STUDIES OF MIND AND BRAIN

BOSTON STUDIES IN THE PHILOSOPHY OF SCIENCE

EDITED BY ROBERT S. COHEN AND MARX W. WARTOFSKY

VOLUME 70

STEPHEN GROSSBERG
Department of Mathematics, Boston University

STUDIES OF MIND AND BRAIN

Neural Principles of Learning, Perception, Development, Cognition, and Motor Control

D. REIDEL PUBLISHING COMPANY
DORDRECHT : HOLLAND / BOSTON : U.S.A.
LONDON : ENGLAND

Library of Congress Cataloging in Publication Data

CIP

Grossberg, Stephen, 1939–
Studies of mind and brain.

(Boston studies in the philosophy of science ; v. 70)
Bibliography: p.
Includes index.
1. Learning–Physiological aspects–Collected works. 2. Mind and body–Collected works. 3. Neural circuitry–Adaptation–Collected works. 4. Memory–Collected works. 5. Neuropsychology–Collected works. I. Title. II. Series.
Q174.B67 vol. 70 [QP408] 501s [153] 81-23453
AACR2

ISBN-13: 978-90-277-1360-5 e-ISBN-13: 978-94-009-7758-7
DOI: 10.1007/978-94-009-7758-7

Published by D. Reidel Publishing Company,
P.O. Box 17, 3300 AA Dordrecht, Holland.

Sold and distributed in the U.S.A. and Canada
by Kluwer Boston Inc.,
190 Old Derby Street, Hingham, MA 02043, U.S.A.

In all other countries, sold and distributed
by Kluwer Academic Publishers Group,
P.O. Box 322, 3300 AH Dordrecht, Holland.

D. Reidel Publishing Company is a member of the Kluwer Group.

To Gail,

My Parents,

and My Friends.

TABLE OF CONTENTS

PREFACE ix

ACKNOWLEDGEMENTS xi

INTRODUCTION xiii

1. How Does a Brain Build a Cognitive Code? 1
2. Some Physiological and Biochemical Consequences of Psychological Postulates 53
3. Classical and Instrumental Learning by Neural Networks 65
4. Pattern Learning by Functional-Differential Neural Networks with Arbitrary Path Weights 157
5. A Neural Theory of Punishment and Avoidance. II: Quantitative Theory 194
6. A Neural Model of Attention, Reinforcement and Discrimination Learning 229
7. Neural Expectation: Cerebellar and Retinal Analogs of Cells Fired by Learnable or Unlearned Pattern Classes 296
8. Contour Enhancement, Short Term Memory, and Constancies in Reverberating Neural Networks 332
9. Biological Competition: Decision Rules, Pattern Formation, and Oscillations 379
10. Competition, Decision, and Consensus 399
11. Behavioral Contrast in Short Term Memory: Serial Binary Memory Models or Parallel Continuous Memory Models? 425
12. Adaptive Pattern Classification and Universal Recoding. I: Parallel Development and Coding of Neural Feature Detectors 448
13. A Theory of Human Memory: Self-Organization and Performance of Sensory-Motor Codes, Maps, and Plans 498

LIST OF PUBLICATIONS 640

INDEX 644

EDITORIAL PREFACE

Throughout the history of philosophy, the project of a naturalistic epistemology - of a theory of knowledge based upon a scientific account of the natural processes of perception and cognition, and of learning - occupied such major thinkers as Aristotle, Descartes, Hume, Reid, Peirce, and recently philosophers and scientists from Helmholtz and Mach to Piaget, Popper and Gibson. The question of how knowledge is acquired is two-sided. On the one hand, there is the epistemological questions *par excellence*: what is truth? by what criteria, or under what conditions, are cognitive claims warranted? On the other hand, there is the question of how the human organism, with its structure of sense perception, language and thought, can acquire veridical knowledge of this world?

With the advent of evolutionary theory in biology, human perceptual and cognitive activity came to be seen in its relation to the more general acquisitions of animal learning or animal intelligence, from which it was believed to have evolved. Attention to the comparative anatomy and physiology of the nervous systems of different species focussed on both the gross structure of behavior as an interaction between organism and environment, and on the fine structure of the neural response subtleties of the sense modalities, and on the cross-modal and higher integrative functions of the brain. In the modern period, naturalistic theories of knowledge therefore have been framed in terms of both biological and psychological description, and have aspired to mathematical formulation in the image of the natural sciences.

Stephen Grossberg's studies, gathered in this volume, lie at the intersection of psychology, neurophysiology, and mathematics. The problem he sets for himself, however, is deeply philosophical and methodological: is a mathematical model of a dynamic, evolving, adaptive system possible? Can such a mathematical model adequately account for such psychological phenomena as arousal, attention, memory, or more generally learning, perception, cognition? Grossberg approaches this not as a formal problem but as a concrete research task. He posits the two major constraints: neural anatomy and function of the brain, and operations in real time. Given these spatial or topological, and temporal constraints, and basing his analysis on the mass of experimental data from current research in psychology and physiology, Grossberg proposes and develops a non-linear mathematics as a model for specific functions of mind and brain. He finds the classic approach to the mathematical modelling of mind and brain systematically inadequate. This inadequacy, he holds, arises from the attempt to describe adaptive systems in the mathematical language of a physics developed to describe "stationary", i.e. non-adaptive and non-evolving systems. In place of this linear mathematics, Grossberg develops his non-linear approach. His method is at once imaginative, rigorous, and philosophically significant: it is the thought experiment. It is here that the richness of his interdisciplinary mastery, and the power of his methods, constructions and proofs, reveal themselves. The method is what C. S. Peirce characterized as the method of abduction, or of hypothetical inference in theory construction: given the output of the system as a psychological phenomenon (e.g.learning, perception, cognition) and interpreting such activities in an evolutionary context, as adaptive behavior with respect to complex and changing patterns of the environment, how can the known structures and properties of neural networks account for the known behavior or features of neural and psychological activity given by the experimental data?

Thus Grossberg deals with such general problems as "how does the brain build a cognitive code?", and such specific ones as, "how does an on-center off-surround anatomy of networks of nerve cells lead to such characteristics of the neural processing as contour enhancement in vision or short-term memory?"

Grossberg's papers in this volume seem to us to make a major contribution to the theoretical formulation of problems in the study of mind and brain, and to their mathematical and empirical solution.

Boston University
Center for the Philosophy
and History of Science
February 1982

ROBERT S. COHEN
MARX W. WARTOFSKY

ACKNOWLEDGEMENTS

The author, the editor, and the publisher are grateful to the following persons and institutions for permission to reprint the papers included in this volume:

"How Does a Brain Build a Cognitive Code?". (First published in *Psychological Review* 87 (1980), 1–51.)

"Some Physiological and Biochemical Consequences of Psychological Postulates". (First published in *Proceedings of the National Academy of Sciences* **60** (1968), 758–765.)

"Classical and Instrumental Learning by Neural Networks". (First published in *Progress in Theoretical Biology*, Vol. 3, Academic Press, New York, 1974, pp. 51–141.)

"Pattern Learning by Functional-Differential Neural Networks with Arbitrary Path Weights". (First published in K. Schmitt (ed.), *Delay and Functional Differential Equations and their Applications*, Academic Press, 1972, pp. 121–160.)

"A Neural Theory of Punishment and Avoidance. II: Quantitative Theory". (First published in *Mathematical Biosciences* **15** (1972), 253–285.)

"A Neural Model of Attention, Reinforcement and Discrimination Learning". (First published in *International Review of Neurobiology* **18** (1975), 263–327.)

"Neural Expectation: Cerebellar and Retinal Analogs of Cells Fired by Learnable or Unlearned Pattern Classes". (First published in *Kybernetik* **10** (1972), 49–57.)

"Contour Enhancement, Short Term Memory, and Constancies in Reverberating Neural Networks". (First published in *Studies in Applied Mathematics* **LII** (1973), 213–257.)

"Biological Competition: Decision Rules, Pattern Formation, and Oscillations". (First published in *Proceedings of the National Academy of Sciences* 77 (1980), 2338–2342.)

"Competition, Decision, and Consensus". (First published in *Journal of Mathematical Analysis and Applications* **66** (1978), 470–493.)

"Behavioral Contrast in Short Term Memory: Serial Binary Memory Models or Parallel Continuous Memory Models?". (First published in *Journal of Mathematical Psychology* **17** (1978), 199–219.)

"Adaptive Pattern Classification and Universal Recoding. I: Parallel Development and Coding of Neural Feature Detectors". (First published in *Biological Cybernetics* 23 (1976), 121–134.)

"A Theory of Human Memory: Self-Organization and Performance of Sensory-Motor Codes, Maps, and Plans". (First published in *Progress in Theoretical Biology*, Vol. 5, Academic Press, 1978, pp. 233–374.)

INTRODUCTION

How is psychology different from physics? What new philosophical and scientific ideas will explicate this difference? Why were the inspiring interdisciplinary successes of Helmholtz, Maxwell, and Mach a century ago followed by a divergence of psychological and physical theory rather than a synthesis? Why has physics rapidly deepened and broadened its theoretical understanding of the world during this century, while psychology has spawned controversy after controversy, as well as dark antitheoretical prejudices?

My scientific work on problems related to mind and brain began in 1958 when I was an undergraduate, too young and enthusiastic to know about, let alone to worry about, these issues. After twenty years of scientific inquiry, answers are emerging which clarify some of the philosophical and scientific questions as well as the sociological ones. The answers suggest the following observations.

The difference between psychology and physics centers in the words evolution and self-organization. Classical physical theory focusses on a stationary world and the transitions between known physical states. Studies of mind and brain focus on a nonstationary world in which new organismic states are continually being synthesized to form a better adaptive relationship with the environment. These new states can thereupon be maintained in a stable fashion to form a substrate for the synthesis of yet more complex states in a continuing evolutionary progression. Perhaps no better example of this evolutionary process exists than language learning, which is one of the defining characteristics of human civilization.

Whereas physics has gradually fashioned a measurement theory for a stationary world, psychology needs to discover an evolutionary measurement theory, or universal developmental code. Whereas physics has been well served by linear mathematics, the evolutionary psychological processes (development, learning, perception, cognition) depend on nonlinear mathematics. Since the time of Helmholtz, Maxwell, and Mach, nineteenth century linear mathematics has stood ready to express and analyse the intuitive insights of physicists interested in electromagnetic theory, relativity, and quantum theory. Students of mind cannot turn to a well-developed body of appropriate mathematics with which to express their deepest intuitions. New nonlinear mathematics must be found that is tailored to these ideas.

Scientific revolutions wherein both physical intuitions and mathematical concepts need to be developed side-by-side are especially complex and confusing, but they also offer special intellectual rewards. In the present instance, understanding self-organizing systems is a necessary step towards understanding life itself, both in its individual and collective forms.

Brain studies play a central role in this pursuit for more than the egocentric reason that brains are the crucibles of all human experience. The brain is a universal measurement device acting on the quantum level. Data from all of our senses, - even a few light quanta! - are synthesized by our minds into a common dynamical coin that supports a unitary experience, rather than a series of dislocated experiential fragments. This universality property is the scientific reason, I believe, that brain studies are starting to play a role as central to evolutionary studies as black body radiation played in the development of quantum theory. This universality property clarifies the usefulness of brain theory laws towards explaining a growing body of data about living systems other than brains.

We find ourselves today in a paradoxical and disturbing situation. After physicists abandoned the study of mind, psychological experimentalists were left with an inappropriate world view for understanding each other's data. Personal experimental replication became a major source of security in an atmosphere of conceptual solipsism. Experimentalists dug into paradigms that were sufficiently narrow to maintain the replication criterion. Experimental approaches to mind hereby shattered into a heap of mutually suspicious fiefdoms, and mind theorists became persona non grata. This tendency has been exascerbated by short-sighted governmental policies that deny adequate funding of both the experimental body and the theoretical mind of our discipline. The same governmental policies encourage the search for easy and fast scientific fame. The nature of the crisis and the opportunity facing the brain sciences suggests that a long-range dialog between data and theory should be fostered instead.

Such a dialog plays a central role in the progress of my scientific work. My method of studying adaptive systems starts by identifying a fundamental environmental constraint, or problem, to which a species must adapt in order to survive. The solution to this problem takes the form of a principle of behavioral organization. The behavioral principle is translated into its minimal realization as a mathematical law. Minimality plays the role of an Occam's razor, or a principle of atrophy due to disuse, in the theory. I shall soon say how the theory overcomes the possibility that the prior evolutionary history of a system prevents the minimal solution from occurring. These mathematical laws have always possessed a vivid interpretation as neural networks. The formal mathematical language hereby bridges the gap between macroscopic psychology and microscopic physiology, much as a mathematical bridge exists between thermodynamics and statistical mechanics.

The reader might well ask: "Why have not all behavioral theories generated neurological insights?" An important part of the answer is this: All the principles in my theory describe how the organism solves the environmental problem in real-time. The theory is not merely formal or probabilistic. It attempts to describe the unfolding of individual behavior through time. This demand for individual real-time laws, simple though it seems, places strong constraints on the form that the solution can take.

Having expressed the behavioral principle in mathematical form, one is now confronted by a nonlinear mathematical system, and one must classify the possibilities inherent in this system. Unaided physical intuition has, time and again, proved unequal to this task. This is because the interactive, or collective, properties of the system control its interesting behavioral properties. The human mind does not easily grasp nonlinear interactions between billions of cells without mathematical tools. A rigorous mathematical method is needed to reveal the implications of the behavioral principle. Among the most comforting and rewarding facts of my life has been that mathematical methods could be invented for the understanding of behavioral principles. These mathematical methods effect a great conceptual simplification by structuring and predicting a large body of complex psychophysiological data as manifestations of a simple behavioral principle. If nothing else, this procedure confronts us with unexpected consequences of our present empirical beliefs, and provides a rigorous and transparent conceptual superstructure with whose aid new concepts can more effectively be fashioned.

It would be hard for me to overemphasize the importance of mathematics in these conceptual advances, although I was myself at first unsure of the need for a rigorous attack, as opposed to an intuitive attack. On many occasions, mathematical work has revealed a totally unexpected property, moreover a property so fundamental that it forced a whole series of new intuitive insights. On other occasions, by being able to recognize a general principle at work in several ostensibly unrelated bodies of data, I could regroup the data in terms of underlying principles, rather than in terms of experimental techniques. Each experimental technique can probe only certain aspects of a principle, but by pooling the results from several techniques that are used in seemingly distinct, but mechanistically related, situations, one can understand the underlying mechanisms much better than one could have by relying only on the techniques applicable in one situation.

The use of thought experiments to derive adaptive behavioral principles from environmental pressures, and the reorganization of data in terms of principles rather than experimental procedures, provide a powerful theoretical method for understanding brain and behavior. This method can detect information that eludes experimental techniques for several reasons: It shows how many system components work together; it compresses into a unified description environmental pressures that act over long, or at least nonsimultaneous times; and most importantly, it explicates design constraints that are needed to adapt in a real-time setting.

The mathematical classification theory approaches the question of minimality by admitting that several principles can simultaneously constrain the adaptive design of a given neural structure. The classification theory expresses its ambivalence towards minimality by suggesting species-specific variations on the same organizational theme which have adapted to principles other than the one under study.

Another important task of a classification theory is to clarify what a behavioral principle cannot achieve. In every case, a sharper understanding of a principle's limitations has suggested which other principles, which solve different environmental problems, are also at work in a given situation. Then the theoretical cycle begins again, and leads us in an evolutionary progression to a small set of adaptive principles and mechanisms capable of organizing and predicting a large variety of psychological and physiological data.

As I mentioned above, the collective or interactive properties of the mathematical laws subserve the adaptive behavioral properties that solve these environmental problems. In this sense, my theory is a 'field' theory which attempts to discover the conceptual level, and the functional transformations acting on this level, that drive particular aspects of the adaptive or evolutionary process. The evolutionary method also 'embeds' the properties of one principle into the properties of several principles acting together. For these reasons, the name *embedding field theory* still seems to be a convenient rubric for the method after the twenty-three years since its inception.

The ensuing papers are loosely grouped according to organizational principles and publication dates. The prefaces that introduce each paper sketch some of the issues, whether about nonequilibrium physical theory, language learning, mental illness, epistemology, or new engineering horizons, that in my mind stand above the scientific results as signposts for further scientific work and philosophical inquiry. The papers in this volume were published between 1968 and 1980. I spent most of the decade between the theory's inception and the first appearance of these articles acquiring the interdisciplinary skills that I knew would be needed. The foundations of the theory were laid while I was an undergraduate at Dartmouth College from 1957 to 1961. The theory continued to expand while I pursued graduate studies at Stanford University until 1964. Then I transferred to the Rockefeller University to write my Ph.D. thesis on this subject. A long monograph marked the first stage of my thesis writing. This experience was torrential and liberating after six years of silent but rapid accumulation of results. My Rockefeller professors generously funded the distribution of this 1964 monograph to one hundred leading laboratories in the U.S. and abroad. The monograph contained many of the physical laws and results which later appeared in papers of 1967-1969, but the theory still lacked a precise mathematical method for analyzing the nonlinear dynamics whereby arbitrarily many cells can learn. I found such a mathematical apparatus while I was a student at Rockefeller and it was the subject of my Ph.D. thesis. To my own surprise, this mathematical theory greatly amplified my physical intuition, and carried me through the first complete cycle of the evolutionary method. The prefaces to the papers sketch the several cycles that the theory has undergone since that time. Because of space limitations, some of the articles that developed a given theoretical cycle and forced the next cycle have been omitted. The prefaces indicate how both enclosed and omitted articles contributed to each cycle.

CHAPTER 1

HOW DOES A BRAIN BUILD A COGNITIVE CODE?

PREFACE

This article provides a self-contained introduction to my work from a recent perspective. A thought experiment is offered which analyses how a system as a whole can correct errors of hypothesis testing in a fluctuating environment when none of the system's components, taken in isolation, even knows that an error has occurred. This theme is of general philosophical interest: How can intelligence or knowledge be ascribed to a system as a whole but not to its parts? How can an organism's adaptive mechanisms be stable enough to resist environmental fluctuations which do not alter its behavioral success, but plastic enough to rapidly change in response to environmental demands that do alter its behavioral success? To answer such questions, we must identify the functional level on which a system's behavioral success is defined.

The article suggests that the functional unit of perception and cognition is a state of resonant activity within the system as a whole. Only the resonant state enters consciousness. Only the resonant state can drive adaptive changes in system structure, such as learned changes. The resonant state is therefore called an *adaptive resonance.* Adaptive resonance arises when feedforward (bottom-up) and feedback (top-down) computations within the system are consonant. The feedback computations correspond to our intuitive notion of expectancies. Feedback expectancies help to stabilize the code against errosive effects of irrelevant environmental fluctuations.

The adaptive resonance concept sheds new light on epistemological problems such as those which Heidegger considered. Is the Johnny I see today the same Johnny that I saw yesterday? Usually not. The resonant state constitutes the recognition act, but it also subverts itself by altering its own defining parameters. Tomorrow's resonance need not be the same as today's, yet certain invariant properties of the resonance remain unchanged, such as being able to say: Here's Johnny!

How Does a Brain Build a Cognitive Code?*

This article indicates how competition between afferent data and learned feedback expectancies can stabilize a developing code by buffering committed populations of detectors against continual erosion by new environmental demands. The gating phenomena that result lead to dynamically maintained critical periods, and to attentional phenomena such as overshadowing in the adult. The functional unit of cognitive coding is suggested to be an adaptive resonance, or amplification and prolongation of neural activity, that occurs when afferent data and efferent expectancies reach consensus through a matching process. The resonant state embodies the perceptual event, or attentional focus, and its amplified and sustained activities are capable of driving slow changes of long-term memory. Mismatch between afferent data and efferent expectancies yields a global suppression of activity and triggers a reset of short-term memory, as well as rapid parallel search and hypothesis testing for uncommitted cells. These mechanisms help to explain and predict, as manifestations of the unified theme of stable code development, positive and negative aftereffects, the McCollough effect, spatial frequency adaptation, monocular rivalry, binocular rivalry and hysteresis, pattern completion, and Gestalt switching; analgesia, partial reinforcement acquisition effect, conditioned reinforcers, underaroused versus overaroused depression; the contingent negative variation, P300, and ponto-geniculo-occipital waves; olfactory coding, corticogeniculate feedback, matching of proprioceptive and terminal motor maps, and cerebral dominance. The psychophysiological mechanisms that unify these effects are inherently nonlinear and parallel and are inequivalent to the computer, probabilistic, and linear models currently in use.

How do internal representations of the environment develop through experience? How do these representations achieve an impressive measure of global self-consistency and stability despite the inability of individual nerve cells to discern the behavioral meaning of the representations? How are coding errors corrected, or adaptations to a changing environment effected, if individual nerve cells do not know that these errors or changes have occurred? This article describes how limitations in the types of information available to individual cells can be overcome when the cells act together in suitably designed feedback schemes. The designs that emerge have a natural neural interpretation, and enable us to explain and predict a large variety of psychological and physiological data as manifestations of mechanisms that have evolved

* This work was supported in part by the National Science Foundation (NSF MCS 77-02958).

Requests for reprints should be sent to Stephen Grossberg, Department of Mathematics, Boston University, Boston, Massachusetts 02215.

to build stable internal representations of a changing environment. In particular, various phenomena that might appear idiosyncratic or counterintuitive when studied in isolation seem plausible and even inevitable when studied as a part of a design for stable coding.

Some of the themes that will arise in our discussion have a long history in psychology. To achieve an exposition of reasonable length, the article is built around a thought experiment that shows us in simple stages how cells can act together to achieve the stable self-organization of evironmentally sensitive codes. If nothing else, the thought experiment is an efficient expository device for sketching how organizational principles, mechanisms, and data are related from the viewpoint of code development, using a minimum of technical preliminaries. On a deeper level, the thought experiment provides hints for a future theory about the types of developmental events that can generate the neural structures in which the codes are formed. It does this by correlating the types of environmental pressures to which the developmental mechanisms are sensitive with the types of neural structures that have evolved to cope with these pressures. References to previous theories and data have been chosen to clarify the thought experiment, to contrast its results with alternative viewpoints, to highlight areas in which more experimentation can sharpen or disconfirm the theory, or to refer to more complete expositions that should be consulted for a thorough understanding of particular results. The thought experiment and its consequences do not, however, depend on these references, and the reader will surely know many other references that can be used to confront and interpret the thought experiment.

1. A Historical Watershed

Some of the themes that will arise were already adumbrated in the work of Helmholtz during the last half of the 19th century (Boring, 1950; Koenigsberger, 1906). Unfortunately, the conceptual and mathematical tools needed to cast these themes as rigorous science were not available until recently. This fact helped to precipitously terminate the productive interdisciplinary activity between physics and psychology that had existed until Helmholtz's time, as illustrated by the perceptual contributions of Mach and Maxwell (Boring, 1950; L. Campbell & Garnett, 1882; Ratliff, 1965) in addition to those of Helmholtz (1866, 1962); to create a schism between psychology and physics that has persisted to the present day; and to unleash a century of controversy and antitheoretical dogma within psychology that led Hilgard and Bower (1975) to write the following first sentence in their excellent review of *Theories of Learning:* "Psychology seems to be constantly in a state of ferment and change, if not of turmoil and revolution" (p. 2).

One illustrative type of psychological data that Helmholtz studied concerned color perception. Newton had noted that white light at a point in space is composed of light of all visible wavelengths in approximately equal measure. Helmholtz realized, however, that the light we perceive to be white tends to be the average color of a whole scene (Beck, 1972). Thus perception at each point is nonlocal; it is due to a psychological process that averages data from many points to define the perceived color at each point. Moreover this averaging process must be nonlinear, since it is more concerned with relative than absolute light intensities. Unfortunately, most of the mathematical tools that were available to Helmholtz were local and linear.

There is a good evolutionary reason why the light that is perceived to be white tends to be the average color of a scene. We rarely see objects in perfectly white light. Thus our eyes need the ability to average away spurious coloration due to colored light sources, so that we can see the "real" colors of the objects themselves. In other words, we tend to see the "reflectances" of objects, or the relative amounts of light of each wavelength that they reflect, not the total amount of light reaching us from each point. This observation is still a topic of theoretical interest and is the starting point of the modern theory of lightness (Cornsweet, 1970; Grossberg, 1972a; Land 1977).

A more fundamental difficulty faced Helmholtz when he considered the objects of perception. Helmholtz was aware that cognitive factors can dramatically influence our perceptions and that these factors can evolve or be learned through experience. He referred to all such factors as ***unconscious inferences***, and developed his belief that a raw sensory datum, or ***perzeption***, is modified by previous experience via a learned imaginal increment, or ***vorstellung***, before it becomes a true perception, or ***anschauung*** (Boring, 1950). In more modern terms, sensory data activate a feedback process whereby a learned template, or expectancy, deforms the sensory data until a consensus is reached between what the data "are" and what we "expect" them to be. Only then do we "perceive" anything.

The struggle between raw data and learned expectations also has an evolutionary rationale. If perceptual and cognitive codes are defined by representations that are spread across many cells, with no single cell knowing the behavioral meaning of the code, then some buffering mechanism is needed to prevent previously established codes from being eroded by the flux of experience. It will be shown below how feedback expectancies establish such a buffer.

Unfortunately, Helmholtz was unable to theoretically represent the nonstationary, or evolutionary, process whereby the expectancy is learned, the feedback process whereby it is read out, or the competitive scheme whereby the afferent data and efferent expectancy struggle to achieve consensus. Helmholtz's conceptual and mathematical tools were linear, local, and stationary.

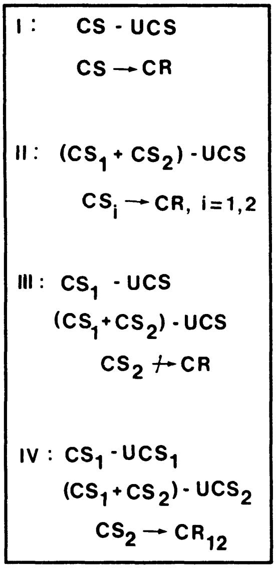
Section 4 begins to illustrate how nonlinear, nonlocal, and nonstationary concepts can be derived as principles of organization for adapting to a fluctuating environment. The presentation is nontechnical, but it will become apparent as we proceed that without a rigorous mathematical theory as a basis, the heuristic summary would have been impossible, since some of the properties that we will need are not intuitively obvious consequences of their underlying principles, and were derived by mathematical analysis. Furthermore, it will emerge that several design principles for adapting to different aspects of the environment operate together in the same structure. One of the facts that we must face about evolutionary systems is that their simple organizational principles can imply extraordinarily subtle properties. Indeed, part of the dilemma that many students of mind now face is not that they do not know enough facts on which to base a theory, but rather they do not know which facts are principles and which are epiphenomena, and how to derive the multitudinous consequences that occur when a few principles act together. A rigorous theory is indispensable for drawing such conclusions.

The next two sections summarize some familiar experiments whose properties will reappear from a deeper perspective in the thought experiment. These experiments are included to further review one of the themes that Helmholtz confronted, and to prepare the reader for the results of the thought experiment. The sections can be skipped on a first reading.

2. Overshadowing: A Multicomponent Adult Phenomenon With Developmental Implications

Psychological data are often hard to analyze because many processes are going on simultaneously in a given experiment. This point is illustrated below in a classical conditioning paradigm that will be clarified by the theoretical development. Classical conditioning is considered by many to be the most passive type of learning and to be hopelessly inadequate as a basis for cognitive studies. The overshadowing phenomenon illustrates the fact that even classical conditioning is often only one component of a multicomponent process in which attention, expectation, and other "higher order" feedback processes play an important role (Kamin, 1969; Trabasso & Bower, 1968; Wagner, 1969).

Consider the four experiments depicted in Figure 1. Experiment 1 summarizes the simplest form of classical conditioning. An unconditioned stimulus (UCS), such as shock, elicits an unconditioned response (UCR), such as fear, and autonomic signs of fear. The conditioned stimulus (CS), such as a briefly ringing bell, does not initially elicit fear, but after preceding the UCS by a suitable interval on sufficiently many conditioning trials, the CS does elicit a conditioned response (CR) that closely resembles the UCR. In this way, persistently pairing an indifferent cue with a



$$\text{I}: \quad CS - UCS$$
$$CS \rightarrow CR$$

$$\text{II}: \quad (CS_1 + CS_2) - UCS$$
$$CS_i \rightarrow CR, \; i = 1,2$$

$$\text{III}: \quad CS_1 - UCS$$
$$(CS_1 + CS_2) - UCS$$
$$CS_2 \not\rightarrow CR$$

$$\text{IV}: \quad CS_1 - UCS_1$$
$$(CS_1 + CS_2) - UCS_2$$
$$CS_2 \rightarrow CR_{12}$$

Figure 1. Four experiments illustrate overshadowing. (Experiment I summarizes the standard classical conditioning paradigm: conditioned stimulus–unconditioned stimulus [CS–UCS] pairing enables the CS to elicit a conditioned response (CR). Experiment II shows that joint pairing of two CSs with the UCS can enable each CS separately to elicit a CR. Experiment III shows that prior CS_1–UCS pairing can block later conditioning of CS_2 to the CR. Experiment IV shows that CS_2 can be conditioned if its UCS differs from the one used to condition CS_1. The CR that CS_2 elicits depends on the relationship between both UCSs, hence the notation CR_{12}.)

significant cue can impart some of the effects of the significant cue to the indifferent cue.

In Experiment 2, two CSs, CS_1 and CS_2, occur simultaneously before the UCS on a succession of conditioning trials; for example, a ringing bell and a flashing light both precede shock. It is typical in vivo for many cues to occur simultaneously, or in parallel, and the experimental question is, Is each cue separately conditioned to the fear reaction or is just the entire cue combination conditioned? If the cues are equally salient to the organism and are in other ways matched, then the answer is yes. If either cue CS_1 or CS_2 is presented separately after the conditioning trials, then it can elicit the CR.

Experiment 3 modifies Experiment 2 by performing the conditioning part of Experiment 1 on CS_1 before performing Experiment 2 on CS_1 and CS_2. In other words, first condition CS_1 until it can elicit the CR. Then present CS_1 and CS_2 simultaneously on many trials using the same UCS as was used to condition CS_1. Despite the results of Experiment 2, the CS_2 does not elicit the CR if it is presented after conditioning trials. Somehow prior pairing of CS_1 to the CR "blocks" conditioning of CS_2 to the CR.

The meaning of Experiment 3 is clarified by Experiment 4, which is the same as Experiment 3, with one exception. The UCS that follows CS_1 is not the same UCS that follows the stimulus pair CS_1 and CS_2 taken together. Denote the first UCS by UCS_1 and the second UCS by UCS_2. Suppose, for example, that UCS_1 and UCS_2 are different shock levels. Does CS_2 elicit a CR in this situation? The answer is yes if the two shock levels are sufficiently different. If the shock UCS_2 exceeds UCS_1 by a sufficient amount, then CS_2 elicits fear, or a negative reaction. If, however, the shock level UCS_1 exceeds UCS_2 by a sufficient amount, then CS_2 elicits relief, or a positive reaction.

How can the difference between Experiments 3 and 4 be summarized? In Experiment 3, CS_2 is an irrelevant or uninformative cue, since adding it to CS_1 does not change the expected consequence UCS. In Experiment 4, by contrast, CS_2 is informative because it predicts a change in the UCS. If the change is for the worse, then CS_2 eventually elicits a negative reaction (Bloomfield, 1969). If the change is for the better, then CS_2 eventually elicits a positive reaction (Denny, 1970).

Thus many learners are minimal adaptive predictors. If a given set of cues is followed by expected consequences, then all other cues are treated as irrelevant, as is CS_2 in Experiment 3. Each of us can define a given object using different sets of cues without ever realizing that our private sets are different, so long as the situations in which each of us uses the object always yield expected consequences. By contrast, if unexpected consequences occur, as in Experiment 4, then we somehow enlarge the set of relevant cues to include cues that were erroneously disregarded.