THE EXPERIMENTAL PSYCHOLOGY SERIES

HUMAN ASSOCIATIVE MEMORY: A Brief Edition

JOHN R. ANDERSON
and GORDON H. BOWER

LEA

HUMAN ASSOCIATIVE MEMORY: A Brief Edition

THE EXPERIMENTAL PSYCHOLOGY SERIES

Arthur W. Melton • *Consulting Editor*

1972 MELTON and MARTIN • *Coding Processes in Human Memory*

1973 ANDERSON and BOWER • *Human Associative Memory*

McGUIGAN and LUMSDEN • *Contemporary Approaches to Conditioning and Learning*

1974 GARNER • *The Processing of Information and Structure*

KANTOWITZ • *Human Information Processing: Tutorials in Performance and Cognition*

KINTSCH • *The Representation of Meaning in Memory*

MURDOCK • *Human Memory: Theory and Data*

1975 LEVINE • *A Cognitive Theory of Learning: Research on Hypothesis Testing*

1976 ANDERSON • *Language, Memory, and Thought*

CROWDER • *Principles of Learning and Memory*

1977 STERNBERG • *Intelligence, Information Processing, and Analogical Reasoning: The Componential Analysis of Human Abilities*

1978 POSNER • *Chronometric Explorations of Mind*

SPEAR • *The Processing of Memories: Forgetting and Retention*

1979 KIHLSTROM and EVANS • *Functional Disorders of Memory*

HUMAN ASSOCIATIVE MEMORY: A Brief Edition

JOHN R. ANDERSON
Carnegie-Mellon University

GORDON H. BOWER
Stanford University


LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS
1980 Hillsdale, New Jersey

This book is dedicated to our wives

Sharon Anthony Bower

and

Lynne Marie Reder

Lawrence Erlbaum Associates, Inc., Publishers
365 Broadway
Hillsdale, New Jersey 07642

Library of Congress Cataloging in Publication Data

Anderson, John Robert, 1947–
Human associative memory.

Bibliography: p.
Includes indexes.
1. Memory. 2. Association of ideas. I. Bower, Gordon H., joint author. II. Title.
BF371.A53 1980 153.1'2 79-28349
ISBN 0-89859-020-5

Printed in the United States of America

CONTENTS

PREFACE TO BRIEF EDITION

We wrote the HAM book six years ago. Six years provides some perspective on one's work. This perspective has led us to construct this new edition of the book. As the reader should be able to tell by comparing the thickness of the two books, the principle transformation between the original and the revision has been deletion. We have dropped some sections that are simply out of date. Others were dropped that seem, in retrospect, not very essential to our message in the original book. Other deletions involve analyses that either proved to be far off target or have been replaced by substantially better analyses. What remains (approximately 50% of the original text) is either what we feel still committed to or, if not that, what we feel is important as background for understanding current important issues in the literature. A major goal of this edition is to make the more important points of the original HAM book available at a more economical price. If someone needs the deleted analyses and experiments, the full version of the book remains available.

This edition contains two major parts. First is the historical analysis of associationism and its countertraditions. This still provides the framework that we use to relate our current research to an important intellectual tradition. This is reproduced without comment from the original book; historical analyses do not need as rapid revision as theoretical analyses.

The second part of the book reproduces the major components of the HAM theory. As we see it today, the major contribution of that theory was the propositional network analyses of memory and the placement of those representational assumptions into an information-processing framework. We have reproduced our specification of the HAM representational assumptions. Although there are problems with certain specifics of this representation, we feel content with most of it.

Also, several assumptions about the processes that operate on this memory representation still seem like good ideas. One is our idea about how pattern recognition operated in such a system. The conception we developed in HAM conceived of recognition as pattern matching. This conception has become even more important in subsequent theoretical developments (e.g., JA's ACT theory). Closely related is the idea that retrieval from long-term memory takes place through this

graph matching process. Although certain essential details have proven wrong, many aspects of this retrieval model are with us still, and even our inadequate assumptions serve as important starting points for organizing current research.

Another major contribution of HAM concerned its analyses of how sentential and other factual knowledge is learned. Although it has attracted a fair number of alternative proposals, our stochastic model of sentence learning is still a real contender. Further, we provided an analysis of many of the traditional analyses of imagery and forgetting. We have changed our minds on some specific claims, but these analyses are still the basic way we conceive of verbal learning phenomena.

Certain aspects of the HAM theory presented here have proven to be incorrect or, if not incorrect, are in considerable dispute. We have appended a new chapter to review briefly these aspects and the issues they raise.

So, in summary, we have tried to use the passage of time to give a more focused rendition of the HAM book. What remains in this book represents, in our mind, what is still important and significant.

Preparation of this revision was supported by grant BNS-78-17463 from the National Science Foundation to John Anderson and by grant MH-13905 from the National Institute of Mental Health to Gordon H. Bower.

John R. Anderson and Gordon H. Bower

PREFACE TO THE FIRST EDITION

This book proposes and tests a theory about human memory, about how a person encodes, retains, and retrieves information from memory. The book is especially concerned with memory for sentential materials. We propose a theoretical framework which is adequate for describing comprehension of linguistic materials, for exhibiting the internal representation of propositional materials, for characterizing the "interpretative processes" which encode this information into memory and make use of it for remembering, for answering questions, recognizing instances of known categories, drawing inferences, and making deductions. This is all a very tall order, and we shall be gratified if a fraction of our specific hypotheses prove adequate for long. However, what is more significant is the overall framework and theoretical methodology within which specific hypotheses are cast: we sincerely hope that this framework would have a singular value that would outlive its specific details.

How have we arrived at the theoretical framework to be proposed? We will answer this question at two levels–first, in terms of a brief autobiography; second, in terms of a broader historical context. When the first author (JA) arrived at Stanford University as a graduate assistant to the second author (GB), there was an ongoing research program concerned with organizational and imaginal factors in various memory tasks. As we tried to become precise, even quantitative, in fitting organizational theory to free recall data, its difference from associationistic models of free recall seemed to evaporate, frankly because neither theory had been formulated with any real precision up to that time. Eventually, JA developed a semi-successful computer simulation model of free recall, FRAN; however, the data base of FRAN (or its memory representation) was fundamentally associationistic in character.

The problem with FRAN, as with other free recall models, is that it could not understand language: it treated a sentence as though it were a string of unrelated words. Consequently, it was decided to put FRAN aside and to search for a theory and model that would be able to represent the information in sentences and describe how they are learned and remembered. This required that both of us learn

a fair amount of linguistics, psycholinguistics, and computational linguistics, a task in which we were aided by Herbert Clark and Roger Schank of the Stanford faculty. We had also begun some empirical investigations of sentence memory, expecting to find support for a Gestalt-like theory but instead finding associationist-like phenomena (these are reviewed here in Chapter 11).

The outcome of rather intensive ruminations and discussions was the theory, HAM (for Human Associative Memory) proposed herein. This was first worked out in detail in a long "dissertation proposal" by JA which had several goals: to present an associative theory of sentence memory, to report evidence relevant to it, to relate the theory to the historical tradition of associationism, and to indicate how a few standard "verbal learning" phenomena might be interpreted in terms of this approach. That document formed the basic outline for this book. The language parser and question-answerer of HAM were written as a LISP program by JA, and its operation is illustrated in Chapter 6 here. That proposal led us into a productive set of discussions and experiments, many of which are scattered throughout this book. Given the volume of results and the number of things we wanted to say about them, it became clear that a book rather than piecemeal publications was the appropriate way to communicate the theoretical framework and its supporting evidence.

In the Spring of 1972, we began collaborative writing of this volume; each day was filled with hours of fruitful discussions followed by our individual writing efforts. In these discussions we came to adopt characteristic roles–JA as the proposer, interpreter, and defender of HAM, and GB as the critic, provider of more problems, the demander of greater generality. However, like most fruitful interchanges, ours were free-wheeling, and we adopted various roles as the occasion demanded. Only a fraction of the analyses and problems solved appear in these pages. The discussions and writing turned out to be both personally and intellectually the most gratifying moments of our collaboration.

Now, let us briefly indicate the historical context of this work. First, our work falls within the tradition of philosophical associationism, which stretches from Aristotle through the British empirical philosophers to current psychology. We found so much of value in that rich intellectual tradition that we felt honor-bound to cite chapter and verse from it to show its contemporary relevance. This we do in Chapter 2, along with criticizing that tradition and anticipating how our theory of memory differs from it.

Second, this work owes a special debt to those scientists doing research on human memory, both researchers from the "verbal-learning" tradition and those using the "organizational" approach. Chapters 14 and 15 here explicitly deal with the verbal learning literature, whereas the influence of the organizational approach to memory should be apparent in chapters 3, 8, 11, 13, and 14.

The third intellectual tradition impinging on our research is the theoretical work in modern linguistics, especially that on transformational generative grammar of Noam Chomsky, his associates, and the whole movement he has promulgated. Although we deal with models of linguistic *performance* for only limited domains, we are nonetheless indebted to the formal analyses of the linguists for suggesting these models. Linguistic theories are reviewed in Chapter 5 and issues concerning

the representation of propositional information occur repeatedly throughout our work (e.g., Chapters 7, 8, 9, 11, 13).

Our final intellectual debt is to the research workers in artificial intelligence, to those like Minsky, McCarthy, Newell, and Simon who have shaped the conceptual development of that entire area, but more specifically to those who have dealt with computer models for natural language understanding and for question-answering. A review of language understanding programs (those of Woods, Winograd, and Schank) is contained in Chapter 5, and a review of models for "long-term semantic memory" (specifically, Quillian's and that of Rumelhart, Lindsay, and Norman) is contained in Chapter 4. Our theoretical framework has a special likeness to that being developed by Rumelhart, Lindsay, and Norman at the University of California at San Diego, and that developed by Walter Kintsch at the University of Colorado. It is indicative of the *Zeitgeist* that our work was begun independently and in relative ignorance of theirs, and only later did we become acquainted with the details of their approach. Special visits to La Jolla and Boulder provided us with detailed information about their theoretical projects, and we are pleased to have this opportunity to thank these scientists and their research students for their intellectual help, encouragement, and hospitality.

These four distinct areas, then, provide the intellectual and historical backgrounds for our theory of human memory. As does every lengthy research project or book, ours has accumulated a number of specific debts to individuals who have helped bring this enterprise to fruition. First, we acknowledge the general support of the faculty and graduate students in cognitive psychology at Stanford University; the general climate of intellectual stimulation there clearly provided the reinforcing and educational contingencies needed to initiate, encourage, and maintain our theoretical enterprise. We appreciate those colleagues–Arnie Glass, Steve Kosslyn, Perry Thorndyke, and Keith Wescourt–who allowed us to report their previously unpublished experiments. Ed Feigenbaum was very helpful in our development of the simulation program and provided us with help when the simulation began to exceed the capabilities of the campus facility.

We solicited and received constructive comments from many colleagues, and the final version of the book is clearly better because of them. Bob Crowder, Jim Greeno, Reid Hastie, Marcel Just, Steve Kosslyn, Alan Lesgold, Elizabeth Loftus, Gary Olson, Lance Rips, Ed Smith, Dave Tieman, and Wayne Wickelgren all have commented on portions of the book. To them we give our thanks. A special note of thanks goes to Lynne Reder who read the book in its entirety and pointed out passages in need of better exposition.

The research reported here and preparation of the manuscript was supported through a research grant to GB, number MH-13950, from the National Institutes of Mental Health. We are pleased to acknowledge Drs. George Renaud and John Hammack of the NIMH staff for their helpful encouragement and support of this research and of its writing. Yale University also helped JA's writing by easing the burden on its new assistant professor and by making resources available for him to supervise the book through its final draft. During the final revisions, GB was supported by the Center for Advanced Study in the Behavioral Sciences and by research funds from NIMH.

We owe a special thanks to the several individuals who have been closely involved with the physical preparation of the manuscript. First among these is Joyce Lockwood, GB's secretary, who typed the first one and a half versions of the book, making sense out of our scrawls while exhibiting patient forebearance in the face of a frustrating barrage of corrections to corrections. The final one and a half versions of the manuscript were typed by JA's secretaries at Yale, Barbara Psotka and Glenna Ames. We appreciate the swift and reliable clerical help they have provided to us. A special thanks also goes to Larry Erlbaum, our publisher, for providing moral support as well as expediating those technical matters associated with shepherding a manuscript through to publication. Finally we are obliged to several authors and publishers who gave permission to quote or to reproduce figures from their publications, and we have acknowledged their contributions in the appropriate places of the text.

John R. Anderson and Gordon H. Bower

March, 1973

1
INTRODUCTION

And I gave my heart to seek and search out by wisdom concerning all things that are done under heaven; this sore travail hath God given to the sons of men to be exercised therewith.

–Solomon

Two years ago, we set out to develop a theory of human memory, a theory which was to span a wide range of mnemonic phenomena. We are now humbled by the immensity of this task; human memory is a complex mental capacity, and our ability to comprehend man's mind appears at times quite limited. But Solomon calls us to the task of understanding, to be "exercised" by its sore travail. And so we tried. In countless hours of conversations, we discussed, proposed, role-played, argued, laughed, cajoled, reasoned, debunked, and just plain talked to one another about the problems of human memory. The time has come for us to commit to print a fraction of the things we have thought about human memory in the hope of helping others to think about this problem–which we consider to be the supreme intellectual puzzle of the century.

The theory of human memory which we will articulate will seem overly ambitious but still terribly programmatic; no one can realize this better than we ourselves. So why bother? What does Psychology need with another fragmentary theory of memory? After all, a long parade of memory theories since Plato's have been offered with great fanfare, hopeful enthusiasm, and persuasive arguments. Most of these were soon consigned to the loneliness of library tombs, accumulating dust to hide their insignificance. A very few of these writings become classics. But no one really believes the classics; they are read only to provide jousting partners for later opponents and voyagers on the seas of the unknown.

It is commonplace that the Zeitgeist in current psychology opposes global theories such as the one to be presented. It is said, instead, that one ought to work

on limited hypotheses for small, manageable problems–categorization effects in free recall, verification latencies for negative sentences, search of items in short-term memory, and so on ad infinitum. Indeed, we have been told by many respected colleagues in psychology that we will surely fail because we "are trying to explain everything." Of course, we are not. Human memory is but a very small part of the psychological domain. To make a salient contrast, a criticism we are apt to receive from colleagues outside of psychology (e.g., artificial intelligence) is that we are far too narrow in our perspective and aims.

The reason for writing a theoretical book on human memory is the belief that we have something important to offer. In rejecting the earlier global theories, modern research on human memory has overreacted to the opposite extreme; it has become far too narrow, particulate, constricted, and limited. There is no overall conception of what the field is about or even what it should be about. There is no set of overarching theoretical beliefs generally agreed upon which provide a framework within which to fit new data and by which to measure progress. Were we describing an unhappy personality, we would say that the contemporary study of memory has lost its sense of direction, its sense of purpose, and it is drifting aimlessly with much talent but little focus. This point was stated forcibly by a recent, informed but highly critical review of the field (Tulving & Madigan, 1970).

1.1. CONCERN FOR SUFFICIENCY CONDITIONS

Laboratory studies of memory appear under the inexorable control of a distinct set of "experimental paradigms," a standard set of "tasks," which seem by their nature to spew out an unending string of methodological variations and empirical studies. But the phenomena studied are becoming further and further removed from the manifestations of memory in everyday life. There would be nothing necessarily wrong with this esoterica provided psychologists had some clear conception of how their research and theories would eventually fit together into a system adequate to explain the complexities of everyday human memory. But, on the contrary, it appears that we psychologists are totally unconcerned about having our psychological theories meet certain *sufficiency conditions.* It is not enough that a theory make adequate ordinal predictions for a particular situation and experiment; in addition, it should be shown that its principles are sufficient to play a part in the explanation of the total complexity of human behavior. For instance, one could require of a model of memory that it be sufficiently powerful to succeed in simulating question-answering behavior.

When we began to concern ourselves with sufficiency conditions, we were forced to fundamental reconceptualizations regarding the nature of memory. We found that memory could no longer be conceived as a haphazard jumble of associations that blindly record contiguities between elements of experience. Rather, memory now had to be viewed as a highly structured system designed to record facts about the world and to utilize that knowledge in guiding a variety of performances. We were forced to postulate entities existing in memory which have no one-to-one correspondence with external stimuli or responses. As discussed in Chapter 2, such

structures violate the Terminal Meta-Postulate of classical associationism and stimulus-response psychology. It also became necessary to postulate the existence in the mind of highly complex parsing and inferential systems which function to interface the memory component with the external world. Furthermore, we were forced to postulate the existence of innately specified ideas in the form of semantic primitives and relations. We will therefore be proposing and arguing for a radical shift from the associationist conceptions that have heretofore dominated theorizing on human memory.

This shift is most apparent in the unit of analysis which we adopt. Unlike past associative theories, we will not focus on associations among single items such as letters, nonsense syllables, or words. Rather, we will introduce *propositions* about the world as the fundamental units. A proposition is a configuration of elements which (*a*) is structured according to *rules of formation*, and (*b*) has a *truth value.* Intuitively, a proposition conveys an assertion about the world. The exact structural properties of our propositional representation will be set forth in Chapter 5. We will suppose that all information enters memory in propositional packets. On this view, it is not even possible to have simple word-to-word associations. Words can become interassociated only as their corresponding concepts participate in propositions that are encoded into memory. However, propositions will not be treated here as unitary objects or Gestalt wholes in memory having novel, emergent properties. Rather, propositions will be conceived as structured bundles of associations between elementary ideas or concepts. However, our insistence that all input to memory be propositional imposes certain *well-formedness* conditions on the structure of the interidea associations. This notion of structural well-formedness is one that was completely lacking in past associative theories and was at the heart of many rationalist attacks on associationism.

1.2. NEO-ASSOCIATIONISM

We shall use the term "neo-associationism" to denote this new conception of human memory. While it introduces substantial deviations from past associationist doctrines, it still maintains a strong empiricist bias. We feel that the full significance of these theoretical assumptions can only be appreciated when one understands the associationist tradition out of which they came. Therefore, we have devoted Chapter 2 to an analysis of the associative tradition that extends from Aristotle through current American psychology. We will argue that a defining feature of associationism has been its *methodological empiricism.* That is, all associationists have accepted as their task the job of taking the immediate sense-data available to them and constructing their theory directly from these, always letting the data dictate the nature of the theory. This is contrasted in Chapter 3 with the *methodological rationalism* which attempts to first arrive at abstract, sufficient conditions, or constraints for the phenomena at hand, and then tries to relate these abstractions and conceptual constraints to the empirical world.

The contrast we are making between methodological empiricism and methodological rationalism corresponds (not surprisingly) to the more frequently made distinction between empiricism and rationalism. In the strong version of

empiricism, the mind begins as tabula rasa, and all knowledge is a consequence of the passive encoding of experience. The strong version of rationalism claims that the mind begins highly structured and all significant knowledge derives from the mind's initial structure. According to the rationalists, the role of experience is simply to stimulate the mind to derive that knowledge. Methodological empiricism and rationalism are not concerned with the origins of human knowledge, but rather with procedures for developing a scientific theory. However, we can almost derive a definition of each by substituting "scientific theory" for "mind" in the above statements of empiricism and rationalism. That is, methodological empiricism claims a scientific theory can be built up from immediate data by the blind procedure of generalization; whereas, methodological rationalism insists the theory builder must bring the essential structure of the theory to the phenomena to be explained.

Neo-associationism represents a profane union of these two methodologies. There is no attempt at a "creative synthesis" of these two positions; we simply pursue both methods in parallel in constructing a theory. The result is a theory that irreverently intermixes connectionism with nativism, reductionism with wholism, sensationalism with intuitionism, and mechanism with vitalism. Depending on the theoretician's propensities, the mixture can be claimed to be either more rationalist than empiricist or vice versa. The mixture we will offer is still strongly empiricist, much more so than the other neo-associationist theories that we will examine.

The various neo-associationist theories of memory (e.g., Simon & Feigenbaum, 1964; Collins & Quillian, 1972b; Rumelhart, Lindsay, & Norman, 1972), including our own, have been cast in the form of computer simulation models of memory. This is no accident. The task of computer simulation simultaneously forces one to consider both whether his theory is sufficient for the task domain to be simulated and also whether it can deal with the particular trends found in particular experiments.

Our therorizing and experimentation are specifically oriented towards memory for linguistically structured material. With such interests, one cannot help but make constant contact with the recent ideas in linguistics. The linguistic work, particularly of Chomsky, Fillmore, Lakoff, Katz, Ross, and their associates, is important for a second reason. These linguists have argued effectively for the importance of sufficiency conditions in linguistics. As a consequence, over the past decade rationalism and mentalism have become strongly entrenched in linguistics. The rationalist "revolution" has been imported from linguistics into psychology. Thus, the developments in linguistics are an important source behind the neo-associationist developments.

Two substantial chapters are being devoted to an extensive historical and theoretical review of efforts related to our own. This is clearly out of character for a typical "research volume." The usual practice for American psychology is to restrict its focus to the last 5 or 10 years of experimentation centered around a narrowly circumscribed topic. This practice is lamentable since true scholarly endeavor would seem to require an appreciation of the historical and intellectual context within which that scholarship occurs. Without knowledge of that context, it is not possible to discriminate between significant theoretical advance as opposed

to elaboration of an established paradigm. Chomsky (1968) has argued persuasively for a similar historical perspective in linguistics.

Our work began in the typical intellectual isolation of experimental psychology, but we constantly found ourselves being led into discussions of issues about which we know very little. Therefore, we have tried to trace the connections between our work and that which had occurred in past centuries or which was occurring in related fields. Our perception of what questions were important changed; similarly, the character of the theory and research to be presented is very different from what we had originally projected and from the typical fare that one finds in psychology. It can only be appreciated in the perspective of the historical context that we set in the first two chapters. One of the incidental advantages of a theory so constructed is that it provides the reader with an integrated viewpoint from which to perceive his own experimental research, related research in psychology and other fields, and the relationship between this research and what has happened in past centuries.

Following these two review chapters, the remainder of the book serves as a forum for presenting our theory and research. We have many experiments to report that have not appeared before in print. We will also review and comment upon a large number of recent experiments that seem particularly interesting with respect to the issues that we are raising. Although there is no attempt to review extensively the literature in human memory, we do hope to establish theoretical connections among many different areas of experimentation in psychology.

To preview the contents of the later chapters, Chapter 4 provides a general overview of our model of long-term memory. We have christened the model HAM, an acronym for Human Associative Memory. The subsequent three chapters set forth most of the substantive theoretical assumptions of that model. The character of presentation varies considerably from one chapter to the next. In Chapter 5, entitled "The Structure of Knowledge," we propose a structure in which information will be encoded and stored in long-term memory. In Chapter 6 we will ask how the memory system recognizes that it has experienced something before. This issue, of how current stimuli contact old traces, is a point of notorious difficulty for other accounts of memory. Finally, in Chapter 7 we will present a stochastic model of how incoming information is encoded into long-term memory.

The remaining chapters will be concerned with relating our theory to various areas of research and experimentation. In Chapter 8 we will examine the question of how long-term memory is searched for information, to decide whether or not some fact is known or some statement is true. This is the problem of *fact retrieval.* In Chapter 9, we will discuss how our model would perform in the typical verbal learning paradigms such as paired-associate learning and free recall. Finally, in Chapter 10 we will discuss how different information inputs interfere with one another to produce forgetting. We will compare our model of this process with past theories of interference and forgetting.

1.3. THE FUNDAMENTAL QUESTIONS

There are well-known advantages to vagueness in constructing a theory; it protects the theory from disconfirmation. The typical strategy is to articulate the

theory at those points where it makes contact with confirming evidence, but otherwise to shroud it in sufficient vagueness so that any other present or future data cannot unambiguously disconfirm the model. We have tried to avoid this tactic. Not only is our theory vulnerable to future disconfirmation; it also clearly fails to handle a number of the existing facts. The points of misfit will be openly acknowledged at the appropriate places. It is difficult to determine how serious these failures are. In a complex model like HAM it is always possible to introduce some special assumption that will handle any particular discrepancy. Also, the misfits may indicate a mistake in one particular assumption rather than a flaw in the grand theoretical design.

The fundamental issue at stake with respect to our theory is its neo-associationist character. This is not to be found in any particular assumption, but rather pervades diffusely throughout the whole enterprise. Our strong computer-simulation orientation has led to a class of controversial assumptions. Information processing in HAM tends to be in terms of discrete units called ideas and associations, and it proceeds in sequential steps, whereas parallel, interactive processes are assumed to be minimal. Can one really claim that a human processes information in this discrete, serial manner? But the physiology of the brain is very different from that of a serial, digital computer, and analogue, parallel processes would not seem out of character for that mysterious organ (cf. Von Neuman, 1958). Perhaps, then, our theory has been too strongly determined by what is easy to simulate on a computer rather than by considerations of psychological plausibility. That is one fundamental question.

Another source of difficulty with our theory may arise from our strong empiricist leanings. We have insisted that all knowledge in memory should be built up from input to the memory. We have denied that memory has any capacity to spontaneously restructure itself into more useful forms. Perhaps we have made memory too passive, too much of a tabula rasa. That is a second fundamental question.

On the other hand, we have granted the mind a great deal of self-structuring power in our assumptions regarding the perceptual parsers that transform external stimuli into memory input, or the various inferential and problem-solving abilities that enable the system to make intelligent use of the information recorded in memory. One is forced to postulate such powerful mechanisms in order to interface a memory with the world. The postulated mechanisms are enormously more complicated than any of the theoretical devices that have been previously postulated in associative theories. Perhaps, if we had complicated the proposed memory system, we could have simplified the interfacing apparatus. That is a third fundamental question.

Another possible flaw in the grand design has to do with our insistence on making the propositional representation fundamental. We will want to encode perceptual as well as linguistic input into this uniform propositional base. Perhaps we are choosing a representation that is too logical and abstract. Perhaps the primary representation of knowledge is of some diffuse, sensory sort; and our ability to encode information propositionally in this original base comes about only

after much conceptual development and training in abstraction. This is a fourth fundamental question.

These are the sorts of questions that will hound us throughout this enterprise. We cannot claim that there is any great initial plausibility to our particular formulation. But we feel it is important that we develop that formulation as explicitly as possible and raise the questions we have about it. Our formulation provides a concrete realization of a certain theoretical position. It provides something definite for research workers to discuss, examine, criticize, and experiment upon. It is hoped that some resolution will be eventually achieved with respect to the fundamental theoretical questions. We hope that others will be encouraged to provide and motivate other explicit models of fundamentally different theoretical positions. If this happens, our goal will have been achieved, whatever is the final judgment with respect to HAM. We will have shifted the focus of experimental psychology from the articulation of narrow paradigms to an analysis of the significant questions concerning human memory. To attempt this may be a pretentious ambition, but it is a primary purpose and justification of this book.

2
ASSOCIATIONISM: A HISTORICAL REVIEW

In our inquiry into the soul it is necessary for us, as we proceed, to raise such questions as demand answers; we must collect the opinions of those predecessors who have had anything to say touching the soul's nature, in order that we may accept their true statements and be on our guard against their errors.

–Aristotle

2.1. ASSOCIATIONISM: AN OVERVIEW

Associationism has a tradition that extends over 2,000 years, from the writings of Aristotle to the experiments of modern psychologists. Despite the existence of this clearly identifiable *theoretical tradition*, there is not a well-defined monolithic *theoretical position* which can be called associationism. Past associative theories differ one from another both in details and in basic assumptions. While all major associative theorists have agreed on a few fundamental points, there are more fundamentals on which there exist no such consensus. So, we are faced with an apparent paradox: How can we identify a coherent associative tradition but no coherent associative theory?

The unifying feature of associationism lies in its empiricist methodology, not in any substantive assumptions that it makes. That is, all associationists have taken as their task the job of using the immediate data available to them (e.g., introspections, stimulus-response contingencies, etc.) and constructing the human mind from these with minimal additional assumptions. Depending upon the data they considered important and upon personal idiosyncracies, different theorists achieved somewhat different mental reconstructions. However, because of the common methodology, their psychological systems tend to share certain metafeatures. Four such features seem to universally typify associationism:

1. Ideas, sense data, memory nodes, or similar mental elements are associated together in the mind through experience. Thus, associationism is *connectionistic.*

2. The ideas can ultimately be decomposed into a basic stock of "simple ideas." Thus, associationism is *reductionistic.*

3. The simple ideas are to be identified with elementary, unstructured sensations. (The meaning we want to assign to "sensation" is rather generous in that we intend to include internal experiences, such as involved in emotion.) Because it identifies the basic components of the mind with sensory experience, associationism is *sensationalistic.*

4. Simple, additive rules serve to predict the properties of complex associative configurations from the properties of the underlying simple ideas. Thus, associationism is *mechanistic.*

We claim that these four features of associative theories are defining features of associationism because they are the highly probable consequences of the empiricist methodology that constructs such theories.

It might seem that the empirical validity of these four metafeatures might then be crucial to evaluating associationism. If one of these assumptions were to be proven false, that would prove that associative theories are wrong. However, it is doubtful whether any of these assumptions is of the sort that it could be subject to empirical falsification. After all, they are metafeatures of the theory rather than definite predictions about observable behavior. These metafeatures become manifest in particular theories in the form of particular predictions that may be falsified or verified, but it seems that the metafeatures are not subject to empirical disconfirmation. But before pursuing this point further, we should examine the four meta-assumptions in more detail.

Connectionism. Regarding connectionism, one must distinguish whether the discussion concerns associationism as a theory of human memory or as a theory of all mental phenomena. Connectionism, with its implicit empiricism, is a controversial assumption within the general associative plan of trying to explain all mental processes with one basic principle. It is not at all obvious that all our mental processes have been connected together through experience. Indeed, in our own model we do not subscribe to the notion that the mind has been totally "wired up" by experience. Some of the important mental processes described in our model are much more naturally viewed as innate rather than acquired mechanisms.

In contrast to the doubtful character of connectionism as the universal principle of mental phenomena, it would seem entirely innocuous as a principle of memory. To say that memory consists of ideas connected by experience would seem to be almost tautological. In this respect, it is interesting to note the uncritical tendency among psychologists to apply the "associationistic" label to any theory of memory that refers to connections or associations. This practice is reducing associationism to an empty descriptive notion. Associationism as a theory of memory gets its cutting edge from the remaining three distinguishing features. They serve to impose some restrictions on the character of the "connections."

Reductionism. This is sometimes called "elementarism," and the doctrine is fairly clear: It is assumed that there are certain elements (the simple ideas) that are

distinguished by the fact that all other ideas are built up from them. The phrase "built up from" is somewhat vague, but a formalization of that notion will be offered in Chapter 5, where the memory structure of HAM is discussed. Some readers might question whether reductionism has any empirical significance; wouldn't every theory of human memory subscribe to such a metaprinciple? The answer is "Definitely not." The classic counterexample (although there are others) is Gestalt theory, which argued many phenomena defied reductionistic analysis (see Chapter 3).

Sensationalism. Certainly no one would quarrel with the claim that representations of sensory data constitute part of the contents of the mind. However, from Plato to Chomsky, there have been radical rationalists who have denied the sensationalist's claim that all knowledge has a sensory base. Indeed, it will turn out that a few non-sensory elements are required in our model.

Mechanism. The mechanistic feature of associative theories is at once the most imprecise and the most controversial. Stated crudely, the claim is that man is a machine and his nature and behavior are to be understood in mechanical terms. Ever since Democritus gave his original mechanistic account of the human soul, the issue has been a controversy of some stature. Since La Mettrie attempted to refute Descartes' claim that man is not machine, the matter has been a violent debate (witness the recent book by Dreyfus, 1972). The problem, however, is that our concept of what it is to be a machine is exceedingly imprecise and is continually being revised as we construct more intelligent automata. However, we do have reliable intuitions about what it means to be mechanistic. Many principles in our model will be unanimously judged as mechanistic and, no doubt, distasteful for that reason to some readers. Mechanistic assumptions such as those in our model tend to display an affinity for simple, linear, and discrete processes and an aversion for mass, interactive, and continuously varying processes.

We have argued that these four features have significance with respect to theories of the mind, although connectionism without the other three is an empty claim with respect to human memory. Nonetheless, it seems unlikely that any single feature can be subject to direct empirical falsification. There is a certain vagueness inherent in each of these metafeatures. One can empirically falsify the manifestation of these features in a particular model (for instance, our own), but it seems that it will always be possible to come up with another set of similar assumptions to explain the offending data. Indeed, much of the history of experimental psychology is the continuing saga of antiassociationists demonstrating the weakness of a particular associative theory, only to find the theory quickly changed and no longer subject to the old attack. This elusiveness of the four metaprinciples should not be surprising since they reflect methodological biases that are not really subject to empirical disconfirmation.

To summarize our conclusions, we claim that associationism is a historical tradition distinguished by its attempts to reconstruct the human mind from sensory experience with minimal theoretical assumptions. This approach contrasts with rationalistic theories which have attempted to work from basic a priori principles. As a consequence of its empiricist methodology, all associative theories have been distinguished by the four metafeatures enumerated above. While these metafeatures

can be manifested in empirical predictions that are subject to disconfirmation, the metafeatures themselves would seem fairly immune.

The Terminal Meta-Postulate

However, there is one feature which tends to haunt associative theories, which can be given precise statement, and which can be proven in error. This is the Terminal Meta-Postulate (TMP) which was so dubbed by Bever, Fodor, and Garrett (1968). This postulate should be viewed as a particularly likely manifestation of associationism's metafeatures. The postulate may be divided into three statements, one statement corresponding to each of three associative metafeatures.

1. *Sensationalist Statement.* The only elements required in a psychological explanation can be put into a one-to-one correspondence with potentially observable elements. These elements may themselves be observable stimuli or responses, or they may be derived from such observables. These derivatives have been variously known as intervening variables, mediating responses, sensations, perceptions, images, or ideas.

2. *Connectionistic Statement.* The elements in Statement 1 become connected or associated if and only if they occur contiguously.

3. *Mechanistic Statement.* All observable behavior can be explained by concatenating the associative links in Statement 2.

While many past associative theories have assented to the TMP, there are theories which are commonly agreed to be associative and which violate this principle. Many of the classical British associationists (see Section 2.3) admitted an irreducible principle of similarity and so would reject Statement 2, although all of the British associationists do seem to have accepted Statements 1 and 3. Aristotle (Section 2.2) rejected all three claims. We have followed his lead and have done likewise in our model. Therefore, to claim that the TMP is a defining feature of associationism, as some of our colleagues have, is just false.

Here we will illustrate a fundamental flaw in the TMP. In our demonstration we will be using the same example as employed by Bever et al. (1968). This is the mirror-image language which is typically employed as a structure that cannot be generated by *finite-state automata* or *regular grammars* (see Hopcroft & Ullman, 1969, for a formal exposition of such technical terms). Any mechanism satisfying the TMP cannot produce behaviors more complicated than can these formal automata or grammars. For instance, Suppes (1969) has established the equivalence between finite-state automata and S-R theory. Rather than becoming enmeshed in the formal theory of automata, however, we will try to make our points at a more conversational level.

In a mirror-image language, the sequence of elements in the first half of a string (or sequence) must be mirrored in the second half. For instance, if *a* and *b* were the only elements of the language, Table 2.1 gives examples of strings which are acceptable and strings which are not. Consider how a TMP system might try to deal with such a mirror-image language. Note that in grammatical strings, *a* can be followed by either *a* or *b*, and *b* can be followed by either *a* or *b*. Then, in accordance with Statements 1 and 2, we would have to postulate that the following

TABLE 2.1

A Mirror-Image Language

Grammatical	Ungrammatical
aa	ba
abba	aaab
bbbb	bbabaa
abaaba	bbabb

four associations are formed: $a \rightarrow a$, $a \rightarrow b$, $b \rightarrow a$, $b \rightarrow b$. These four associations do suffice to generate all the grammatical strings of our language. The first grammatical string in Table 2.1 could be generated simply with the association $a \rightarrow a$; the second could be generated by concatenating $a \rightarrow b$, $b \rightarrow b$, and $b \rightarrow a$, and so on for other strings. But the reader has probably already noted the difficulty. This TMP system generates too much. For instance, from the association $b \rightarrow a$ we can generate the ungrammatical string *ba*.

The basic problem with the TMP system is that it has no means of recording what it did early in the string so it can unwind the mirror image of that sequence in completing that string. To use a term popular in some psychological circles, a TMP system can have no "plan of action." It is easy enough to construct a system capable of generating all and just the strings of our mirror-image language. A context-free grammar to do this is given in Table 2.2. Also in Table 2.2 is a tree structure generated by the grammar. Bever et al. (1968) argue that it is the element X in the rewrite rule of the grammar which violates the TMP. However, this is to confuse a description of the formal grammar with the mechanism that implements the grammar. Nonetheless, when we examine such a mechanism we will find ample violations of the TMP.

However, before we turn to that mechanism, the reader should be clear about what is the problematical aspect of the mirror-image language in Table 2.1. The

TABLE 2.2

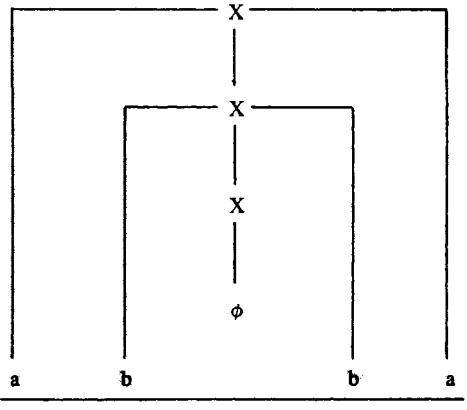
A Grammar for the Mirror-Image Language

Rewrite Rules

$X \rightarrow a\ X\ a$

$X \rightarrow b\ X\ b$

$X \rightarrow \phi$

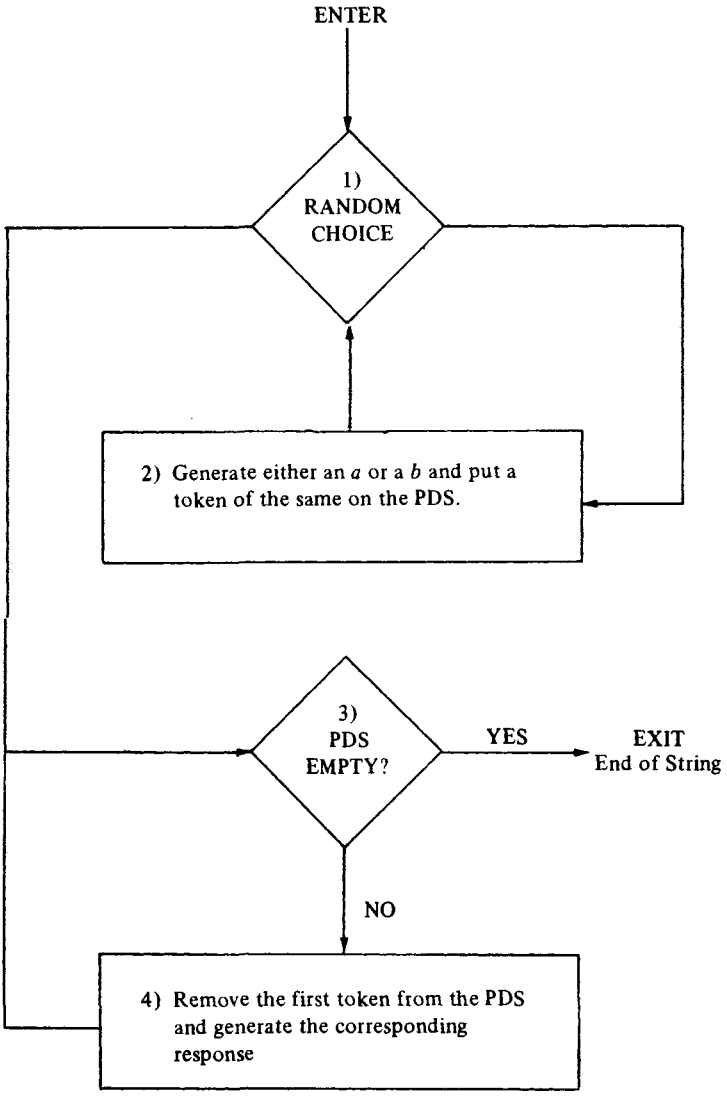


difficulty is that this language permits an indefinite number of embeddings of strings within strings. This embedding introduces dependencies between elements at arbitrary distances in the final string. For instance, the first and last element of a mirror-image string must match. Such dependencies cannot be captured in a finite-state automaton.

In Table 2.3 is the flow chart of a minimal system that is adequate to generate mirror-image languages. This system requires a push-down stack (PDS), a device

TABLE 2.3

Flow Chart of Push-Down Stack Machine for Mirror-Image Language



that stores objects and returns them according to the principle of last-in, first-out. This PDS clearly violates Statement 1 of the TMP. When we examine the flow chart, prescribing use of the PDS, we find further violations of the TMP. Consider how the string *abba* would be generated: The mechanism enters the random-choice box 1, and decides to move to box 2. Here it generates an *a* and puts an *a* token on the PDS. It cycles through the choice-box 1 and returns to box 2, this time to generate a *b*. Correspondingly, a *b* token is stored on the PDS. The contents of the PDS are now the tokens *b* and *a* in that order. The mechanism returns to choice-box 1 and then proceeds this time to decision-box 3. As the PDS is not empty, our mechanism proceeds to box 4, where it takes the first element from the PDS. This is a *b* token, and the system correspondingly generates a *b* response. It cycles through the choice-box 3 and back to box 4. Here it removes the last token from the PDS and generates an *a*. Upon returning to box 3, it finds the PDS empty and exits having successfully generated the string *abba*. The various operations we have been describing are clearly not encoded according to Statement 2 of the TMP or performed according to Statement 3. So we have violated all three conditions. Further violations of Statement 1 are the *a* and *b* tokens which were stored in the TMP. These *a* and *b* tokens are not responses nor response derivatives. Responses are not elements that reside for indefinite periods on push-down stacks. They rather occupy a brief moment in time and space.

Therefore, to generate the mirror-image language, we were forced to postulate a number of structures and processes only *abstractly* related to external observables. This is just what the TMP cannot abide. To end our discussion of the TMP on a technical note: In our "conversational" exposition we have been basically using the fact that a push-down automaton can and a finite-state automaton cannot accommodate the mirror-image languages. However, it is well known that any push-down automaton can be replaced by an equivalent finite-state automaton if we set a finite limit to the length of the push-down stack. A push-down automaton like that in Table 2.3 with a stack of length n can only generate (or recognize) mirror-image strings of length $2n$. But this is not objectionable, since there are certainly short-term memory limitations on the length of strings that we realistically can generate or recognize. Hence, it might be questioned whether we have shown the TMP to be in error, since there is a finite-state automaton that will handle mirror-image strings bounded by some upper length. However, this argument overlooks two important facts. First, the translation of a finite-stack push-down automaton into a finite-state automaton involves an enormous complication in terms of number of states. Essentially, each possible mirror-image string up to length $2n$ must be individually recorded by a distinct set of states. Secondly, by Statement 2 of the TMP, each transition in the state diagram must encode a contiguity in experience. But this is nonsense. We do not learn the mirror-image language by being exposed to all possible sequences of length up to $2n$; rather, a minute's study of the rules in Table 2.2 is sufficient. So this is one example of the importance of distinguishing questions of logical equivalence of two models from questions of their relative empirical plausibility. It is on the latter basis that we would reject any model based on the TMP.