentricity. This requires that the largest channels be active, for example by using bars or other large bandwidth stimuli.

*Comments.* (1) Mitchell (1966) used small flashed line targets and found, in keeping with earlier studies, that the maximum amount of convergent or divergent disparity without diplopia is 10–14′ in the fovea, and about 30′ at 5° eccentricity. The extent of the so-called Panum fusional area is therefore twice this.

Under stabilized image conditions, Fender & Julesz (1967) found that fusion occurred between line targets (13′ by 1° high) at a maximum

disparity of 40′. This value probably represents the whole extent of Panum's fusional area. Using the same technique on a random dot stereogram, Fender & Julesz arrived at a figure of 14′ (6′ displacement and 8′ disparity within the stereogram). Since the dot size was only 2′, one may expect more energy in the high frequency channels than in the low, which would tend to reduce the fusional area. Julesz & Chang (1976), using a 6′ dot size over a visual angle of 5°, routinely achieved fusion up to $\pm$ 18′ disparity. Taking all factors into account, these figures seem to be consistent with our expectations.

(2) Prediction 6 should hold for dynamic stereograms with the following caveats. First, motion cues must be eliminated. Secondly nonlinear temporal summation between frames at the receptor level may introduce unwanted low spatial-frequency components in the two images.

7 (P**). In the absence of eye movements, the perception of rivalrous random dot stereograms is subject to certain limitations. For example, for images of sufficiently high quality, fig. 2*b* of Mayhew & Frisby (1976) should give rise to depth sensations, but fig. 2*c* should not. In the presence of eye movements, fig. 2*c* gives a sensation of depth. This could be explained if vergence eye movements can be driven by the relative imbalance between the numbers of unambiguous matches in the crossed and uncrossed pools over a small neighbourhood of the fixation point.

8 (A). As measured by disparity specific adaptation effects, the optimum stimulus for a small disparity is a high spatial frequency grating, whereas for large disparities, the most effective stimulus is a low spatial frequency grating. Furthermore, the adaptation effect specific to disparity is greatest for gratings whose periods are twice the disparity (Felton, Richards & Smith 1972). (In our terms, in fact, $\lambda$ is approximately $2.2w$ where $\lambda$ is the centre frequency of the channel.)

9 (A). Evidence for the two pools hypothesis (Richards 1970, 1971; Richards & Regan 1973) is consistent with the minimal requirement for the second of the matching algorithms we described (figures 6 and 7).

10 (P***). In the absence of eye movements, the perception of tilt in stereoscopically viewed grating pairs of different spatial frequencies is limited by 4, 5, and 6 above.

11 (A). Individuals impaired in one of the two disparity pools show corresponding reductions in depth sensations accompanied by a loss of vergence movements in the corresponding direction (Jones 1972).

12 (P*). Outside Panum's area, the dependence of depth sensation on disparity should be roughly proportional to the initial vergence velocity under the same conditions.

13 (P***). For a novel two planar stereogram, vergence movements should exhibit a random-search-like structure. The three star status holds when the disparity range exceeds the size of the largest masks activated by the pattern.

14 (P***). The range of vergence movements made during the successful and precise interpretation of complex, high frequency, multi layer, random dot stereograms should span the range of disparities.

15 (P*). Perception times for a random dot stereogram portraying two small planar targets separated laterally and in depth, against an uncorrelated background, should be longer than the two planar case (13). Once found, their representation in the memory should be labile if an important aspect of the representation there consists of local disparity differences.

*Neurophysiology*

16 (*partly* A). At each point in the visual field, the scatter of bar mask receptive field sizes is about 4 : 1 (Hubel & Wiesel 1974, figs. 1 and 4; Wilson & Gieze 1977, p. 27). More data are however needed on this point. This range is spanned by four populations of receptive field size.

17 (P**). There exist binocularly driven cells sensitive to disparity. A given cell signals a match between either a zero-crossing pair or a termination pair, both items in its pair having the same sign, size and rough orientation.

18 (P**). Each of the populations defined by (17) is divided into at least two main disparity pools, tuned to crossed and uncrossed disparities respectively, with sensitivity curves extending outwards to a disparity of about the width of its corresponding receptive field centre (see figure 7). Being sensitive to pure disparity, these cells are sensitive to changes in disparity induced by vergence movements. In addition, there may be units quite sharply tuned to near-zero disparities.

19 (P*). In addition to the basic disparity pools of (18), there may exist cells tuned to more outlying (diplopic) disparities (compare figure 7). These cells should be inhibited by any activity in the basic pools (cf. Foley, Applebaum & Richards 1975).

20 (P**). There exists a neural representation of the $2\frac{1}{2}$-D sketch. This includes cells that are highly specific for some monotonic function of depth and disparity, and which span a depth range corresponding to about 2° of disparity. Within a certain range, these cells may not be sensitive to disjunctive eye movements. This corresponds to the notion that the plane of fixation can be moved around within the 2° disparity range currently being represented in the $2\frac{1}{2}$-D sketch.

21 (P*). The diplopic disparity cells of (20) are especially concerned with the control of disjunctive eye movements.

*Comments.* Because of the computational nature of this approach, we have been able to be quite precise about the nature of the processes that are involved in this theory. Since a process may in general be implemented in several different ways, our physiological predictions are more speculative than our psychophysical ones. They should perhaps be regarded as guidelines for investigation rather than as necessary consequences of the theory.

Unfortunately, the technical problems associated with the neuro-

physiology of stereopsis are considerable, and rather little quantitative data is currently available. Since Barlow, Blakemore & Pettigrew's (1967) original paper, relatively few examples of disparity tuning curve have been published (see for example, Pettigrew, Nikara & Bishop 1968; Bishop, Henry & Smith 1971; Nelson, Kato & Bishop 1977). Recently however, Poggio & Fischer (1978, in the monkey), and von der Heydt, Adorjani, Hanny & Baumgartner (1978, in the cat) have published properly controlled disparity tuning curves. On the whole, these studies (see also Clarke, Donaldson & Whitteridge 1976) favour the pools idea (see prediction 18).

## Discussion

Perhaps one of the most striking features of our theory is the way it returns to Fender & Julesz's (1967) original suggestion, of a cortical memory that accounts for the hysteresis and which is distinct from the matching process. Consequently fusion does not need to be cooperative, and our theory and its implementation (Grimson & Marr 1979) demonstrate that the computational problem of stereoscopic matching can be solved without cooperativity. These arguments do *not* however forbid its presence. Critical for this question are the predictions about the exact extent of Panum's fusional area for each channel. If the empirical data indicate a fusable disparity range significantly larger than $\pm w$, false targets will pose a problem not easily overcome using straightforward matching techniques like algorithm (2) of figure 6. In these circumstances, the matching problem could be solved by an algorithm like Marr & Poggio's (1976) operating within each channel, to eliminate possible false targets arising as a result of an extended disparity sensitivity range.

As it stands, there are a number of points on which the theory is indefinite, especially concerning the $2\frac{1}{2}$-D sketch. For example:

(1) What is its exact structure, and how are the various constraints implemented there?

(2) What is the relationship between the spatial structure of the information written in the memory and the scanning strategy of disjunctive and conjunctive eye movements?

(3) Is information moved around in the $2\frac{1}{2}$-D sketch during disjunctive or conjunctive eye movements, and if so, how? For example, does the current fixation point always correspond to the same point in the $2\frac{1}{2}$-D sketch?

Finally, we feel that an important feature of this theory is that it grew from an analysis of the computational problems that underlie stereopsis, and is devoted to a characterization of the processes capable of solving it without specific reference to the machinery in which they run. The elucidation of the precise neural mechanisms that implement these processes, obfuscated as they must inevitably be by the vagaries of natural evolution, poses a fascinating challenge to classical techniques in the brain sciences.

We are deeply indebted to Whitman Richards for many remarks that we understand only in retrospect. We are especially grateful to Jack Cowan, John Frisby, Eric Grimson, David Hubel, Bela Julesz, John Mayhew and Hugh Wilson, and to Werner Reichardt and the Max Planck Society for their kind hospitality in Tübingen. Karen Prendergast prepared the illustrations. The Royal Society kindly gave permission for reproduction of figure 3, and *Science* and the American Association for the Advancement of Science for figure 2. This work was conducted at the Max-Planck-Institut für Biologische Kybernetik in Tübingen, and at the Artificial Intelligence Laboratory, a Massachusetts Institute of Technology research program supported in part by the Advanced Research Projects Agency of the Department of Defense, and monitored by the Office of Naval Research under contract number N00014-75-C-0643. D.M. was partly supported by NSF contract number 77-07569-MCS.

## References

Barlow, H. B., Blakemore, C. & Pettigrew, J. D. 1967 The neural mechanism of binocular depth discrimination. *J. Physiol., Lond.* **193**, 327–342.

Bishop, P. Q., Henry, G. H. & Smith, C. J. 1971 Binocular interaction fields of single units in the cat striate cortex. *J. Physiol., Lond.* **216**, 39–68.

Blakemore, C. & Campbell, F. W. 1969 On the existence of neurons in the human visual system selectively sensitive to the orientation and size of retinal images. *J. Physiol., Lond.* **203**, 237–260.

Blomfield, S. 1973 Implicit features and stereoscopy. *Nature, new Biol.* **245**, 256.

Campbell, F. W. & Robson, J. 1968 Application of Fourier analysis to the visibility of gratings. *J. Physiol., Lond.* **197**, 551–566.

Clarke, P. G. H., Donaldson, I. M. L. & Whitteridge, D. 1976 Binocular visual mechanisms in cortical areas I and II of the sheep. *J. Physiol., Lond.* **256**, 509–526.

Cowan, J. D. 1977 Some remarks on channel bandwidths for visual contrast detection. *Neurosci. Res. Progr. Bull.* **15**, 492–517.

Dev, P. 1975 Perception of depth surfaces in random-dot stereograms: a neural model. *Int. J. Man–Machine Stud.* **7**, 511–528.

Felton, T. B., Richards, W. & Smith, R. A. Jr. 1972 Disparity processing of spatial frequencies in man. *J. Physiol., Lond.* **225**, 349–362.

Fender, D. & Julesz, B. 1967 Extension of Panum's fusional area in binocularly stabilized vision. *J. opt. Soc. Am.* **57**, 819–830.

Foley, J. M., Applebaum, T. H. & Richards, W. A. 1975 Stereopsis with large disparities: discrimination and depth magnitude. *Vision Res.* **15**, 417–422.

Frisby, J. P. & Clatworthy, J. L. 1975 Learning to see complex random-dot stereograms. *Perception* **4**, 173–178.

Frisby, J. P. & Mayhew, J. E. W. 1979 Spatial frequency selective masking and stereopsis. (In preparation.)

Georgeson, M. A. & Sullivan, G. D. 1975 Contrast constancy: deblurring in human vision by spatial frequency channels. *J. Physiol., Lond.* **252**, 627–656.

Grimson, W. E. L. & Marr, D. 1979 A computer implementation of a theory of human stereo vision. (In preparation.)

von der Heydt, R., Adorjani, Cs., Hanny, P. & Baumgartner, G. 1978 Disparity sensitivity and receptive field incongruity of units in the cat striate cortex. *Exp. Brain Res.* **31**, 523–545.

Hines, M. 1976 Line spread function variation near the fovea. *Vision Res.* **16**, 567–572.

Hirai, Y. & Fukushima, K. 1976 An inference upon the neural network finding binocular correspondence. *Trans. IECE* J59-D, 133–140.

Hubel, D. H. & Wiesel, T. N. 1974 Sequence regularity and geometry of orientation columns in monkey striate cortex. *J. comp. Neurol.* **158**, 267–294.

Jones, R. 1972 Psychophysical and oculomotor responses of manual and stereoanomalous observers to disparate retinal stimulation. Doctoral dissertation, Ohio State University. Dissertation Abstract N. 72-20970.

Julesz, B. 1960 Binocular depth perception of computer-generated patterns. *Bell System Tech. J.* **39**, 1125–1162.

Julesz, B. 1963 Towards the automation of binocular depth perception (AUTOMAP-1). *Proceedings of the IFIPS Congres, Munich* 1962 (ed. C. M. Popplewell). Amsterdam: North Holland.

Julesz, B. 1971 *Foundations of cyclopean perception.* The University of Chicago Press.

Julesz, B. & Chang, J. J. 1976 Interaction between pools of binocular disparity detectors tuned to different disparities. *Biol. Cybernetics* **22**, 107–120.

Julesz, B. & Miller, J. E. 1975 Independent spatial-frequency-tuned channels in binocular fusion and rivalry. *Perception* **4**, 125–143.

Kaufman, L. 1964 On the nature of binocular disparity. *Am. J. Psychol.* **77**, 393–402.

Leadbetter, M. R. 1969 On the distributions of times between events in a stationary stream of events. *J. R. statist. Soc.* B **31**, 295–302.

Longuet-Higgins, M. S. 1962 The distribution of intervals between zeros of a stationary random function. *Phil. Trans. R. Soc. Lond.* A **254**, 557–599.

Marr, D. 1974 A note on the computation of binocular disparity in a symbolic, low-level visual processor. *M.I.T. A.I. Lab. Memo* 327.

Marr, D. 1976 Early processing of visual information. *Phil. Trans. R. Soc. Lond.* B **275**, 483–524.

Marr, D. 1977 Representing visual information. *AAAS 143rd Annual Meeting. Symposium on Some Mathematical Questions in Biology*, February. Published in *Lectures on mathematics in the life sciences* **10**, 101–180 (1978). Also available as *M.I.T. A.I. Lab. Memo* 415.

Marr, D. & Hildreth, E. 1979 Theory of edge detection. (In preparation.)

Marr, D. & Nishihara, H. K. 1978 Representation and recognition of the spatial organization of three-dimensional shapes. *Proc. R. Soc. Lond.* B **200**, 269–294.

Marr, D., Palm, G. & Poggio, T. 1978 Analysis of a cooperative stereo algorithm. *Biol. Cybernetics* **28**, 223–229.

Marr, D. & Poggio, T. 1976 Cooperative computation of stereo disparity. *Science, N.Y.* **194**, 283–287.

Marr, D. & Poggio, T. 1977*a* A theory of human stereo vision. *M.I.T. A.I. Lab. Memo* 451.

Marr, D. & Poggio, T. 1977*b* Theory of human stereopsis. *J. opt. Soc. Am.* **67**, 1400.

Mayhew, J. E. W. & Frisby, J. P. 1976 Rivalrous texture stereograms. *Nature, Lond.* **264**, 53–56.

Mitchell, D. E. 1966 Retinal disparity and diplopia. *Vision Res.* **6**, 441–451.

Nelson, J. I. 1975 Globality and stereoscopic fusion in binocular vision. *J. theor. Biol.* **49**, 1–88.

Nelson, J. I., Kato, H. & Bishop, P. O. 1977 Discrimination of orientation and position disparities by binocularly activated neurons in cat striate cortex. *J. Neurophysiol.* **40**, 260–283.

Papoulis, A. 1968 *Systems and transforms with applications in optics.* New York: McGraw Hill.

Pettigrew, J. D., Nikara, T. & Bishop, P. O. 1968 Binocular interaction on single units in cat striate cortex: simultaneous stimulation by single moving slit with receptive fields in correspondence. *Exp. Brain Res.* **6**, 311–410.

Poggio, G. F. & Fischer, B. 1978 Binocular interaction and depth sensitivity of striate and prestriate cortical neurons of the behaving rhesus monkey. *J. Neurophysiol.* **40**, 1392–1405.

Rashbass, C. & Westheimer, G. 1961*a* Disjunctive eye movements. *J. Physiol., Lond.* **159**, 339–360.

Rashbass, C. & Westheimer, G. 1961*b* Independence of conjunctive and disjunctive eye movements. *J. Physiol., Lond.* **159**, 361–364.

Rice, S. O. 1945 Mathematical analysis of random noise. *Bell Syst. Tech. J.* **24**, 46–156.

Richards, W. 1970 Stereopsis and steroblindness. *Exp. Brain Res.* **10**, 380–388.

Richards, W. 1971 Anomalous stereoscopic depth perception. *J. opt. Soc. Am.* **61**, 410–414.

Richards, W. 1975 Visual space perception. In *Handbook of Perception*, vol. 5, *Seeing*, ch. 10, pp. 351–386 (ed E. C. Carterette & M. D. Freidman). New York: Academic Press.

Richards, W. A. 1977 Stereopsis with and without monocular cues. *Vision Res.* **17**, 967–969.

Richards, W. A. & Regan, D. 1973 A stereo field map with implications for disparity processing. *Invest. Ophthal.* **12**, 904–909.

Riggs, L. A. & Niehl, E. W. 1960 Eye movements recorded during convergence and divergence. *J. opt. Soc. Am.* **50**, 913–920.

Saye, A. & Frisby, J. P. 1975 The role of monocularly conspicuous features in facilitating stereopsis from random-dot stereograms. *Perception* **4**, 159–171.

Schiller, P. H., Finlay, B. L. & Volman, S. F. 1977 Quantitative studies of single-cell properties in monkey striate cortex. III. Spatial frequency. *J. Neurophysiol.* **39**, 1334–1351.

Sperling, G. 1970 Binocular vision: a physical and a neural theory. *Am. J. Psychol.* **83**, 461–534.

Sugie, N. & Suwa, M. 1977 A scheme for binocular depth perception suggested by neurophysiological evidence. *Biol. Cybernetics* **26**, 1–15.

Waltz, D. 1975 Understanding line drawings of scenes with shadows. In *The psychology of computer vision* (ed. P. H. Winston), pp. 19–91. New York: McGraw-Hill.

Westheimer, G. & Mitchell, D. E. 1969 The sensory stimulus for disjunctive eye movements. *Vision Res.* **9**, 749–755.

Williams, R. H. & Fender, D. H. 1977 The synchrony of binocular saccadic eye movements. *Vision Res.* **17**, 303–306.

Wilson, H. R. 1978*a* Quantitative characterization of two types of line spread function near the fovea. *Vision Res.* **18**, 971–981.

Wilson, H. R. 1978*b* Quantitative prediction of line spread function measurements: implications for channel bandwidths. *Vision Res.* **18**, 493–496.

Wilson, H. R. & Bergen, J. R. 1979 A four mechanism model for spatial vision. *Vision Res.* (in the press).

Wilson, H. R. & Giese, S. C. 1977 Threshold visibility of frequency gradient patterns. *Vision Res.* **17**, 1177–1190.

Wilson, H. R., Phillips, G., Rentschler, I. & Hilz, R. 1979 Spatial probability summation and disinhibition in psychophysically measured line spread functions. *Vision Res.* (in the press).

## Appendix. Statistical analysis of zero-crossings

We assume that $f(x) = \int I(x,y)\, h(y)\, \mathrm{d}y$, where $I(x,y)$ is the image intensity and $h(y)$ represents the longitudinal weighting function of the mask, is a white Gaussian process. Our problem is that of finding the distribution of the intervals between alternate zero-crossings by the stationary normal process obtained by filtering $f(x)$ through a linear (bandpass) filter.

Assume that there is a zero-crossing at the origin, and let $P_0(\xi)$, $P_1(\xi)$ be the probability densities of the distances to the first and second zero-crossings. $P_0$ and $P_1$ are approximated by the following formulae (Rice 1945, § 3.4; Longuet-Higgins 1962, eqns 1.2.1 and 1.2.3; Leadbetter 1969):

$$P_0(\xi) = \frac{1}{2\pi}\left[\frac{\psi(0)}{-\psi''(0)}\right]^{\frac{1}{2}} \frac{M_{23}(\xi)}{H(\xi)} (\psi^2(0) - \psi^2(\xi)) [1 + H(\xi) \operatorname{arccot}(-H(\xi))],$$

$$P_1(\xi) = \frac{1}{2\pi}\left[\frac{\psi(0)}{-\psi''(0)}\right]^{\frac{1}{2}} \frac{M_{23}(\xi)}{H(\xi)} (\psi^2(0) - \psi^2(\xi)) [1 - H(\xi) \operatorname{arccot}(H(\xi))],$$

where $\psi(\xi)$ is the autocorrelation of the underlying stochastic process, a prime denotes differentiation with respect to $\xi$, and also

$$H(\xi) = M_{23}(\xi)\,[M_{22}(\xi) - M_{23}(\xi)]^{-\frac{1}{2}},$$

$$M_{22}(\xi) = -\psi''(0)\,(\psi^2(0) - \psi^2(\xi)) - \psi(0)\,\psi'^2(\xi),$$

$$M_{23}(\xi) = \psi''(\xi)\,(\psi^2(0) - \psi^2(\xi)) + \psi(\xi)\,\psi'^2(\xi).$$

These approximations cease to be accurate for large values of $\xi$ (i.e. of order $\lambda$, where $2\pi/\lambda$ is the centre frequency of the channel; see Longuet-Higgins (1962) for a discussion of various approximations), where they overestimate $P_0$ and $P_1$. The autocorrelation $\psi(\xi)$ can be easily computed analytically for the two filters of figure 5.

**Ellen C. Hildreth and W. Eric L. Grimson**

Commentary on

# *Binocular Depth Perception*

David Marr's work on binocular stereopsis, which evolved through a fruitful collaboration with Tomaso Poggio, represents a landmark in the emerging field of Computational Vision. The culmination of this work, the model of human stereopsis that appeared in the Proceedings of the Royal Society, is special in many regards. The first is its attempt to incorporate and account for the large, diverse body of psychophysical and physiological observations available at the time, and the extensive predictions for future experimental work that derived from this model. The second is the strong theoretical foundation used to justify particular aspects of the model. The third is the innovation captured in the algorithm for stereo correspondence, which subsequently formed the basis of a successful computer stereo system that has been replicated many times over, both in research and in commercial systems. In this commentary, we first provide some historical perspective on the evolution of Marr's ideas on stereo vision, and then address the impact of this work on the study of stereo processing in biological and computer vision systems over the past decade.

Marr's first venture into the problem of stereo correspondence is captured in the MIT Artificial Intelligence Laboratory Memo, "A Note on the Computation of Binocular Disparity in a Symbolic, Low-Level Visual Processor." In a style that became a trademark of Marr's work, he emphasizes the importance of being very precise about the goals of the computation and about the constraints imposed on it by the structure of the three-dimensional world. Following this prescription, he first lays out the basic steps for calculating stereo disparity: (a) items corresponding to locations on a physical surface must be identified in one image, (b) the corresponding features must be identified in the second image, and (c) the relative positions of the corresponding features in the two images must be measured. He then makes a strong case that the features matched between the two images should not be the raw intensity measurements used in previous stereo models, but low-level symbolic tokens such as intensity edges, bars, terminations, and so on. Finally, he formally identifies three constraints on the matching process itself: uniqueness (the "use once" condition), continuity (the "suggestion interaction"), and maximization of some measure of goodness of fit. In this early paper, Marr presents only a few vague ideas about how one might embody such constraints in an algorithm, but it is valuable to see the top-down nature of Marr's thinking, which often evolved from a formal statement of a problem, through the identification of constraints, and finally to the formulation of an algorithm that embodies these constraints.

The final stage of formulating an algorithm for stereo correspondence emerged in the form of a cooperative network described in the paper by Marr and Poggio, "Cooperative Computation of Stereo Disparity." This first model was driven in part by the observation that the highly interactive organization of neural hardware suggests algorithms with a parallel structure, requiring many local operations on large data arrays, and in part by evidence from Julesz and his colleagues that human stereo fusion exhibited cooperative behavior. The network embodies a literal implementation of the constraints of uniqueness and continuity in an iterative process that operates on a "primal sketch" representation of the right and left images. With a view that Marr and Poggio must later retract, they boldly state, "non-iterative local operations cannot solve the stereo problem in a satisfactory way." Marr and Poggio admit that the analysis of the behavior of such cooperative algorithms is very difficult, although they later provide some analysis of its convergence properties for a limited class of random-dot stereograms (Marr et al., 1978).

The cooperative algorithm was interesting for its simplicity and for the clarity with which it embodied Marr's original statement of the goals of the stereo process and necessary physical constraints. A primary implication of the model for biological stereo processing, namely the existence of numerous sharply tuned binocular disparity detectors sensitive to a wide range of disparity values, was of course not born out through later physiological studies. In addition, the cooperative algorithm did not embody other key observations about human stereo vision, such as the need for vergence eye movements to fuse complex stereograms and the role of multiple spatial frequency tuned channels. The Marr-Poggio cooperative algorithm was in some ways a culmination of several influential models of the past, by Dev, Sperling, Julesz and others. The next stage in Marr's work on stereopsis, however, represented a radical change of view.

It was always a driving force in Marr's work to understand the particular nature of the visual processing underlying biological systems, and his ideas could change dramatically in the face of "challenging" experimental evidence. This drive led Marr and Poggio to develop their second algorithm described in the proceedings of the Royal Society paper, "A Computational Theory of Human Stereo Vision." The essence of this new model is the following. First, intensity edges are extracted from the result of filtering the left and right images with a range of second derivative filters of varying size. The matching of features at coarser scales provides rough disparity information over a broad range, while the matching of features at finer scales provides more accurate disparity measurements over a narrower range. When the disparities within a particular region of the image are outside the range of fusion for the finer channels, the coarser channels can initiate appropriate vergence eye movements to bring the left and right images within a range of fusion for the finer channels. If the disparities are outside the range of fusion for the coarsest channel, random eye movements are performed until portions of the image fall within the necessary range. Within all of the channels, disparity information is initially represented

only qualitatively, indicating whether potential matches can be found within two broad pools of crossed and uncrossed disparities and a single narrower pool around zero disparity. Mechanisms are also postulated for disambiguating multiple local disparities by examining the disparities assigned to neighboring features, both within the same channel and at coarser scales. Finally, a statistical analysis is presented that determines an appropriate choice for the range of disparities that can be detected as a function of filter size, and provides a criterion for determining whether corresponding regions of the two images are really within the needed range of disparity.

As we mentioned earlier, one of the special qualities of this second model is the diverse set of observations from psychophysical and physiological studies that it attempts to bring together. Among these are the general observations by Julesz and others concerning the role of multiple, independent spatial-frequency-tuned channels in stereo fusion, the critical role of vergence eye movements, the emerging psychophysical and physiological evidence for the existence of a small class of broadly tuned disparity sensitive mechanisms, and measurements of the limited size of Panum's fusional area. Early experimental observations that suggested a cooperative process in stereo fusion were instead accounted for by the existence of a permanent representation of depth (or surface orientation) called the 2 1/2D Sketch, into which the results of the stereo computation were placed.

With regard to further study of stereo processing in biological systems, another trademark of Marr's work is the formulation of specific predictions from his models for future experimentation in psychophysics and physiology. This contribution both reflects the power of the style of computational models that he pursued, as well as his special talent for relating theory and experiment. Marr and Poggio's Royal Society article ends with an extensive list of such predictions, many of which have been addressed in later experimental work (for review, see Poggio and Poggio, 1984; Mayhew and Frisby, 1981).

One focus of Marr and Poggio's predictions regarded the coupling between the range of disparity that can be fused and the spatial frequency range (or the size) of the underlying filters that are active. Subsequent experiments provide mixed support for these implications of the theory. There is some evidence that the overall extent of the range of fusable disparities scales with the spatial frequency content of the stimulus, but less support for an increase in the minimal disparity difference that can be detected as coarser channels alone become active (for example, Mayhew and Frisby, 1979; Schumer and Julesz, 1982; see also Poggio and Poggio, 1984). It appears that even coarser channels can represent disparity information with relatively high precision. With regard to disparity range, it is not necessary to assume the strict cutoff in range as a function of filter size that was embodied in the original theory. This range can be expanded within individual channels, at the expense of relying more critically on the disambiguation mechanisms. In support of this relationship is a physiological study in the cat by Ferster (1981) that indicates that there is a strong coupling between receptive field size and range of disparity tuning.

With regard to vergence eye movements, it is clear that the model must be expanded to incorporate a broader set of mechanisms that can initiate appropriate vergence (for review, see Poggio and Poggio, 1984). For example, large, correct vergence eye movements can be initiated when only high spatial frequency channels are presumed to be active (Mowforth et al., 1981), and the detection of the disparity of features such as texture boundaries can initiate these eye movements (Kidd et al., 1979).

Many recent models focus on particular aspects of human stereo processing not considered explicitly in Marr and Poggio's work, such as our ability to cope with transparency, the existence of a limit on the gradient of stereo disparity and the ordering constraint. These models may better reflect the way in which these particular factors enter into the stereo correspondence process. In our view, however, subsequent models do not surpass the second Marr-Poggio model in the broad range of stereo phenomena that they attempt to incorporate. It is clear that in detail, there is much evidence to question particular aspects of the theory (e.g., Mayhew and Frisby, 1981), but it is less clear to what extent these observations could be accounted for through modification of the original model versus fundamental changes.

Besides its impact on models of human stereopsis, Marr and Poggio's stereo theory also had a major impact on computer vision. In part this follows from Marr's paradigmatic view that designing and testing a specific algorithm for a computational theory is essential in the development of that theory. This step forces one to be explicit about all the details of the theory, and often provides feedback about inappropriate assumptions or overlooked cases in the theoretical development. Today, this view of the relevance of empirical testing is commonly accepted, but at the time of the development of Marr's theories, it was a minority view.

Marr and Poggio's second stereo model embodied a number of seminal ideas that have been influential in computer vision. Among these ideas are the use of multi-scale edge detectors, including filters with large support for image smoothing, and the use of a coarse-to-fine processing strategy to control the complexity of matching (an idea developed concurrently by Moravec). Today, many automated stereo systems use multi-resolution feature based algorithms. The second Marr-Poggio stereo model was formalized and expanded further by Grimson (1981, 1985). In its final form, it reflected a strong influence from other computational studies of stereo vision, such as those of Mayhew and Frisby (1981), which, among other things, suggested the use of a non-oriented filtering process before the extraction of matching features and the use of the so-called "figural continuity" constraint. The final algorithm that emerged remains one of the most successful computer stereo algorithms currently in use. It has been extensively tested on a broad range of synthetic and natural stereo imagery, has been replicated by many researchers and formed the basis for successful commercial stereo systems that perform tasks such as automated stereo cartography.

To summarize, Marr's work on stereo vision is among his most lasting

and influential pieces of work, having a strong impact both on the course of computational work in stereo and on experimental work in psychophysics and physiology. It highlights his formal approach to the study of visual processing detailed in his book, Vision, and gives us a particularly clear look into the evolution of his thinking.

## REFERENCES

Ferster D (1981): A comparison of binocular depth mechanisms in areas 17 and 18 of the cat visual cortex. *J Physiol* 311:623-655

Grimson WEL (1981): *From Images to Surfaces: A Computational Study of the Human Early Visual System.* Cambridge, MA: MIT Press

Grimson WEL (1985): Computational experiments with a feature based stereo algorithm. *IEEE Trans Patt Anal Machine Intell PAMI* 7:17-34

Kidd AL, Mayhew JEW, Frisby JP (1979): Texture contours can facilitate stereopsis by initiating vergence eye movements. *Nature* 280:829-832

Marr D, Palm G, Poggio T (1978): Analysis of a cooperative stereo algorithm. *Biol Cybern* 28:223-239

Mayhew JEW, Frisby JP (1979): Convergent disparity discriminations in narrow-band-filtered random-dot stereograms. *Vision Res* 19:63-71

Mayhew JEW, Frisby JP (1981): Psychophysical and computational studies toward a theory of human stereopsis. *Artif Intell* 16:349-385

Mowforth P, Mayhew JEW, Frisby JP (1981): Vergence eye movements made in response to spatial-frequency-filtered random-dot stereograms. *Perception* 10:299-304

Poggio GF, Poggio T (1984): The analysis of stereopsis. *Ann Rev Neurosci* 7:379-412

Schumer RA, Julesz B (1982): Disparity limits in bandpass random-grating stereograms. *Invest Ophthal Visual Sci* 22 (Suppl) 272

*Ellen C. Hildreth*
*Associate Professor*
*Brain and Cognitive Sciences*
*Artificial Intelligence Laboratory*
*Co-Director, Center for*
*Biological Information Processing*
*Massachusetts Institute of Technology*
*Cambridge, Massachusetts*

*W. Eric L. Grimson*
*Associate Professor*
*Electrical Engineering and*
*Computer Sciences*
*Artificial Intelligence Laboratory*
*Massachusetts Institute of*
*Technology*
*Cambridge, Massachusetts*

# David Marr: A Pioneer in Computational Neuroscience

**Terrence J. Sejnowski**

# *David Marr: A Pioneer in Computational Neuroscience*

David Marr advocated and exemplified an approach to brain modeling that is based on computational sophistication together with a thorough knowledge of the biological facts. The pioneering papers in this collection demonstrate that a combination of computational analysis and biological constraints can lead to interesting neural algorithms. The recent developments in computational models of neural information processing systems is an extension of this seminal research: Marr has influenced the latest generation of network models through both his models and his emphasis on the computational level of analysis (Marr, 1975, 1982). Progress has been made by adopting an integrated approach in which constraints from all three of Marr's levels of analysis—the computational, algorithmic and the implementational—are applied at many different levels of investigation (Sejnowski and Churchland, 1989).

These early papers are not easy to read. Marr gives the reader too many concrete details and too little overall guidance. He demands of the reader a deep understanding of probability theory and an encyclopedic knowledge of neuroanatomy. Even those who are steeped in the current generation of neural network modeling will find terms in the early papers difficult to translate into recent usage. Still, they are reminiscent of the style found in Maxwell's papers, which were written before the invention of vector notation. Just as Maxwell introduced specialized terminology and drew analogies with mechanical concepts such as gears and idler wheels, there are unusual terms in Marr's papers, such as "codons," borrowed from molecular genetics, and a novel set of concepts that must be mastered before the papers can be appreciated. Although some of the central ideas in these early papers are well known, there are important insights into neural computation that will handsomely repay the reader the effort taken to master the terminology.

The first four papers in this collection are from Marr's Cambridge period in the 1960s, among the earliest in his career, and articulate a remarkably ambitious theory of memory. These models were firmly based on what was then known about the structure of the brain. Even those who know that Marr had a model for motor learning in the cerebellum (Marr, 1969) may not be aware that his models of neocortex (Marr, 1970) and the hippocampus (Marr, 1971) were even more detailed than his cerebellar model. In a personal conversation with me in 1975, Marr singled out the neocortex paper as the one among these early papers that he was most proud of. The last set of papers on vision in this collection, are transitional and reflect a new, computationally-motivated approach to vision that guided his later research at MIT. At the same time they reflect an evident fascination with the structure of the brain, a fascination that Marr retained throughout his career.

The learning algorithm in the cerebellar model requires an external "teacher" to instruct the synapses and would today be called a supervised learning procedure. This type of error-correction learning has been extended to multiple layers of processing units (Rumelhart et al., 1986). The archicortex model uses a form of unsupervised learning that is based on competition among the output units. Similar unsupervised learning algorithms for neural network models have recently been studied for clustering data (Grossberg, 1976; Kohonen, 1984; Rumelhart and Zipser, 1985). They are "simple" memory systems that do not address the central issue of how the information is represented on the inputs and outputs.

In contrast to the simple memory models, the neocortical model is about category formation and discovery of high-order patterns in data based on unsupervised learning. Several recent models have been proposed to attack the problem of extracting information from sensory data (Hinton and Becker, 1990; Linsker, 1990). Marr's approach was different and depended on recruiting new neurons to represent high-order statistical structure or "concepts." Unfortunately, the computational resources available to Marr in the late 1960s were minimal, and one of the frustrations when reading these papers is the lack of numerical simulations. Promising algorithms do not always perform as expected when confronted with data from the real world; too many simplifications must be made so that the analysis is tractable. Until simulations of the neocortex model are performed we will not be able to assess its effectiveness.

The retina paper is an attempt to match a computational problem—finding the lightness of a surface—to anatomical substrates in the retina (Marr, 1974). This paper is transitional: Marr's subsequent papers in vision would emphasize more and more the computational aspects while relying less and less on anatomy. The predictions in this paper were not borne out by subsequent and more recent physiological recordings from single neurons make it likely that the locus of lightness and color constancy is in visual cortex (Zeki, 1983). Interestingly, recent algorithms based on neural network models are similar to those proposed by Marr (Hurlbert and Poggio, 1988; Land, 1986). Old ideas often come back in contexts that their originators might not even recognize.

Perhaps the best known papers in this collection are the models of binocular depth perception (Marr and Poggio, 1976, 1979). These models were appealing because they were based on plausible biological mechanisms, were constrained by psychophysical data and were tested by simulations on real-world data. In the first model, Marr and Poggio performed simulations to demonstrate the effectiveness of their algorithm on random-dot stereograms, first introduced by Bela Julesz (Julesz, 1971). The elegant simplicity of the network model was anticipated in earlier research (Dev, 1975; Nelson, 1975), but the interpretation of the constraints and the convincing demonstration made this a landmark paper. This approach to constraint satisfaction was an inspiration for connectionist-style models of visual computation (Ballard et al. 1983).

One of the remarkable properties of the Marr-Poggio stereo network was that it always converged. The stereo network is a highly nonlinear system of

equations and attempts to analyze the nonlinear equations came to the conclusion that they were as difficult to analyze as Conway's game of "life" (Marr et al. 1978). In 1982, John Hopfield pointed out that symmetric networks like the Marr-Poggio model were a special case because they possess an energy or Lyapunov function that guarantees convergence to a local energy minimum (Hopfield, 1982). The stereo network was designed in such a way that the local energy minima are solutions to the problem. It is also possible to design network models to handle transparent surfaces in random-dot stereograms (Qian and Sejnowski, 1988). Subsequent developments showed how even more difficult constraint satisfaction problems can be solved by globally minimizing the energy (Hopfield and Tank, 1986; Kienker et al., 1986; Poggio et al., 1988). Marr later felt that the time delays inherent in the relaxation of a network to a solution were unsatisfactory given the speed with which our visual system can interpret most images (Marr, 1982: see p. 107). The shortcomings of the first stereo model were addressed in the second model (Grimson, 1981; Marr and Poggio, 1979), which was much faster and took into account multi-resolution filters that could be applied to real images. However, there are other aspects of stereo vision that cannot be handled by this algorithm. The human visual system is even more clever than these early stereo algorithms (Poggio and Poggio, 1984).

David Marr continued to make major contributions to the study of vision. Those who have been influenced primarily by Marr's book on the computational approach to vision (Marr, 1982) may be surprised by the extraordinary attention given in this collection of papers to neuroanatomy. In rereading them, it is possible to put into perspective Marr's later work on the computational approach to vision. Although Marr made fewer appeals to detailed biological mechanisms in his later vision papers, there were still many examples of inspiration and confirmation of computational approaches from neuropsychological and psychophysical data and constraints from physiological measurements. The scientific style of the papers in this collection make it clear that these intrusions from the biological realm were not incidental.

One of the inevitable problems of building models and theories in neuroscience is that new facts about the brain are continually being discovered and old ideas are sometimes modified or discarded. The striking advances in neuroscience since these early papers are most evident in our present view of neurons. In 1970, dendrites were thought to be passive cables and ideas on synaptic mechanisms were based primarily on the neuromuscular junction. Today, dendrites are known to have voltage-dependent conductances that make them dynamical entities (Llinas, 1988); a gallery of channels and neurotransmitter receptors with a wide range of time scales allow neurons to burst and oscillate, and synapses to potentiate and habituate (Kandel et al., 1987). Marr's models need to be updated to take these new properties into account. However, the insight that a powerful computational system could be built from a sophisticated model of memory remains an exciting idea, and the goal of incorporating anatomical constraints into network models of vision is now being

actively pursued (Sejnowski et al., 1988).

Finally, how is one to reconcile the research direction implicit in these papers and the explicit statements found in Marr (1982) regarding the independence of the computational level from the implementation level? This principle, taken out of the context provided by Marr's research style, gives the misleading impression that constraints from the algorithmic and implementation levels found in biological systems are unnecessary. A remarkable feature of Marr's book is the degree to which biological considerations enter on almost every page in inspiring computational analysis, in choosing between algorithms and in providing the ultimate measure of success. Computational explanations for our visual and mental abilities eventually may be found, and seeking such explanations is essential—this was Marr's message. However, he was far from abandoning biological and psychological data in reaching this goal.

The performance of our perceptual and cognitive systems and the way that brains are organized provide essential constraints on possible computational explanations (Churchland and Sejnowski, 1988; Sejnowski et al., 1989). Neural circuits and how they function clearly inspired Marr and they continue to be rich sources of inspiration for many of us.

## REFERENCES

Ballard D, Hinton G, Sejnowski T (1983 ): Parallel visual computation. *Nature* 306:21-26

Churchland PS, Sejnowski TJ (1988): Perspectives on cognitive neuroscience. *Science*, 242:741-745

Dev P (1975): Perception of depth surfaces in random-dot stereograms: a neural model. *Int J Man-Machine Stud* 7:511-528

Grimson E (1981): *From Images to Surfaces.* Cambridge, MA: MIT Press

Grossberg S (1976): Adaptive pattern classification and universal recoding: I: Parallel development and coding of neural feature detectors. *Biol Cybern* 23: 121-134

Hinton GE, Becker S (1990): *An unsupervised learning procedure that discovers surfaces in random-dot stereograms.* New Jersey: Lawrence Erlbaum Associates, pp 218-222

Hopfield JJ, (1982): Neural networks and physical systems with emergent collective computational abilities. *Natl Acad Sci USA* 79: 2554-2558

Hopfield JJ, Tank DW (1986): Computing with neural circuits: A model. *Science* 233: 625-633

Hurlbert AC, Poggio TA (1988): Synthesizing a color algorithm from examples. *Science* 239: 482-485

Julesz B (1971): *Foundations of Cyclopean Vision.* Chicago: University of Chicago Press

Kandel ER, Klein M, Hochner B, Shuster M, Siegelbaum SA, Hawkins RD, Glanzman DL, Castellucci VF (1987): Synaptic modulation and learning: new insights into synaptic transmission from the study of behavior. In *Synaptic Function*, Edelman GM, Gall WE, Cowan WM, eds, New York: John Wiley and Sons, pp 471-518

Kienker PK, Sejnowski TJ, Hinton GE, Schumacher LE (1986): Separating figure from ground with a parallel network. *Perception* 15:197-216

Kohonen T (1984): *Self-Organization and Associative Memory.* New York: Springer Verlag

Land EH (1986): An alternative technique for the computation of the designator in the retinex theory of color vision. *Proc Natl Acad Sci USA* 83:3078-3080

Linsker R (1990): Perceptual neural organization: some approaches based on network models and information theory. *Annu Rev Neurosci* 13:257-281

Llinas RR (1988): The intrinsic electrophysiological properties of mammalian neurons: Insights into central nervous system function. *Science* 242:1654-1664

Marr D (1969): A theory of cerebellar cortex. *J Physiol Lond* 202:437-470

Marr D (1970): A theory for cerebral neocortex. *Proc R Soc Lond, B* 176:161-234

Marr D (1971): Simple memory: a theory for archicortex. *Phil Trans Roy Soc B* 262: 23-81

Marr D (1974): The computation of lightness by the primate retina. *Science* 14:1377-1387

Marr D (1975): Approaches to biological information processing. *Science* 190: 875-876

Marr D (1982): *Vision.* San Francisco: Freeman

Marr D, Palm G, Poggio T (1978): Analysis of a cooperative stereo algorithm. *Biol Cybern* 28: 223-239

Marr D, Poggio T (1976): Cooperative computation of stereo disparity. *Science* 194::283-287

Marr D, Poggio T (1979): A computational theory of human stereo vision. *Proc R Soc Lond Ser. B* 204: 301-28

Nelson, JI (1975): Globality and stereoscopic fusion in binocular vision. *J Theor Biology* 49:1-88

Poggio G, Poggio T (1984): The analysis of stereopsis. *Annu Revs Neurosci* 7:379-412

Poggio T, Gamble EB, Little JJ (1988): Parallel integration of vision modules. *Nature* 242:436-439

Qian N, Sejnowski TJ (1988): *Learning to solve random-dot stereograms of dense transparent surfaces with recurrent backpropagation.* Pittsburgh, PA: Morgan-Kaufmann Publishers, pp 235-443

Rumelhart D, Zipser D (1985): Feature discovery by competitive learning. *Cogn Sci* 9:75-112

Rumelhart DE, Hinton GE, Williams RJ (1986): Learning internal representations by error propagation. In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations* Cambridge, MA: MIT Press

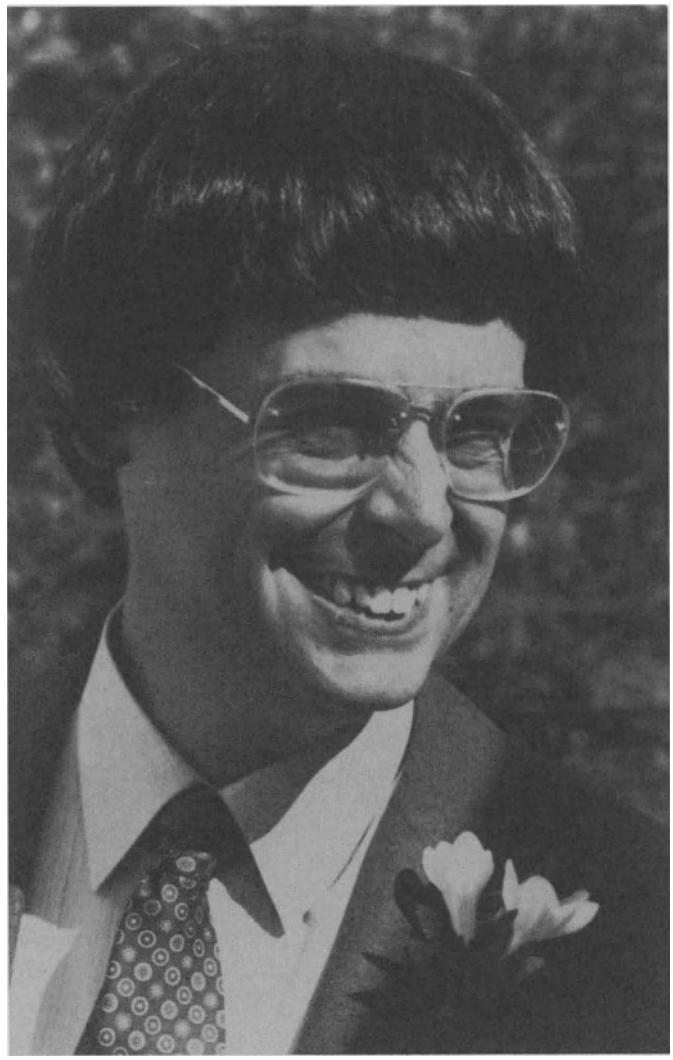
Sejnowski T, Koch C, Churchland P (1988): Computational neuroscience. *Science* 241:1299-1306

Sejnowski TJ, Churchland PS (1989): Brain and cognition. In: *Foundations of Cognitive Science*, Posner MJ, ed. Cambridge, MA: MIT Press

Zeki S (1983 ): Colour coding in the cerebral cortex: the reaction of cells in monkey visual cortex to wavelengths and colors. *Neuroscience* 9:741-765

*Professor*
*The Salk Institute*
*and the University of California, San Diego*
*La Jolla, California*

# Epilogue:
# Remembering David Marr



David Marr
1980 – Cambridge, Massachusetts

## Peter Rado

I was a student at Trinity with David and saw a lot of him until I left Cambridge in 1967. David was a great source of strength to me—I have been ill myself (with kidney problems) since 1958, and when I first was given a place on a kidney machine, David offered me one of his kidneys if it would be any use—it is very ironic to think that I have survived him; neither he nor I would ever have expected that (until recently).

David's thoughtful approach to things was one of the things that I valued a lot. He would never give a snap decision, but always thought carefully about everything. He was always very generous with his time and efforts: he took me camping when I was told that I should go away for my last ever holiday before being tied down to a kidney machine for the rest of my life. He immediately suggested we go camping, borrowed a tent and whisked me off, for a very memorable "last fling" in Scotland.

Another holiday together was in Germany—he borrowed a Land Rover and a group of us went to a Youth Music Festival at Bayreuth. I vividly remember David prowling around with his clarinet, looking for a soprano ("any soprano!") to sing Schubert's "Shepherd on the Rocks" with him. He was a superb musician, who seemed to be able to play the clarinet quite effortlessly—and without practice.

*November, 1980*

*Friend from college years.*

## Tony Pay

David was a person who brought excellence in one form or another to everything he did.

His excellence as a clarinet player was, to begin with, an excellence of attitude and response. He was always committed to what he did. He valued the immediate impulse as highly as considered action when he played, or when he discussed others' playing. Music sprang from the heart, and though he possessed a formidable intellect, he never tried to reduce this side of his life to anything theoretically manageable or systematizable. Indeed, his relationship with his instrument was a very physical one, and he obviously loved the act of performing. As sometimes happens, this physicality in one sense stood in the way of his further technical and expressive development, and it was a little time before he overcame the problem.

He once auditioned for a place in the University Orchestra, and was passed over (wrongly, we later thought!) in favor of another more obviously technically able instrumentalist. He told me afterward that he'd promised his supervisor that if he didn't play in the orchestra he would devote the time to a project involving a learning network. I never found out what happened to that particular project.

When we met again, much later, he had obviously developed further as a player and musician. He would sometimes say to his musician friends that when he was burnt out as a scientist he would take up the clarinet seriously. Few of us doubted that he would do rather well if such a moment came. The idea of performances being 'wrong' in some sense seemed not to be very important to him, and he was often excited by the possibilities revealed by other people's ideas. On the other hand, I remember him saying to me, "How could you do that to my slow movement?!" after a performance of the Brahms Clarinet Trio in which we had thrown away some of the more melting moments for the sake of showing the long phrases. But the remark was characteristically softened in his subsequent giggle. I also recall his glee at being included in the Orchestra to play Berlioz's 'Beatrice and Benedict'— "They think I'm good enough!"

Most of all, though, I remember his friendship, and his winning blend of seriousness and joy, and his laugh when there were no words to express what he felt. Though he said many memorable things, when there was nothing to say, David didn't say it.

*March, 1990*

*Friend from college years.*

*Principal Clarinetist*
*Royal Symphony Orchestra*
*London, England*

# G. S. Brindley

David gained first-class Honours in Part III of the Mathematical Tripos in June, 1966. In the same month he came to me saying that he wanted to do theoretical research on the brain. I advised him that he must spend one year of full-time study on what was known empirically about the brain. In the academic year of 1966-1967 he attended courses on anatomy, physiology and biochemistry, all new subjects to him.

In the summer of 1967 he began theoretical research. I was nominally his supervisor, but I gave him no ideas beyond a few that I had published (IBRO Bulletin 3 , 80 (1964) and Proc. Roy. Soc. B. 168, 361 (1967) or was preparing for publication (Proc. Roy. Soc. B. 174, 173, [1969]). David read and wrote, and then brought me drafts of long chapters, in which, after hard work, I managed to find a few minor errors. We met to discuss these minor errors, and I gave David one or two pep talks whose purpose was to persuade him that unless his work led to experimentally testable predictions whose prior probability was neither almost zero nor almost unity, no experimenter would read his work.

With these exceptions, our only contact was musical. I was a mediocre bassoonist, David a much better clarinetist. Inevitably we read through the three Beethoven duos and the Poulenc sonata for clarinet and bassoon. We even rehearsed one of the Beethoven duos and played it in a college concert. Once or twice we joined three other wind players, and read through some quintets.

In August 1968, after just 13 months of research, David submitted a dissertation for a Title A Fellowship of Trinity College, and was elected. This was a high honor; only those familiar with Trinity will know just how high.

I spent the fall quarter of 1968 in Berkeley, California, and migrated permanently to London, after 16 years of Cambridge, in December, 1968. David mqved, for that fall quarter, into the house that I had just bought for myself in London, and before my return from Berkeley he had rewritten as a paper for the Journal of Physiology the chapter on the cerebellum from his fellowship dissertation. It appeared in 1969 (Vol. 202, p. 437), and was the first substantial theoretical paper that the journal had ever published. David remained in London for 18 months after my return from Berkeley, at first full-time and then part-time, commuting from Cambridge. During these months, he wrote his paper on the classifying and memorizing functions of the cerebral

cortex, Proc. Roy. Soc. B 176, 161-234, the outline of which was already in his fellowship dissertation.

One reminiscence from this London period: Local amateur players performed Mozart's concerto for flute and harp. My six-year-old daughter, who had a remarkable musical memory and an excellent whistling technique, had heard recordings and some rehearsals and wanted to hear the performance. My wife and I were unavailable, and David offered to take her. She sat on his lap and, he reported, whistled the flute part, faultlessly and very quietly, throughout almost the whole performance. He didn't stop her, judging that it spoiled neither his enjoyment nor that of people in neighboring seats.

*March, 1990*

*David's doctoral thesis advisor.*

*Professor*
*Fellow of the Royal Society*
*Honorary Director*
*Medical Research Council*
*Neurological Prostheses Unit*
*Institute of Psychiatry*
*De Crespigny Park*
*London, England*

## Benjamin Kaminer

I never anticipated that the "workshop" I organized in May, 1972 would have been a turning point in the life of David Marr. And many might wonder why I, not being a neurobiologist, organized this workshop centered around David's ideas at the time, on the organization of the cerebral cortex. True, when I became Chairman of the Department of Physiology in 1970, I decided to develop a section of neurophysiology, but why gather a group of eminent molecular biologists, theoretical physicists, computer scientists and mathematicians together with classical neurobiologists? The time, I thought, appeared ripe for such interdisciplinary interaction and interchange in the search for an understanding of the human brain. As was well known, a number of molecular biologists, having cracked the genetic code, had challenged themselves with the task of unraveling the mysteries of the brain and mathematicians and computer scientists were designing networks and models creating "artificial intelligence." Such intellectual reasoning on the timeliness for a meeting of the minds, however, would not have been enough impetus for me to arrange it. Superimposed personal factors played a decisive role.

Sydney Brenner, Seymour Papert, and I were friends in South Africa. In the 1950s Sydney and I were in the Department of Physiology in the University of Witwatersrand, Johannesburg and Seymour, from the Department of Mathematics, became associated with our department, learning and experimenting on the nervous system as he developed his interests in psychology. In the latter part of that decade the three of us left South Africa. Sydney joined Francis Crick at the MRC, Cambridge, England. Seymour went to the Department of Mathematics, Cambridge, England for further study, then to France, and later joined Piaget in Switzerland. I went to work in Albert Szent-Gyorgyi's laboratory at the Marine Biological Laboratory, Woods Hole, in Massachusetts. Sydney and I remained in touch and we saw each other frequently in Woods Hole, but I had lost contact with Seymour Papert. When I took my current position in 1970, I heard that Seymour was at MIT, visited him at the Artificial Intelligence Laboratory and there met Marvin Minsky and learned of the inroads they were making in their field. I, of course, was aware of the progress Sydney was making on the nervous system of the nematode C. elegans and soon my idea emerged of getting these two "foreign" groups together with the "native" neurobiologists on common territory. When I discussed the proposition with Sydney he told me that David Marr had joined

their group and suggested that a good take-off point for discussion would be David Marr's recent papers on the cerebral cortex.

After a few more visits to MIT, I soon realized that many students in artificial intelligence, smart mathematicians, had never seen a synapse or a human brain and students in biology or neurobiology had little idea of the language of artificial intelligence. Since I decided to invite selected students from Harvard, MIT and our school to sit in at the workshop, I arranged for a series of lectures and demonstrations for these students to prepare them in advance. Marvin Minsky lectured on concepts in artificial intelligence to the biology students and the students from the Artificial Intelligence Lab learned about the ultrastructure and organization of the brain from Alan Peters and Walle Nauta and about basic neurophysiology from Ken Muller. The workshop was held at Boston University School of Medicine for three days on May 24-26, 1972. The main participants were David Marr, Sydney Brenner, Francis Crick and Stephen Blomfield (MRC, Cambridge); Freeman J. Dyson (Institute of Advanced Studies, Princeton); Seymour Papert, Marvin Minsky (Artificial Intelligence Lab, MIT); Stephen Kuffler, David Hubel, Torsten Wiesel, John Dowl- ing and Ken Muller (Harvard); Horace B. Barlow (University of California, Berkeley); Anthony Gorman and Alan Peters (Boston University); and I acted as chairman of the meeting. Together with 20 students, a total of 50 attended.

My first meeting with David Marr several days before the workshop, I shall never forget. His gentle manner, lovely smile, and joyful laughter remain as vivid perceptions in my brain. He agreed with my plan not to have a formal program except for introductory talks, first by David and then by Seymour Papert. I opened the first session with the briefest survey of past developments in neurobiology beginning with Cajal, and emphasized the main purpose of the workshop, to exchange ideas freely without any publication of the proceedings. David in his opening remarks applied the "inverse square law" to theoretical research, suggesting that its value varies inversely with the square of the generality (!) stressing the need to establish structure-functional relationships either from the bottom up, from structure to function, or top down, from function to structure. He developed his ideas by posing the questions: What does the brain do? What are the logical equivalents? What are the actual mechanisms?

Seymour Papert conceptualized on "artificial brains," "perceptrons" and analyzed a "simple system" involved in catching a piece of chalk. And for the rest of the three days we had lively and provocative discussion interspersed with mini-talks. During tea breaks and lunchtime, I would approach various individuals encouraging them to give short presentations. The main participants continued with informal discussion at a dinner in my house and after the workshop social interaction and intense intellectual discussion continued during the weekend, which was spent by some of the participants in Woods Hole. Albert Szent-Gyorgi invited us to his beach cottage, where he also exchanged

ideas, and we also went across to Martha's Vineyard by boat. Relaxation and fun did not hinder the intellectual vibrations in our brains.

During these five days David Marr spent much time with Minsky and Papert, and as I said, this event was a turning point that led to David subsequently joining the Artificial Intelligence Laboratory. That in itself was a very positive outcome of the workshop. Since the proceedings were not published, any other outcome is intangible. But let me include here two of the letters I received:

MRC Laboratory of Molecular Biology
9 June, 1972

Dear Bennie:

Thank you so much for such a marvelous few days in Boston. And your hospitality was quite wonderful. I was almost embarrassed by the trouble to which you went. The meeting itself was very useful for me. I learned a good deal from contact with the AI Laboratory and hopefully some good will come of it. I hope the neurophysiologists also gained, though I doubt if any will admit that they did. Many, many, thanks once again. It was really wonderful.

With very best wishes to you both,
David Marr

MRC Laboratory of Molecular Biology
31 May, 1972

Dear Bennie:

Many thanks for such a relaxed and informal meeting and for the very smooth organization behind it. I'm sorry I didn't come to Woods Hole but I felt really wretched and by the time I got to London my sore throat turned to a streaming cold. It seemed to me that the meeting really brought the various groups into intellectual speaking distance of each other—I hope the continuations at Woods Hole cemented the process. Of course I was really a spectator, but I think that for Sydney and David it was especially stimulating. Perhaps history will consider it seminal. If so, the credit will be yours. It was certainly much less of a strain than any meeting I've attended for years and that allows people to develop ideas at the back of their heads.

Best wishes,
Yours sincerely,
Francis
F.H.C. Crick

When David moved to MIT we kept in touch. He gave a wonderful seminar in our department and he visited me in Woods Hole in the summers. I became very fond of David and admired his intellectual prowess (coupled with modesty). He was well recognized and respected for his important scientific contributions, which continued during his brave struggle with a fatal condition. The ending of his life prematurely was a sad loss to all who knew, admired and loved him, and also to the science of neurobiology.

*March, 1990*

*Ben provided one side of the bridge for David's crossing the Atlantic. Sydney Brenner provided the other side.*

*Professor and Chairman*
*Department of Physiology*
*Boston University School of Medicine*
*Boston, Massachusetts*

## Francis H. Crick

I first heard of David under unusual circumstances. One day in the early 1960s I was standing about in the basement of the University Arms Hotel with a group of scientists, waiting to be photographed for a London newspaper. I had recently been dipping into several papers on theoretical neuroscience and I remarked to Alan Hodgkin that much of it seemed rather pretentious stuff to me. "I agree with you," he said, "but what about David Marr?" "Who," I said, "is David Marr?"

Sydney Brenner and I got in touch with David and one afternoon he came to talk to us in our joint office at the Molecular Biology Lab on Hills Road. David had been working on his theory of the cerebellum. At that time I didn't know the difference between a parallel fiber and a climbing fiber, so he had to explain it all from the beginning. This took several hours. When he left, Sydney and I were exhausted but undeniably impressed. A little later, when David was wondering where to go, we provided a home for him in a small office down the corridor.

I didn't see much of him during this period, though I remember struggling to follow several difficult seminars based on his cortical papers. Then came the meeting at the other Cambridge that brought together several neurophysiologists (Horace Barlow, David Hubel, etc.), members of the MIT AI group, together with David, Sydney and myself. The boat trip after the meeting was David's Road to Damascus. (I wasn't there, as I had a wretched cold.) He became converted to AI and before long moved to MIT.

I moved to the Salk Institute, in La Jolla in 1976 and decided to switch to the neurosciences. David by then had become close friends with Tomaso Poggio, and I was lucky enough to persuade them to visit me for a month. (Happy days! Who could spare a whole month now for such a visit?) It was April, 1979. The weather was perfect for the entire time. David and Tommy worked together during the morning. We all had lunch together and the three of us talked each day until tea-time.

It was very educational for me. They kept telephoning somebody called Westheimer, whom I'd never heard of. Apparently he did something called psychophysics, but what was that? I didn't entirely agree with David's rather functionalist approach. You can find echoes of our conversations in the last chapter of his book Vision. Most of the time David is arguing against me, though I also detect traces of Marvin Minsky in his imaginary antagonist. It

was a happy period in David's life. He thought he was cured of his leukemia. Only a few weeks after he left it recurred again.

David's work clearly falls into two phases. Marr I was concerned with neural circuitry and what it might compute. Marr II (the AI phase) was more functional. The emphasis was on the theory of the process and possible algorithms, with much less attention to realistic implementation. I believe that if he had lived he would have moved to a synthesis of these two approaches.

His early death was a great loss. He was trained as a mathematician and had outstanding intellectual powers. Most of his papers are not easy reading, as he was striving for precision and vigor, but behind it there was always a well thought out set of ideas. His book, much of which he wrote during his last illness, is written clearly and in a compelling style. It remains to be seen how much of his detailed work will survive, but his influence is everywhere. If he had lived, I have no doubt he would have come to dominate the field. His early death was a great loss, both for the subject and particularly for his close friends.

*March, 1990*

*Collaborator and friend, from Cambridge (UK) to Cambridge (USA)*

*Kieckhefer Professor*
*The Salk Institute*
*San Diego, California*

## Whitman Richards

I can't remember exactly when I first met David Marr. Perhaps the first time was when he and Tommy Poggio attended an NRP meeting on Neuronal Mechanisms in Visual Perception in December, 1973, where they spoke about "Levels of Understanding." I know for certain that this first brush made little impact. My own work on stereopsis was going very well then, with the discovery that stereo resembled color vision in having three distinct "channels" or "pools," any one of which could be absent in an observer, just as in color blindness. About that same time, or perhaps a year or so later, a squash friend from Cambridge asked what I thought of Marr's work. Somewhat embarrassed, I was motivated to review the cerebellar and archicortex papers, but again gained little profit because they were remote from my psychophysical studies. It would be some ten years later that I would return again to David's paper on the archicortex, and, with Aaron Bobick, see its relevance to perception. In the meantime, others, of course, had already come to appreciate Marr's early understanding of the cerebellum.

In 1975 David had targeted stereopsis as an information processing module for study. He needed someone to critique his ideas from a psychophysical viewpoint. I recall vividly our first real encounter, where he challenged me to explain my "three pool" model of stereopsis with sufficient precision that it could be run on a computer. This challenge struck home, for it then became quite clear to me that simple circuit or block diagrams were woefully inadequate. Left unspecified were a host of "details" such as the features to be matched, the hemisphere of eye that serves as the base representation, epipolar problems, false target elimination, etc. From that moment on, David had hooked me on the value of a computational approach to perception.

During those early years, my role in Marr's group was largely as a psychophysical bangboard. The most profitable exchanges were at the Newbury steak house just across the river in Boston, where we would walk for lunch. (Later, the original Legal's restaurant at Inman square became a favorite.) We usually loosened up with a gin and tonic, followed by a $2.00 steak and fries special, with coffee as an antidote for the appetizer. On a hot day we'd pick up a cone at Steve's across the street before heading back over the Charles River basin to MIT. Most of our discussions revolved around stereopsis—the hottest problem, the size and number of channels, early primal sketch issues, and later, object recognition and why "segmentation" was misguided. At these

lunches, we rarely spoke about anything but vision. The one marked exception was flying, which had become his principal hobby.

The uniqueness of these times is very difficult to express in writing. There were a dozen or so in Marr's group. The energy level, excitement, and dedication was exceptional. When David was not in Tubingen with Tommy Poggio, we would meet weekly to discuss each individual's progress or special topics such as the Retinex, short and long range motion, grouping, axis finding, and the building of a vision machine. Once a month or so a visitor came through and offered their latest ideas. Typical of MIT, the visitor was battered with questions from the small audience, each anxious to find the soft spot. Yet, after all of us had had our chances, David would sum up the work in a few sentences and then proceed to point out any serious weakness that all of us had missed. He had a remarkably clear view and an exceptional ability to cut to the quick in a Mozart-like manner.

In late 1976, we approached NSF with a proposal called "Vision Algorithms and Psychophysics" which was to be the principal vehicle for funding (with AFOSR) of the more biological aspects of the group's work. (The machine-oriented aspect was provided by the AI Lab through the efforts of Patrick Winston and Mike Brady.) This was a significant award not only because of its interdisciplinary nature, but because it proposed a specific strategy for attacking the visual system that included both computational and experimental components. While writing the introduction, we realized that the "three levels" of understanding could also be recast as a scientific protocol for attacking a vision problem: Step 1: Propose a computational theory; Step 2: Write an algorithm embodying the theory; Step 3: Check out its (biological) merit with psychophysics (or neurophysiology). This simple paradigm was used first on the stereoscopic problem, leading to a rejection of the cooperative model by psychophysics in favor of the later noncooperative models culminating with the Grimson-Marr-Poggio version. The impact of psychophysics on this development was critical, and a major benefit of the interdisciplinary approach. (As an aside, a visit by the two Johns—John Frisby and John Mayhew—who presented extensive psychophysical findings in support of matching features other than zero crossings, also had caused considerable impact, especially on the development of later models.) Unfortunately, over the course of the years, as the problems became harder, the enforcement of the 1-2-3 method became lax, and often we did not get closure when studying a problem. However, I believe the best theses of the group were those that completed at least one iteration of these three steps.

During the next seven years, the activity level of Marr's vision group was prodigious. In this period not only was stereo seen from a different, more computational viewpoint, but also motion, color, edge detection, some aspects of object recognition, and the age-old grouping problem. Our confidence was enormous (and unfortunately, overbearing to many!). In October, 1977, we decided to invade the annual Optical Society meeting in Toronto to "show the

others just how vision should be studied." I arranged a symposium on "Vision by Man and Machine" that ended with Marr as the key speaker. (At this time, David was still relatively unknown.) Unfortunately, the presenta- tions were late starting and ran too long, leaving David only half his allotted time. So with the approval of Lorrin Riggs, the presider, we arranged for David to complete his talk in the technical discussion session that evening. This was a disaster! Already most were irritated by our arrogance, and the last straw was for the technical discussion period now to be replaced by still more lectures from MIT. Leo Hurwich spoke strongly against this policy change, and left in protest. Nevertheless, a majority stayed, realizing that no one leaves a technical discussion session unscathed. Quite true! Richard Blackwell, always ready with penetrating questions, led the attack, targeting Michael Riley's work on texture boundaries. (Michael was the youngest member of the group, just in the process of completing his bachelor's thesis on texture.) Finally, the discussion moved away from the MIT papers, and eventually the discussion/critique session ended. Afterward, I recall that we all, still undaunted, headed off to a restaurant opposite the Needle, where we had a grand evening, still convinced of the merit of our new computational wave. Now, the Optical Society hosts one of the most positive and supportive groups espousing computational approaches to biological vision.

The reaction of the Optical Society to our approach was not unusual. However, by 1980 there was a significant shift in the balance of work on vision in favor of computational approaches. David's activities were stirring interest, and groups at other universities and laboratories were developing interdisciplinary groups similar to ours, with competing ideas. Unfortunately, on occasion this led to some strain, but overall the competition was a very positive factor, for it challenged us to remain productive and thorough. Within MIT the recognition of David's work of course came earlier, and in 1977 there was sufficient interest that a faculty appointment was proposed in the Department of Brain and Cognitive Sciences (then inappropriately named "Psychology"). This was to be a joint appointment with the AI Lab, where he was currently a Research Scientist. As with all new faculty positions, the candidate must give a "job talk" to the faculty. David elected to speak on one aspect of his Primal Sketch paper, namely to show what would be required for a so-called feature detector like a bar mask ("simple cell") to assert the presence of a line. The talk was another disaster. First, most of the audience did not see why, in the first place, a simple bar mask couldn't report the presence of a single line (a collection of such masks is required). And second, the mathematical details and computational recipes were boring to most experimental psychologists, and especially so to the philosophers! However, in spite of the negative reception to his talk, there were enough respected supporters for David to get the appointment. The result was formal recognition of the potential of building a bridge between AI and Psychology—a bridge that David worked hard to maintain while at MIT.

The future looked bright and conquerable in 1977 until December. Then we learned that David would probably have only another three years. It may be hard to understand that these years, although painful, were also rewarding, exciting, and at times a lot of fun. Most of the dark days were spared us, for David now spent much time in England for treatment and recovery with his parents. Ideas still flowed at a fast pace; we had tremendous momentum. In one three-year span alone there were 120 publications. During this period David brought much of this together in his book Vision, ably assisted in the transcription by our secretary Carol Bonomo, who was the world's fastest talker and a perpetual optimist. We were all excited by the prospect of David's book, recognizing it would be a milestone and hence eager to learn of its progress. We looked forward to David's return and to the days he would lead our research meetings, which continued even in his absence. Any little victory was celebrated, reserving lobsters and champagne for special occasions. Now, with Lucia, the simple things of life were enjoyed, which previously had often been passed by.

Yet the vision never wavered. David's principal aim was to unravel the mysteries of the human mind, choosing as his route the understanding of the information processing carried out by the human nervous system. To this end, David accomplished more in a few years than most of us can in a lifetime, and set in motion a wave for the future. He believed that these first steps toward understanding how our brains work would eventually "change man's image of himself, and that most current philosophies of life and thought would have to undergo a profound transformation to deal with this new knowledge." His work represents the beginnings of this great new adventure.

*March, 1990*

*Colleague, friend and "guardian angel" who protected David from bureaucratic nuisances.*

*Professor*
*Department of Brain and Cognitive Sciences*
*Massachusetts Institute of Technology*
*Cambridge, Massachusetts*

## Tommy Poggio

I met David for the first time in the fall of 1973 when I came to Boston to chat with Marvin (Minsky) at the Artificial Intelligence Lab. Boston was wet, foreign and dark. David came out of his office in the "playroom." We exchanged a few words. His name was known to me, of course, because of his cerebellum paper, which was highly praised by many VIPs in the biological sciences.

Three weeks later I was again in Boston for an NRP meeting. David was also at this meeting. We had both been invited at the last moment and were not scheduled to speak. David sat quietly the whole time, listening to what people were saying about psychophysics and physiology. John Dowling joked with him about David's red Mustang. Back at the AI lab for the first time, a scientific conversation took place about the ideas on the retina he was then developing. I was the messenger of an invitation for him to give a series of lectures at the Spring course on Biophysics in Erice, Sicily. I was happy that he accepted immediately, and so our next meeting was already arranged.

Erice is a beautiful old village on top of a mountain overlooking the Mediterranean sea. For two weeks the participants–among them Mike Fuortes, David Hubel, Bela Julesz, John Szentagothai, Sir John Eccles, Michael Arbib—gathered together mornings and afternoons. At lunch and dinner we divided into small groups to explore the five restaurants of Erice, all above scientific average. There were also several expeditions to the various beaches down the hill. I was impressed and obviously pleased by the interest and the respect David had for my lectures and my comments. David's approach was by far the most unconventional and for me the most interesting. We discussed at length, dining, lying on the beach.

One year later, I came to MIT to the AI lab to work with David, to clear my head and decide what to do next. For the first week or so I was left relatively alone, free to play with Macsyma and Lisp. I spoke about the visual system of the fly at David's vision seminar. It went well. After the lecture we had a beer together in Harvard Square. David was very happy about my lecture. His enthusiasm, his praise, were contagious. I felt great and alive!

Our lunches together—at the Tech cafeteria—were still trying to define the nature of our approaches to the problem of the brain. Our views were already very near and converged rapidly.

In the meantime, I was understanding more and more David's work on

vision. In retrospect it took a lot of time. Really new ways of thinking cannot be understood at once. A thousand different facets must be communicated with the magic of a language and the fascination of a style. David's papers on early vision all have these rare properties.

A good way to understand something new is to try to criticize it, playing the role of the advocat du diable. In doing this with David, I found myself at some point defending recurrent networks. David formulated the challenge of solving stereopsis in his way. He told me in terms of his analysis of the computational problem of stereopsis which "cells" had to be excited and which ones had to be uninhibited. On the back of a paper napkin at the Tech cafeteria I wrote down the obvious equations for the recurrent network. I claimed that it would be very easy to prove its convergence. Liapunov-like functions constructed from conditional expectations was what I had in mind. On the same evening, back at the lab, after a dinner at the Greek restaurant in Central Square, David programmed the recurrent algorithm which seemed to work well in 1-D. The day after, the 2-D version of the algorithm gave encouraging signs of liking Julesz' stereograms. One week later, when I had finally understood David's computational analysis, I also realized that an analysis of the convergence of the algorithm was going to be very difficult. All general stand- ard methods failed. When I told David that I had to turn to the last resort—a probabilistic approach—he teased me. The teasing became even more intense when I had to write a program to compute the result of the probabilistic analysis.

In the meantime, creative life was exciting, despite the headaches from the convergence problem. We started working closely together. It was a fantastic experience. David was very sharp; he had clear ideas about almost everything and they were usually right. At that time I was also slowly discovering various facets of him. He was passionate about music—Italian opera—for instance. I heard him improvising a few times on the piano. He played with ease and emotion. I was impressed. I had to wait another year, however, before experiencing him playing his instrument, the clarinet.

During those three months in Boston I often went sailing on the Charles. But the really new experience was flying. David used to fly and my presence triggered anew this passion. I took a few flying lessons. Several times we flew together with a rented Cessna. On those occasions I used to stay overnight at David's house. Early in the morning we heard the weather forecasts and then drove to Hanscom where we would get a plane from Patriot Aviation for the day and share expenses. One of the most beautiful flights brought us to the Lakes Region. The sky was clear and deep blue as the water beneath us. We landed on a grass strip on a little island. It was very quiet. We walked to the water a few hundred yards away. There we sat for a few hours. We reviewed what we had done on stereopsis and decided to write a short paper for Science about it. David already had the opening sentence and the overall formulation was clear. We just had to sit down the next day and write. There was only

one white sail on the lake. Blue air. Green, and blue, and silence. David was happy and relaxed. So many more ideas and flights and forests and lakes were waiting for us!

On our way back to Hanscom the weather changed quickly. The rain started. On the final approach, David was very tense, his mind totally concentrated on the plane, the control tower and the instruments. Several people were afraid of this concentration, which they misled as a sign of unfriendliness. I knew well the alertness of David's mind when he was discussing science, lecturing, playing music or flying I could physically feel the presence of his thoughts, his total concentration. There was often an incredible "intensity" in his thinking. His reactions and his answers were then incredibly quick and at the same time crystal clear, sure and sharp. Our landing on the wet field at Hanscom was perfect. Five minutes later the airport closed down.

On another flight expedition to Nantucket and Martha's Vineyard, we got sunburned on the beach. Two days later, red like two lobsters, we gave a lecture on stereo at Harvard Medical School. We were making grand plans of flying across America for one month or so. Life was going to be a lot of fun!

A few weeks before my return to Germany, we were right in the middle of bicentennial time. The tall ships were coming into Newport on their way to New York. On the weekend (July 2nd) the weather was beautiful and David decided to fly down to Newport. We came above the bay with our Cessna 170 to find out that the blue sky was filled with flying objects. Balloons, choppers, a Goodyear "blimp" and many other planes. Tower instructions were to circle at x feet above the tall ships. The surface of the sea was covered by little white traces, glittering under the sun. There were hundreds of boats of all sizes that came to meet the tall ships. The scenery was superb. It was simply great, circling above the ships, together with so many other planes and boats. Down in Newport airport hundreds of planes were scattered all over the field, many of them old timers, happy and colorful. In the afternoon the weather deteriorated quite suddenly. There was a storm and ghastly winds. Back to the airport, we thought for a while to leave the plane and go back to Boston in some other way. David phoned several times to inquire about the weather at Hanscom. It was clear and so he decided to start. Airborne again, drops of rain slashed across the windscreen until we came from the low clouds out in the sun. It was the eternally beautiful weather to which poets have accustomed us. But the feeling in a small plane without instruments is quite different. David, however, was relaxed. There was nothing to do but fly straight and wait for the clouds to dissolve. Near Boston, the ground started to appear at short intervals through foggy holes in the white carpet above which we were flying. When we landed at Hanscom the sun was setting down against a clear sky.

It is not by chance that my deep friendship with David was associated with flying together. Flying and friendship, joy and beauty, freedom and living are things that are made of the same substance. I did not fly anymore with a light

plane after David got ill. I don't know whether I am going to do it ever again.

David came to Tubingen in the beginning of 1977. He stayed in the guest room of the Institute and walked over to our home every morning for breakfast. We worked on the probabilistic analysis of stereopsis, every day discovering some more difficulties. David wanted us to think about a theory of human stereopsis. Eye movements were important. At that time I had just heard from Jack Cowan of his and Hugh Wilson's work on spatial frequency channels. David brought Mayhew and Frisby's Nature paper on rivalrous stereograms. These two ingredients, together with the refusal of our first algorithm and the need of eye movements, formed our starting point. We read everything on stereo from Barlow's seminal paper to all of Julesz'. At some point we were suffocating in my office under piles of bound volumes of the Journal of Physiology and Vision Research. We even did some informal experiments. At the end of the three weeks we had written three-fourths of the analysis of the cooperative algorithm paper (I had to write the final quarter with Gunther Palm) and had some rough ideas about a new model of stereopsis.

David brought his clarinet. I introduced him to Eric Buchner, a good cello player. With another friend, a very good pianist, they played together several times. All of us were deeply impressed by David's music. During David's visit in Tubingen, the members of the Scientific Curatorium of the Institute came one day to meet with the members of our Institute. In the evening after dinner, David and other friends played one of Beethoven's Quartets. I never was so deeply struck by music as I was that evening by David's clarinet. It was so beautiful and perfect, so full of emotion as to be almost unbearable. The audience—it was quite clear afterwards—had a similar experience.

He was quite alone in his work at that time. He did not have anybody back at the lab with whom to work in the same way we did. I suggested to him to try to work with Shimon and share with him responsibilities of the group and of the students. At that time I knew Shimon only superficially but my feelings and what David thought about him and his work left no doubts. David promised he would do it. It was an easy promise to fulfill. He also promised several times that he would finally get out of his "craziness" and his "women problem." But he never managed until he met Lucia, one year later.

Those three weeks in Tubingen were a lot of fun; life was full, warm and happy.

In June, 1977, David again came to Tubingen. He stayed a full month in "his" room in the Institute. We worked hard, developing our stereo ideas and writing them down at the same time. The days were productive. The theory took form. Through all my work with David it was often impossible to say who had a specific idea; almost everything came out from discussions and thinking together and reciprocal criticisms. David had the power of vetoing: if I was unable to convince him, that was it. He also had the ability of keeping us right on course.

I remember the origin of part of the zero-crossing idea. Coming out of

the cafeteria I expressed my uneasiness about taking zero-crossing and peaks of the filtered images, since filtering the images was roughly equivalent to making their second derivative zero-crossing correspond to extrema of the first derivative. This made sense. But peaks were something strange, at least at this level. For simplicity, and because of the relations between derivatives, difference of gaussians and bandpass channels, I wanted to flush peaks and retain zero-crossing only. David thought a while and then decided that—for reasons I did not think of—the idea was not too bad. It is still unclear whether he was right.

We finished our manuscript right on schedule with some time left to take Polaroid pictures of the two authors sitting with the title in one hand and stereoglasses in the other. (In the original draft of the manuscript there were a few lines warning the secretary that at that particular point we had just had too much Courvoisier and therefore the following sentences were going to be particularly immortal.)

The whole month was continuous, concentrated, happy playing. As so often with David, science was fun and freedom! I often ask myself why David's presence had this fascination, this incredible power. I still find it very difficult to give a full answer. But I know that part of it was the clarity and especially the force of his mind, of his thoughts. To think with David was for me an inebriating experience, a special feeling of playing and creating. Skiing beautifully downhill on a sunny day in the Alps gives me some hint of this intellectual fun.

Werner (Reichardt) organized a Neurobiology meeting for the 500 years of Tubingen University. I had helped in setting the framework of the lectures. Many friends came: David, Vincent Torre, David Hubel, Dennis Baylor, Emilio Bizzi, Gunther Stent, Jack and Max Cowan, Bela Julesz and others.

David's lecture was beautiful, crystal clear, a jewel of intellectual brilliance and improvisation. We had the feeling that the world was there for us to play.

In the middle of October I flew to Toronto for the annual meeting of the Optical Society of America. I was invited by Whitman Richards who organized a special session. The whole MIT Vision group came. It was fun, although short and chaotic. A couple of days later I flew to Boston to work with David for three weeks. It was a fight with LISP and probability (again!). At the end of my stay, we drove together in a rented car through a colorful New Jersey down to Bell Labs. I gave a lecture for Bela Julesz and his small group on a topic that was completely uninteresting to them, synapses. When we mentioned our probabilistic analysis of zero-crossings, Bela named some mathe- maticians at Bell Labs who had worked on somewhat similar topics. Among them there was a name that we did not know, Ben Logan. We asked for the paper and Bela sent his secretary to get reprints. Glancing through it I saw that his theorem was very suggestive of our notion of independent bandpass channels. In the hotel and later, in the car, I tried to convince David, who remained quite skeptical. The zero-crossing idea and its connection with

Logan's theorem is of the kind I immediately like. Unfortunately, such ideas are often too nice to be correct and David was certainly right in his skepticism.

*November, 1980*

*David's closest collaborator and great friend.*

*Uncas and Helen Whitaker Professor*
*Brain and Cognitive Sciences*
*Artificial Intelligence Laboratory*
*Co-Director, Center for*
*Biological Information Processing*
*Massachusetts Institute of Technology*
*Cambridge, Massachusetts*

## Shimon Ullman

When I came to MIT as a graduate student in the summer of 1973, David Marr was already there, having arrived from England a couple of months before to work at the AI lab. This was extremely fortunate for me. I came to the AI lab with the intention of studying brain functions, and in particular visual perception, using mathematical models and computer simulations. From the limited literature I had seen about MIT's AI lab I had the impression that this was the main focus of the scientific activity there. As it turned out, however, the emphasis at the time was primarily on machine intelligence, and nobody at the lab was actively involved in the study of biological brain functions. When I started to talk with David soon after my arrival at the lab, it immediately became clear to me that he was the person I wanted to do my Ph.D. work with. We had similar interests, and a similar background that started in an interest in pure mathematics, then shifting to biology, with an interest in the brain and its functions, and then to artificial intelligence in an attempt to model some aspects of the human visual system. David could not be my formal thesis advisor at the time, since he was not yet a faculty member. Soon, however, he became my unofficial advisor, with Marvin Minsky's tacit blessing. Although he was my advisor, he was only slightly older than me, and after a short while we also became personal friends.

Working with David was always challenging, exciting and rewarding. It was hard work, but it was a lot of fun. We had the feeling that our small group, centered around David (that included Whitman Richards, Tommy Poggio and a number of David's students), was creating something new and exciting. Around the time I had finished my Ph.D. work, David and I worked together extensively for a period of a few months on some problems in motion perception, and later wrote a paper on this work. It was for me the first, and in fact still the only time, that I wrote a paper with someone in this mode, actually sitting together for long hours at a time, composing sentence after sentence, and discussing each paragraph as we wrote it. The experience was very intense and enjoyable. I think we both enjoyed it, and we were both exhausted by the time the paper was finished.

David was extremely quick, and expected others to be equally quick and alert. We once went down to Washington, D.C., to meet one of the sponsors of our work at the lab. We discussed with him some of our vision work, and then he asked if we had any views regarding new directions his agency should

perhaps be looking into. David snapped, "grow wires," without offering any additional explanation. I knew what he meant: he became interested at the time in possible hardware implementation of vision devices, and thought that the restrictions imposed by standard semiconductor technology on the number of interconnections among functional elements (much smaller than the number of connections among neurons in the brain) was a severe limitation on the way to producing compact and practical vision devices. I could see the puzzlement in the other person's expression, but David saw no need to elaborate the issue further.

David's illness came as a shock. He called me from the MIT infirmary on the day he was diagnosed with the disease, and asked me to close my office door. He then said briefly and without any introduction that he had acute leukemia. The period that followed was very painful. Twice during his illness we thought that there was some hope. The first was during his first remission. Everyone hoped that perhaps, by some miracle, the disease would not come back. We took a vacation together in Vermont, and he resumed his work with his usual intensity. After a period, he felt weaker and went to the hospital for some tests. He came to my office to call the hospital about the tests' results, and found out that it was indeed a relapse. We sat in my office for a long time, devastated by the news.

The second hope came when a physician in Cambridge, England, had some initial success with a vaccine against leukemia. David was hospitalized in Cambridge. He was very weak, and worked on this book. When I came to visit, I met the physician, who was very supportive and promised to help as much as he could. When David came back to the U.S., Tommy Poggio managed to bring some of the Cambridge vaccine with him, but it was too late to actually use it (and it did not prove effective in later clinical trials anyway).

The final period, when David already suspected that the battle was lost, was in fact a quietly happy one. He was happily married to Lucia, and was working intensively on his book and a number of other projects. In his premature death, the scientific world lost an intellectual giant, who, in a short time, made a huge impact on his field. We also lost a warm, brilliant, exciting, unusual friend.

*March, 1990*

*Colleague, studentt, close friend*

*Professor*
*Artificial Intelligence Laboratory*
*and Department of Brain and Cognitive Sciences*
*Massachusetts Institute of Technology*
*Cambridge, Massachusetts*

## Ellen Hildreth

When I first came to work with David, I was overwhelmed by him; by both his brilliance as a scientist, and his personal magnetism. It was perhaps a year before I could really feel comfortable with David. We have an expression for being overwhelmed by something, which is being "blown away." I sometimes had visions of myself going to talk with David, and being whooshed out of his office, chair and all, by a big gust of wind, and hanging on for dear life to the edge of his door, in order to hear every word he had to say. You always remembered the things David would tell you; it was often the case that you wouldn't understand him at first, but something about the way he said things would make the words stay in your mind, and hours, days, even weeks later, bright lights would snap on in your head, as you figured out what he meant by his cryptic message.

David was always very generous with his ideas; he'd solve half the problem for you, and then later insist that you did it all yourself. Ideas would come to him any time; sometimes the phone would ring at 9:00 on a Sunday morning (only my mother, or David would call at 9:00 Sunday morning)—David would greet us with a glorious "Hello!" and "I was just wondering if by chance you might have planned to come into work today; I have a new idea you might like to try out." Whether I had planned to come in or not, without a moment's thought, I'd answer "Of course, David! I was just on my way out the door!"

It was the enthusiasm that David instilled in us that made us want to do these things; everything we did was so important to him, so vital. This always made us feel somewhat under pressure; it wasn't a pressure that David placed on us directly—his enthusiasm just made us want to be always working madly at our research. And he was always so busy himself, you felt bad if you weren't working at least as hard.

David always thought big, and tried to teach his students to do this as well; it wasn't enough to study an aspect of stereo, a subproblem of motion, or a particular type of texture problem. You gritted your teeth, and went in to tackle the whole problem of stereo, motion or texture head on. He'd always prefer to present a "whole theory" of something (which might be a bit lacking in detail), than to present an explanation for any part of a problem.

He'd have little "favors" for you to do make a glossy of something, a demo, run an experiment and was so overly courteous and charming in the

way he asked. With a twinkle in his eye, he'd ask us to do it whenever you had the time—there was no rush, take a week, a month, whatever. The moment he was out the door, you'd drop everything, work on his project non-stop, and have the results on his desk the next day. (Do you suppose he knew you'd do this?)

Getting back a paper with David's comments was really something that took getting used to—we were never quite sure how to interpret the "GUF-FAWs" and "tee hees" and "oh really?'s", but the bright, red, bold "rubbish" and "No!'s" were a little more obvious—David could really be devastating at times. We were all very much "in tune" with David's moods. If he was in a jovial mood, so were we, but if he was unhappy about something (particularly if it was something we did), our emotions could be destroyed for days.

David meant so much to us, and his teachings were so important, but I must admit that I was quite taken aback one day when a visiting scientist asked me if David was like a "guru" around the lab, because that he was not. He was human, like the rest of us, and could be wrong sometimes too (he just made mistakes with so much more style than the rest of us). What he believed in, he believed very strongly, but if you presented a convincing argument for the other side, he'd change his mind. He trained his students to stand on their own two feet, and be their own people—sometimes playing the devil's advocate, just to get us to argue with him.

David held a strong presence in the lab; you always knew when he was in; the word would get around. Someone would spot him up at the XGP (printer), or wandering down to the Xerox machine, or logged in, or would notice that his office door was shut (a sure sign that he was both in, and not wanting to be disturbed), and would spread the word around that David was in. He'd make a point of popping in from day to day to see what you were up to, so you had to be sure you looked busy. We'd feel horrible if he caught us chatting in the playroom or bullshitting about politics in the office. He wasn't a slave driver by any means (although we used to kid him about the 30-foot bullwhip he kept hidden in the office); on the contrary, he was one of the most gentle and gracious people I have ever met (second only to his mother). He was just a very stimulating person; the energy level in the lab would suddenly double when David walked in (it would quadruple if Tommy [Poggio] was around too we used to refer to the two of them as the "Dynamic Duo").

I had worked with the Logo group for three years before I came to work with David. I came to him knowing almost nothing about human vision, and very little in my background to offer, except some applied math. But that didn't matter. He felt it was important for a person to have some background in an analytic discipline, but beyond that, all that was necessary was an eagerness to learn; everything else would come. It took a tremendous amount of time and patience for David to work with someone with so little background in his field of research, and I just can't say in words how much I admire David for having that time and patience; it's helped me to establish my life's work, and

to develop me in many personal ways, as well.

The quiet courage with which he faced the last three years was very hard on us. David would try so hard to not let his illness interfere with his work and interacting with his students. He always kept things to himself; the time that we had with him was so valuable that every moment was spent talking about vision, and how we were doing in our work. His students were always so important to him. When he had so many more important things in his life to be concerned about, he'd be worrying about the Vision group. He always worried about me much more than I worried about myself. But in the times that I've had with David, he's given me far more than I need to keep going for a lifetime.

*November, 1980*

*Student, collaborator, friend.*

*Associate Professor*
*Brain and Cognitive Sciences*
*Artificial Intelligence Laboratory*
*Co-Director, Center for*
*Biological Information Processing*
*Massachusetts Institute of Technology*
*Cambridge, Massachusetts*