THE UNIVERSITY OF ALBERTA

COMPUTATIONAL COGNITIVE MODELING OF CONCEPT ATTAINMENT

by

MICHAEL DAVID CARBONARO ©

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY

EDMONTON, ALBERTA

FALL, 1997

0-612-22960-2

THE UNIVERSITY OF ALBERTA

LIBRARY RELEASE FORM

NAME OF AUTHOR: Michael David Carbonaro

TITLE OF THESIS: Computational Cognitive Modeling of Concept Attainment

DEGREE: Doctor of Philosophy

YEAR THIS DEGREE GRANTED: 1997

11306-79 Ave
Edmonton, AB
Canada, T6G OP3

Date: Sept 2, 1997

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommended to the Faculty of Graduate Studies and Research, for acceptance, a thesis entitled Computational Cognitive Modeling of Concept Attainment submitted by Michael David Carbonaro in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

S. Hunka
S. Hunka, Supervisor

M. R. W. Dawson

RH Short
R. Short

Fern Snart
F. D. Snart

L. Stewin

D.S. per
J. H. Mueller

Date: July 10, 1997

Dedication

To Frank and Elisa

Abstract

Traditionally cognitive psychology has used a set of qualitative and quantitative procedures to formalize investigative strategies of cognitive processes. More recently the modeling of cognitive processes by computer simulation has become an important addition to these strategies by providing another set of investigative procedures for understanding cognition. The goal of this thesis is to extend our knowledge of the use of computational cognitive modeling to issues of concept attainment.

This thesis is divided into three major parts. Chapter 2 examines the role of computational cognitive modeling in researching cognitive processes. The evolution of the field of cognitive science and the major types of computational techniques used to model cognition are described in this Chapter. The Chapter concludes that computational cognitive modeling has the potential to provide a more detailed understanding of conceptual development, an area which is of special concern in educational psychology.

In Chapter 3 Bruner, Goodnow, and Austin's (1956) research on concept attainment is re-examined from a connectionist perspective. A network was constructed which associates positive and negative instances of a concept with their corresponding attribute values. Two methods were used to help preserve the *ecological validity* of the input: (a) closely mapping the input to the actual visual stimuli, and (b) structuring the output layer based on Gagne's (1962; 1985) work on human concept learning. This resulted in the addition of output units referred to as *attribute context constraints.* These units required the network to demonstrate the identification of attributes both relevant and irrelevant to the task of classification. Furthermore these additional output units guided the network in constructing a faster and more generalizable representation than those networks in which the constraints were absent.

Chapter 4 analyzes the internal structure of the connectionist networks constructed in Chapter 3 to model Bruner et al.'s (1956) concept attainment task. The networks' hidden

cellular responses were examined using graphical and singular value decomposition. This examination was carried out on the dynamic and final state data produced by the network. It was concluded that constrained networks learned a set of rules which produced greater discrimination among exemplars without any loss to correct categorization.

Acknowledgements

Many thanks to Dr. Steve Hunka who allowed me the privilege of being his graduate student. Without his support and dedication in the pursuit of knowledge this thesis would not have been possible! A special note of thanks to Dr. Mike Dawson whose guidance was invaluable and who first introduced me to the field of connectionist modeling. For their insightful suggestions and personal friendship thoughout my period as graduate student I would like to thank: Drs. Rob Short, Fern Snart, and Len Stewin. Thank you to Dr. J. H. Mueller for agreeing to be my external reader. I am also grateful to the Department of Educational Psychology for the financial support that provided me with the opportunity to complete this thesis. Finally, for their special support throughout the years I would like to thank Norma Nocente and my children Frank and Elisa.

Acknowledgements (cont)

Being Dr. Hunka's last "official" graduate student I would like to take this opportunity on behalf of all those who have worked under his guidance and supervision to offer this special thanks. During his last thirty six years Dr. Hunka has demonstrated the important contribution a University can make to our society. The term University has come under criticism recently for being a place too far removed from the pragmatics of daily life. Unlike many individuals Dr. Hunka has never lost sight of the reason this institution was created: a search for knowledge and the truth. His desire to learn was not only motivational but infectious.

The following is a list of students Dr. Hunka has supervised and institutions that they were associated with during their careers.

Year	Degree	Name	Thesis Title	Position
1963	MEd	Eastman, Mervyn Norman Guy	A semantic differential analysis of the concept school	
1964	MEd	Maguire, Thomas Owens	Component curve analysis of concept attainment	University of Alberta
1964	MEd	Quinn, Joseph William	An investigation of the validity and dimensionality of anxiety scales	(see below)
1965	PhD	Quinn, Joseph William	An investigation of personality and cognitive correlates of religious devoutness	Calgary Separate School Board
1966	PhD	Strutz, Peter George	A study of choice behavior of three age groups under three different treatments of a probability learning task	Loma Linda University and Medical Center
1967	PhD	Wahlstrom, Merlin Walter	A factor analytic item selection procedure	University of Toronto
1968	PhD	Carlson, James Eugene	Effects of differential weighting on the inter-reader reliability of essay grades	University of Iowa
1968	PhD	Fehlberg, Dieter August	Student achievement under Alberta's semester system	Alberta Education
1968	PhD	Phillips, Nicholas William	Personality correlates of cognitive styles	University of Winnipeg
1969	PhD	Brown, Kenneth Gordon	The relation between intelligence and achievement using computer-assisted instruction	University of New Brunswick
1969	PhD	Flathman, David Paul	List processing simulation of computer-assisted instruction	University of Calgary
1969	MEd	Hazlett, Clarke B.	The storage and retrieval of multiple choice items on computer	(see below)

1970	PhD	Powell, James Charles	A study of achievement information from the wrong answers given to multiple choice tests	University of Windsor
1970	PhD	Romaniuk, Eugene William	A versatile authoring language for teachers	University of Alberta
1971	PhD	Bay, Kyung Sun	An empirical investigation of the sampling distribution of the reliability coefficient estimates based on alpha and KR20 via computer simulation under various models and assumptions	University of Alberta
1971	PhD	Burnett. J. Dale	Component curve analysis of student performance on a computer-based simulation game	University of Lethbridge
1972	PhD	Hazlett, Clarke B.	Estimating construct validity in multiple choice, essay, and simulation graduate achievement examinations	Health Science, University of Alberta
1973	MEd	Boyle, John Ernest	Individualized intelligence testing by computer	Northern Alberta Institute of Technology
1973	PhD	Cartwright, Glenn F.	Social, personality, and attitudinal dimensions of individual learning with computer-assisted group instruction	McGill University
1973	PhD	Petruk, Milton W	The infrared computer based oculometer	University of Alberta
1978	PhD	Kearsley, Greg P.	A study of learner control in computer based instruction	George Washington University
1980	PhD	Harasym, Peter H.	The analysis of various techniques used for scoring patient management problems	University of Calgary
1981	PhD	Pagliaro, Louis Anthony	CAI in pharmacology: Student academic performance and instructional interactions	University of Alberta
1983	MEd	Garraway, Robert William Thomas	Microcomputer based computer-assisted learning system: CASTLE	(see below)
1983	MEd	Nesbit, John Cale	Approximate string matching applied to response analysis in computer assisted instruction	(see below)
1984	PhD	Harley, Dwight David	Simulated tailored testing of the CCAT	Edmonton Catholic School Board
1988	PhD	Nesbit, John Cale	Applications of learning hierarchies in adaptive instructional systems	Technical University of British Columbia
1990	MEd	Beaulne, Albert Leo	The effect of higher order latent spaces on the robustness of unidimensional nonlinear item response models	Workers' Compensation Board - Alberta

1990	PhD	Olson, Karin	Factors associated with the practice of breast self examination	Cross Cancer Institute University of Alberta
1992	PhD	Cameron, Judy	Intrinsic motivation revisited	University of Alberta
1992	PhD	Buck, George Henry	Instructional devices: Development, learning theories, and deployment	University of Alberta
1993	PhD	Garraway, Robert William Thomas	Hierarchical control and management in a CAI visual authoring environment	Bahai School of Canada
1994	PhD	Kivilu, Joseph Mbithi	Perceived causal attribution factors to academic performance: A multi-level analysis	Kenyata University
1994	PhD	Beckie, Theresa Marie	Quality of life after coronary artery bypass graft surgery	Florida State University
1995	MEd	Leighton, Jacqueline	The Johnson-Neyman method and "Mathematica"	University of Alberta
1997	PhD	Carbonaro, Michael David	Computational cognitive modeling of concept attainment	University of Alberta

Table of Contents

List of Figures

List of Tables

Chapter 1

1. Introduction

In this thesis three issues are examined with respect to computational cognitive modeling. The first issue concerns the implications such models have for research in learning and education. Central to this issue is the role of computational modeling in furthering our understanding of a given cognitive phenomenon under investigation.

The second issue addresses the actual building of such models. Computational cognitive models take a wide variety of forms (Simon & Halford, 1996). The model's architecture can greatly influence its performance and thus affect the results provided by the model. In contrast to the first issue discussed, where results from computational models are used to inform the researcher, it may be the case that existing theoretical knowledge (e.g., Gagne's (1984) instructional theory) can significantly contribute to the actual structural design of the model. Therefore, how a model is built is just as important as the reasons why such a model is built.

The final issue examined in this thesis concerns the internal representations generated by the computational model. Assuming that a particular cognitive phenomenon is chosen to be modeled (issue one) and given a particular modeling architecture (issue two), what do the representations produced by the model look like? In other words, at some level of analysis, the representations instantiated by the model must be interpreted and made explicit to the investigator. An understanding of the model's process of constructing this representation is therefore an essential part of the investigative procedure.

All three of the above issues are, to a large extent, issues of cognitive science. Furthermore they follow the historical trend in psychological research of formalization (Gardner, 1985; Johnson-Laird, 1988). The relationship between research in cognitive science and educational psychology is still in its infancy. The work in this thesis represents an effort to demonstrate that the relationship is maturing—especially with respect to the computational modeling of learning and development processes. What follows in this introduction is: (a) a general perspective on computational cognitive modeling, and (b) an explanation of the thesis format.

1.1 A general perspective on computational model building

Models may lead to theories or may be derived from theories but in all cases models (computational or otherwise) share the common goal of attempting to formalize an explanation of a phenomenon (Churchland, 1986). Cognitive models vary with the methods used to build them and to interpret their outcomes. Figure 1 shows a diagram that provides a general description of cognitive model building with respect to the model's degree of granularity. The concept of granularity refers to the increasing degree of precision and explicitness in which the phenomenon is described by the model. Precision means the accuracy of the explanation whereas explicitness indicates the clarity of the explanation. To some extent there is always a tradeoff between the two. Assuming it is possible to produce a function (computation) which describes an overall information processing system (IPS) (Turing, 1936)—such a functional description should not obscure the internal relationships between different states of the system (Marr, 1982). Although it might be argued that a mathematical description of a phenomenon represents the finest level of granularity, with respect to cognition, it can also be argued that such descriptions must be formalized as part of the total human information processing system (Klahr, 1992). Therefore mathematical descriptions embedded within an overall computational system are viewed as being a more precise and explicit method of modeling cognition.

Figure 1 is divided along three continuums. One continuum, **Y**, outlines general modeling approaches. At a coarse level of granularity modeling approaches take the form of verbal or written descriptions. Ascending on this continuum towards a finer level of granularity involves a process of numerical formalization and computational simulations. A physical (neurological) instantiation of the actual phenomenon represents the finest level of model building. Within each approach constraints on the modeling can take on many forms (environmental, biological, logical, mathematical). Regardless of which form, adherence to these constraints is intended to increase the model's validity. In this sense adherence to constraints helps prevent the model from being easily falsifiable.

A second continuum, **X**, outlines some of the pragmatic techniques used to build models. Model building techniques take many forms from textual descriptions to flow-charts and decision trees. At a finer degree of granularity both rules and numerical formalization overlap to form a computational explanation of a given cognitive phenomenon. This type of model building is often referred to as symbolic artificial intelligence and connectionism.

The third continuum, **Z**, concerns the granularity of interpretation for the model's behavior. In the simplest terms it involves comparing the model's performance with data obtained from qualitative and quantitative investigations. Moving along the continuum involves first defining the physical and environmental structural assumptions of the model. For example, what is the structure of the input, what is the structure of the output, and how does the actual model reflect the actual IPS being modeled. Finally, the model's representation must be given an interpretation.

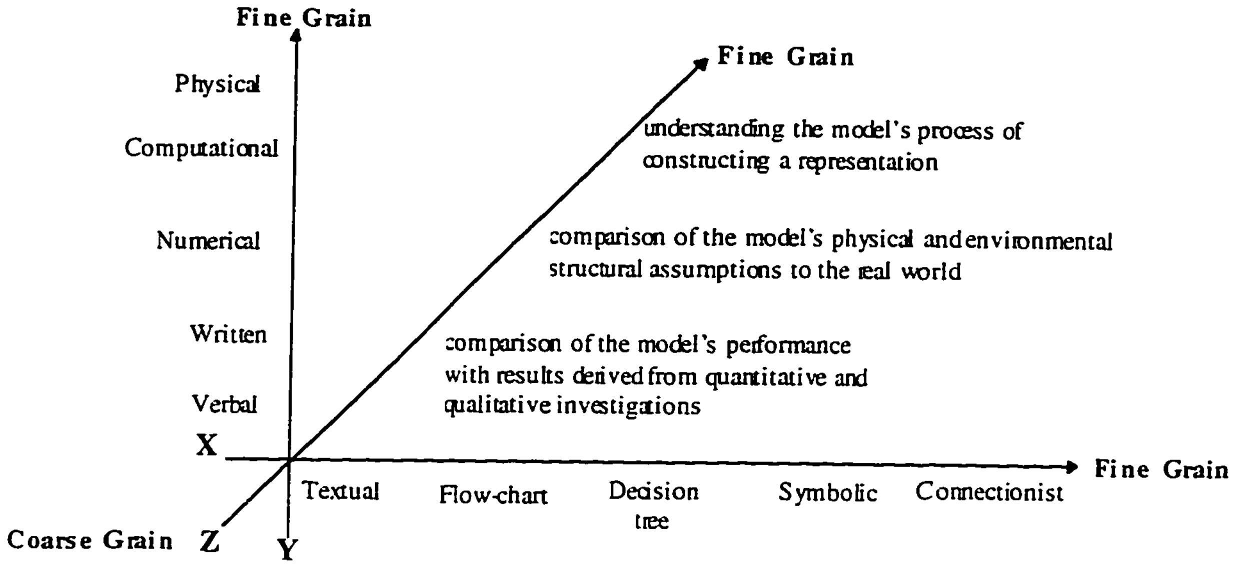

Fine Grain
Physical
Computational
Numerical
Written
Verbal
X
Fine Grain
understanding the model's process of constructing a representation
comparison of the model's physical and environmental structural assumptions to the real world
comparison of the model's performance with results derived from quantitative and qualitative investigations
Fine Grain
Textual
Flow-chart
Decision tree
Symbolic
Connectionist
Coarse Grain
Z
Y

Figure 1: Dimensions of granularity used in cognitive modeling.

In summary, the concept of granularity outlines a general framework in which to view computational cognitive modeling. The three issues discussed in this thesis can analogously be placed along a coarse to fine grain continuum. That is, issue one presents the "big picture" view of the relationship between computational modeling and education, issue two looks at the influence instructional theory can have on the structure of computational models, and issue three examines the representations formed by the model.

1.2 Overview of the thesis

The purpose of this work reported here was to develop the role of computational modeling in educational research. It is hoped that results from this thesis will demonstrate the importance of computational cognitive modeling in educational research.

Chapter 2 outlines the role of computional modeling in the context of cognitive science and educational research. It describes the two basic computational modeling architectures, symbolic artificial intelligence and connectionism. Finally it focuses on the implications connectionist modeling can have in areas of learning and development.

Chapter 3 recasts Bruner, Goodnow, and Austin's (1956) classic concept attainment task in a connectionist framework. A network was constructed which associates positive and negative instances of a concept with their corresponding attribute values. Two methods were used to help preserve the ecological validity of the input: (a) closely mapping the input to the actual visual stimuli; and (b) structuring the output layer based on Gagne's (1962; 1985) work on human concept learning, which lead to the addition of attribute context constriants.

Chapter 4 presents an analysis of the network representations formed by the network during learning. A variety of mathematical, statistical, and graphical techniques are used to provide an interpretation to the internal network weight space for various network configurations. Specific attention is given to the effects of attribute context constraints on the network's process of representation intantiation.

1.3 References

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). A study of thinking. New York: John Wiley & Sons.

Churchland, P. S. (1986). Neurophilosophy: Towards a unified science of the mind-brain. Cambridge, MA: MIT Press/Bradford Books.

Gagne, R. (1962). The acquisition of knowledge. Psychological Review, 69(4), 355-365.

Gagne, R. M. (1984). Learning outcomes and their effects. American Psychologist, 39(4), 377-385.

Gagne, R. M. (1985). The conditions of learning and theory of instruction (4 ed.). Toronto, ON: Holt, Rinehart and Winston.

Gardner, H. (1985). The mind's new science: A history of the cognitive revolution. New York: Basic Book Inc.

Johnson-Laird, P. N. (1988). The Computer and the mind: An introduction to cognitive science. Cambridge, MA: Harvard University Press.

Klahr, D. (1992). Infromation-processing approaches to cognitive development. In M. H. Bornstein & M. E. Lamb (Eds.), Developmental psychology: An advanced textbook (pp. 273-335). Hillsdale, NJ: Lawrence Erlbaum Associates.

Marr, D. (1982). Vision: A computional Investigation into the human representation and processing of visual information. San Francisco: W. H. Freeman.

Simon, T. J., & Halford, G. S. (Ed.). (1996). Developing cognitive competence: New approaches to process modeling. Hillsdale, NJ: Lawrence Erlbaum Associates.

Turing, A. M. (1936). On computable numbers, with an application to the entscheidungs problem. *Proceedings of the London Mathematical Society*, *Series 2*(42), 230-265.

Chapter 2

2. The Implications of computational cognitive modeling

Recently research in cognitive science has drawn the attention of people in the field of education (Elman, Bates, Johnson, Karmiloff-Smith, Parisi, & Plunkett, 1996; Schneider & Graham, 1992; Simon & Halford, 1996b). This research is directed at understanding processes of the mind.[1] Clearly, any field that offers insight into such processes has implications for furthering our understanding of learning and development, issues that lay at the center of education. At the core of research in cognitive science is the designing and building of computational models of mental phenomena. Although modeling is often deemed to be important, it is not always obvious: (a) What constitutes a computational cognitive model? (b) Where do these models fit in the overall framework of research? (c) How are such models constructed? (d) Why these models are potentially significant for educational researchers? This chapter attempts to answer these questions and provide a rationale for computational cognitive modeling[2] in the context of educational research.

The chapter is organized as follows: Section one briefly reviews the traditional approaches to educational research and postulates a position computational cognitive modeling might occupy relative to such approaches. The second section addresses some of the definitional issues in cognitive science and includes a more specific definition of a cognitive scientist that is intended to describe what a typical researcher in the field of cognitive science does. The third section describes the position of cognitive science in relation to the disciplines of philosophy, psychology, and neuroscience, and emphasizes the interdisciplinary relationship among different levels of explanations of understanding the mind. The fourth section explains why cognitive processes can be viewed as computational processes and realized in the form of computer programs. The fifth section examines the two most prevalent forms of computational cognitive modeling architectures, symbolism (symbolic artificial intelligence), and connectionism (artificial neural networks or parallel distributed processing). To gain insight into the discipline of cognitive science it is necessary to have an understanding of these cognitive modeling architectures and their

[1] The concept of mind is a nebulous one. In the context of this paper the term mind refers to the mental processes concerned with learning, cognition, knowledge representation, and development.

[2] The term comuptational cognitive modeling refers to the simulation of cognitive processes by a computer program.

associated techniques. The final section describes examples of cognitive-modeling which are directly related to educational research issues such as child development and concept formation. Emphasis will be given to models which are realized within a connectionist framework.

2.1 Educational research

The domain of educational research is generally placed within the broader category of social science research. Within this context two general methodological approaches predominate: quantitative and qualitative research (Borg & Gall, 1989). Traditionally, quantitative approaches to research design and data collection have been the mainstay of educational research. Essentially quantitative research involves the use of numbers to help measure, interpret and describe the phenomenon under investigation. A variety of analytical techniques are employed by quantitative researchers to collect and process data. Basically these techniques can be grouped into the following two categories: (a) descriptive and correlational studies that attempt to demonstrate relationships among variables, or (b) experimental studies that attempt to demonstrate a cause-and-effect relationship between two or more phenomena.

Quantitative research is concerned with instrumentation, validity and reliability, and design controls that emphasize methods which minimize threats to the internal and external validity of the study. For example, Campbell and Stanley's (1963) classic work describes twelve threats to internal and external validity in experimental and quasi-experimental research designs and explains how such threats may be controlled (e.g., various configurations for experimental and control groups). Data gathered in quantitative research studies are often analyzed using statistical procedures such as multiple regression or analysis of variance. The end results of such studies are usually predictive models or inferential conclusions about a population based on findings from a sample.

Qualitative research is usually associated with textual or narrative descriptions of a phenomenon. For example, a qualitative researcher might use analytical techniques often associated with ethnographic research or with case studies. The process of data collection relies heavily on both the subjects' and researchers' perceptions and interpretations of environmental events that develop in a natural context. The credibility of qualitative research is determined by the degree to which conclusions are thought to be believable. Qualitative research methodology has become increasingly popular in the educational community over the last twenty years (Frankel & Wallen, 1993).

Often qualitative and quantitative approaches to research are mistakenly interpreted to be mutually exclusive. This is not the case as many studies mix both approaches during the course of an investigation (Wiersma, 1995). For example, researchers might combine the techniques from verbal protocol analysis with assessment instruments to collect both narrative and quantifiable descriptions of a problem solving task. Overall, educational research presently employs varying degrees of both quantitative or qualitative methodologies and, to a large extent, educational researchers have borrowed and adapted quantitative and qualitative analytical research techniques from the disciplines of Anthropology, History, Science, Psychology, and Sociology. Furthermore they have relied heavily on measurement and statistical procedures for the collection and interpretation of data (Borg & Gall, 1989).

2.1.1 Role of cognitive science in educational research

Given the present approaches to educational research, two questions appear to be most relevant with respect to research in Cognitive Science: (a) What analytical techniques does the field of cognitive science have to offer educational research? (b) How might such techniques be assimilated and accommodated into the existing framework of qualitative and quantitative research methodologies?

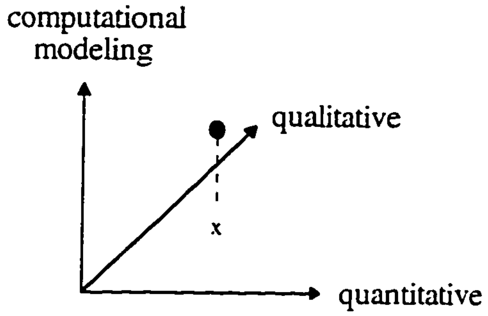
Unfortunately the answer to the first question does not come in the form of a set of analytical procedures that can be readily applied, as is often the case with statistical techniques. The analytical techniques employed by cognitive scientists are far from "cook book" in nature, and require the researchers to have some understanding of computation and its relationship to the cognitive phenomena under investigation. The computational aspects of the cognitive scientist's work are realized in the form of computer simulations. Even experienced educational researchers of cognitive phenomena can be confounded by the technical language of computer science that is often used to describe such simulations. As a result, a "communication gap" often exists between cognitive scientists and the more traditional researchers involved in the study of cognitive processes.

Almost fifteen years ago Bower and Hilgard (1981), in their chapter on Information-Processing Theories of Behavior, referred to such a gap as "serious" but then suggested that this gap appears to be closing over time. It appears their early assessment of the gap closing may have been somewhat optimistic.[3] Part of the reason for the gap not

[3] It is interesting to note that Bower and Hilgard's (1981) chapter on information processing (IP) is almost entirely composed of descriptions of IP models along with their associated computer simulations. Although, somewhat dated, their account of the relationship between IP models and computer simulation

closing has to do with the unprecedented growth in the computational power available to construct computer simulations. This power has often resulted in exceedingly complicated computer simulations being constructed, like the SOAR type models proposed by Newell (1990). Further exacerbating this gap has been the development of a wide variety of different approaches to computational cognitive modeling. For example, at the time of their writing Bower and Hilgard discuss information processing models with respect to the early computer simulation work of Newell and Simon (1956; 1972). Work of this type is often referred to as "symbolic artificial intelligence" or "production rule-based reasoning" (Luger & Stubblefield, 1989) and represents only one approach to modeling cognitive behavior. Recent advances in cognitive modeling techniques have resulted in the development of brain-like connectionist models such as the Parallel Distributed Processing (PDP) models of Rumelhart and McClelland (1986b), along with a variety of other types of neural network models (Hecht-Nielsen, 1990). Although a good deal of learning effort is still required by researchers interested in using specific cognitive modeling techniques—much insight can be gained by describing the premises of such techniques and their method of implementation.

One way to approach answering the second question regarding the relationship between computational modeling and existing forms of educational research (qualitative and quantitative) is to view each methodology as sharing some aspects of the other. Figure 2 presents a three dimensional view of the possible relationship between qualitative, quantitative, and computational modeling. The black circle is positioned to indicate that the three methodologies may share varying degrees of overlap with respect to how they investigate and explain a given phenomenon. More importantly, results from research using any or all of the methodologies can act as constraints on further research and thereby guide investigative procedures.



models is still highly relevant and informative.

Figure 2: A three dimensional view of the relationship between qualitative, quantitative, and computational modeling research methodologies.

With respect to Figure 2 computational models can be viewed as formalized models of specific cognitive phenomena. The words formalized models are emphasized for the following two reasons:

1. Constraints. Both qualitative and quantitative research results act as constraints with respect to computational models. Constraints can be viewed as behavioral assumptions under which the model must function. For example, Inhelder and Piaget (1958) first postulated a qualitative description of how children attempt to solve a balance scale problem. Their research observations provided important insight into our understanding of the development of children's deductive reasoning. Siegler (1976) later hypothesized four rules children might use to solve the same balance scale problem. Siegler used an experimental (quantitative) procedure, rule assessment method, to develop a more formal understanding of the children's balance scale problem solving approach at different developmental stages. Building on this work, Klahr and Siegler (1978) built a computer simulation (production rule-based reasoning system) which embodies the rules identified by Siegler. Langley (1987) later modified the work of Klahr and Siegler by constructing a computer model, referred to as an adaptive production system, that was concerned with explaining the transition between rule stages. More recently, McClelland and Jenkins (1991) constructed a connectionist network that exhibits the same rule based behavior on the balance scale task first postulated by Siegler.

The path from Inhelder and Piaget (1958) to McClelland and Jenkins (1991) depicts a continuum of research endeavors designed to formalize an explanation of children's reasoning during a problem solving activity. The computational model appears as the final step in this process and represents a system that behaves similarly to the observed human behavior under investigation. In other words, such models can be viewed as having the capability to describe, predict, and algorithmically represent, specific human behaviors.

2. Mathematical. Computational models are composed of representations that are grounded in logical reasoning (deductive or inductive logic) or numerical values (parameters or coefficients) and thus rely on computable procedures to describe cognitive phenomena. In a mathematical sense, the act of building a computational model, in the form of a computer simulation, is a formal representation process that can be described in terms of computable procedures.

2.2 Cognitive science

The field of cognitive science has been gradually developing for the past 40 years and involves the integration of ideas from a number of disciplines. The major contributors to this trend are from the fields of neuroscience, psychology, computer science (artificial intelligence), philosophy, anthropology and linguistics (Gardner, 1985). An obvious question is, "what exactly is cognitive science?" Ironically, many researchers within the field often find themselves asking the same question.

One of the problems in trying to pin down a clear definition of cognitive science is its interdisciplinary nature. This is further confounded by the rapid growth of research in the associated disciplines of computer science and neuroscience. The field is truly a victim of the current phenomenon colloquially referred to as the 'information explosion.' One of the consequences of the information explosion is that the terminology used by the cognitive scientist appears somewhat ill-defined and context sensitive.

Such ambiguity of definition can also be traced to the different backgrounds of the researchers in cognitive science. A review of the literature failed to provide a single, unambiguous definition. For example, in the Dictionary of Psychology (Reber, 1985) cognitive science is defined as:

> A newly coined name for the cluster of disciplines that studies the human mind. The term refers to an amalgamation; it is an umbrella term which includes a host of once disparate approaches such as cognitive psychology, epistemology, linguistics, computer science, artificial intelligence, mathematics and neuropsychology. (p. 130)

This definition is somewhat lacking in that it gives little insight into what a cognitive scientist actually does. Howard Gardner's (1985) classic book The Mind's New Science, provides a more precise definition:

> I define cognitive-science as a contemporary, empirically based effort to answer long-standing epistemological questions—particularly those concerned with the nature of knowledge, its components, its sources, its development, and its deployment. (p. 6)

Gardner continues, later in the book, to offer a much richer and, of course, much longer conceptual definition of cognitive science, based on the philosophical and scientific disciplines from which it is derived. Gardner offers five distinguishing features of cognitive science, many of which have been discussed in one form or another by other cognitive scientists (Collins & Smith, 1988; Johnson-Laird, 1988; Posner, 1989).

1. Representations. An essential feature of cognitive science is that there exists a "level of representation" at which mental states can be formalized with representational entities such as symbols, rules, logic, or numerical values. These entities are used to construct cognitive models of processes that are usually postulated to exist in the input and output flow charts of cognitive psychology. The appropriate representational level at which to model the phenomenon being investigated is a critical issue for any cognitive scientist (Marr, 1982; Newell, 1982; Smolensky, 1988).

2. Computers. The development of the first primitive computers in the 1930s and 1940s had a significant influence on how psychologists believed the mind might operate (Pylyshyn, 1989). In this context the computer was viewed as a metaphor for how the mind might function. This idea led to the development of information processing psychology later called cognitive psychology (Collins & Smith, 1988). For many cognitive scientists building a computer model of cognitive processes is central to their research work. Initially this was often left to the field of artificial intelligence. More recently, the distinction has blurred between the artificial intelligence researcher who is interested in building systems that claim to account for human processes, and the cognitive scientist. More and more universities are developing programs of study in cognitive science—so much so that cognitive science can now be said to have absorbed many of the artificial intelligence researchers from computer science who are interested in studying human cognitive processes.

3. De-Emphasis on Affect, Context, Culture, and History. Traditionally, cognitive scientists have tended to factor out the elements of context, culture and history. To build operational computational models which attempt to account for the cognitive processes, one must constrain these models within a pragmatic realm. One would like to avoid trying to explain everything because in doing so one explains nothing (i.e., essentially attempting to account for too many variables).

4. Belief in Interdisciplinary Studies. As indicated above, cognitive science is influenced by a number of research disciplines. This mixture of ideas allows for the development of more powerful insights into cognitive phenomena in the sense that it considers research results obtained in a variety of contexts. Complex processes are often understood at a multitude of levels each of which provides its own explanatory power. For

example, the biochemist is interested in the microlevel of chemical energy transformation that occurs in Kreb's cycle to produce Adenosine Tri Phosphate (ATP) from glucose. Meanwhile at a macro level a coach is interested in an energy efficient diet which maximizes athletic performance. Furthermore by working from an interdisciplinary perspective, one is more likely to induce explanations which are constrained by the knowledge accumulated from other related disciplines. An important belief among many cognitive scientists is that efforts to explain cognitive processes must be embodied in a constraint satisfaction relationship between various research disciplines (Churchland & Sejnowski, 1992).

5. Rootedness in Classical Philosophical Problems. Like any discipline cognitive science appeals to philosophical explanations for reasons of justification and as a starting point for initial research inquiries (Fetzer, 1991). For example, questions such as: Why do we think? How do we think? What is the nature of knowledge? These are philosophical questions of the epistemological workings of the mind that date back to ancient Greece. According to Gardner (1985) these questions appear as relevant for the present day cognitive scientist as they were for Plato and Aristotle.

2.2.1 The cognitive scientist

The above features provide a general perspective on cognitive science but lack the specific details of what a cognitive scientist actually does. In order to understand a cognitive scientist's work one needs to consider the mind as though it were a "computational device" (Churchland & Sejnowski, 1992; Johnson-Laird, 1988; Pylyshyn, 1989; Simon & Kaplan, 1989). It then follows that: (a) computational cognitive models can formalize abstract ideas such as learning, knowledge representation, concept formation, and development, so that a representational approach to mental states and processes can be instantiated in a computational world; and (b) computational models offer insight in mental processes not available through other avenues of research. For example, in the study of development, Simon and Halford (1996a) make the following statement in support of computational models:

> Microgenetic methods involve high-density observations of children's behavior over an extended period of time and were found to be effective in providing insights into behavioral changes the children exhibit. However, the mechanisms responsible for those changes are not available to inspection. They can only be inferred from the effects they have on the initial competence of the children as a consequence of observed activities.

> What is missing from the microgenetic approach, but present in a running computer model, is a physical embodiment of the actual representations and processes with a lid one can open and look inside. This enables the theorist to unequivocally state what the set of mechanisms are that transform the initial representations into those that support the target behavior. So, whereas microgenetic or other kinds of longitudinal methodology may track the time and experience aspect of cognitive change very closely, they do not provide direct access to candidate mechanisms of change—for us, the objects of greatest interest. (p. 13)

In summary, a somewhat restrictive definition of a cognitive scientist is proposed: A cognitive scientist is a person who builds cognitive-models, in the form of computer programs, and where these models are constrained by knowledge from any discipline which has gathered data (qualitative, quantitative, and physical) with respect to how mental processes may operate. The idea of constraining these computational models with knowledge from other disciplines is important and represents the primary vehicle for the integration of educational research and cognitive science. The more research data that provides constraints for the model, the greater the model's explanatory power. Building precise cognitive models of mental processes allows researchers to go beyond the observational level of investigation. Essentially, the work of a cognitive scientist is directed at providing a more precise accounting of the observable data and to help deal with the common problem of neuroscience and psychology research of being 'data rich and theory poor' (Churchland & Sejnowski, 1992).

2.3 The cognitive science level of explanation

Theoreticians in philosophy, psychology, and neuroscience have postulated a wide range of ideas for understanding the mind. Essentially these theoretical approaches can be envisaged as representing a particular general level of explanation of the mind. Selecting the correct level at which to postulate an explanation of the mind is a critical factor in determining the richness and applicability of the explanation. For example, there is little doubt that neurological and biochemical factors ultimately underlie the functioning of the mind—at the same time this may not always be the most appropriate level to explain a given mental phenomenon (e.g., the semantic meaning of a visual representation).

Two features appear to be significant when selecting a level: (a) the pragmatic functional value of the explanation, and (b) the elegance of the explanation. Pragmatic functional value means that the explanation embodies predictive power such that over time causal relationships, within and between phenomena, are relatively invariant. The notion of having an elegant explanation is subjective and depends on the context and constraints which encompass the phenomenon being explained. The key feature of an elegant explanation is the degree to which the explanation makes the phenomenon explicit, in other words, understandable and simple. An example of both a pragmatic and elegant explanation, in the context of mathematics, would be the use of row reduction to obtain a solution to a set of simultaneous linear equations. For this type of problem the constraints are well defined by the mathematical rules of linear algebra and the explanation clearly identifies the phenomenon of a linear relationship among equations.[4] Analogously, in the context of genetics, a high school student may be asked to work out a pea plant inheritance problem using the simple Punnet square method. The problem is constrained by a well known procedural methodology and by the initial parameters. The pea plant level of explanation of inheritance is pragmatic, in that the causal relationships of pea plant inheritance are relativity invariant and the method proves to be a good predictive model of inheritance outcomes. Furthermore, the Punnet square explanation is elegant in that it easily conveys the phenomenon of inheritance.

There will always be arguments over the robustness and depth of understanding an explanation can provide. The reason for this is primarily attributable to the degree of insight any given explanation can provide. One can always question whether a higher degree of insight can be obtained by appealing to lower (more micro) levels of explanation at the cost of giving up the generality of the explanation. For example, some developmental psychologist might postulate, at a general level, that development occurs in a series of stages (Bruner, Olver, & Greenfield, 1966; Piaget, 1954; Vygotsky, 1934/1986). Others might take a more detailed analysis of the developmental process and postulate explanations that developmental change is not discrete but continuous (Eimas, 1994; Siegler, 1995). Still others may study specific changes to the neural structure of the brain to account for the developmental process (Greenough, Black, & Wallace, 1987).

Clark (1993), a modern day connectionist philosopher, states the following with respect to explanations:

[4] The solution set should be a point at which all other equation lines (planes or hyperplanes) intersect assuming the problem is well formed (non-singular).

> Explanation, it seems, is a many-leveled thing. A single phenomenon may be subsumed under a panoply of increasingly general explanatory schemas. On the swings and roundabouts of explanation, we trade the detailed descriptive and explanatory power of lower levels for a satisfying width of application at higher levels. And at each such level there are virtues and vices; some explanations may be available only at a certain level; but individual cases thus subsumed may vary in ways explicable only by descending the ladder of explanatory generality. (p. 42)

The following discussion expands upon the notion of levels of explanation as they pertain to understanding the mind. Three major premises underlie this discussion: (a) the influence of science and technology are major factors in pursuing a reductionist explanation of the mind, (b) there is a co-evolution of research at each respective level, and (c) all levels are subject to the constraining knowledge obtained at other levels. These themes are essentially a synthesis of ideas taken from the works of Newell (1982), Marr (1982), Churchland (1986b), and Smolensky (1988).

2.3.1 General levels: philosophy, psychology, neuroscience

Figure 3 shows the three disciplines of philosophy, psychology, and neuroscience that have traditionally been responsible for generating theories of the mind which can be considered to occupy different levels of explanation. Figure 3 also illustrates an evolutionary trend toward reductionistic investigative inquiries. Reductionism is highly correlated with the use of a formalized approach to explain a phenomenon, such as those used in science. Essentially, reductionism maintains that a complex phenomenon (system) can be best understood by breaking the phenomenon into its fundamental elementary parts and studying these parts separately. The bold arrow, on the right-hand side, in Figure 3, highlights this underlying reductionist research trend both within and between the three disciplines.

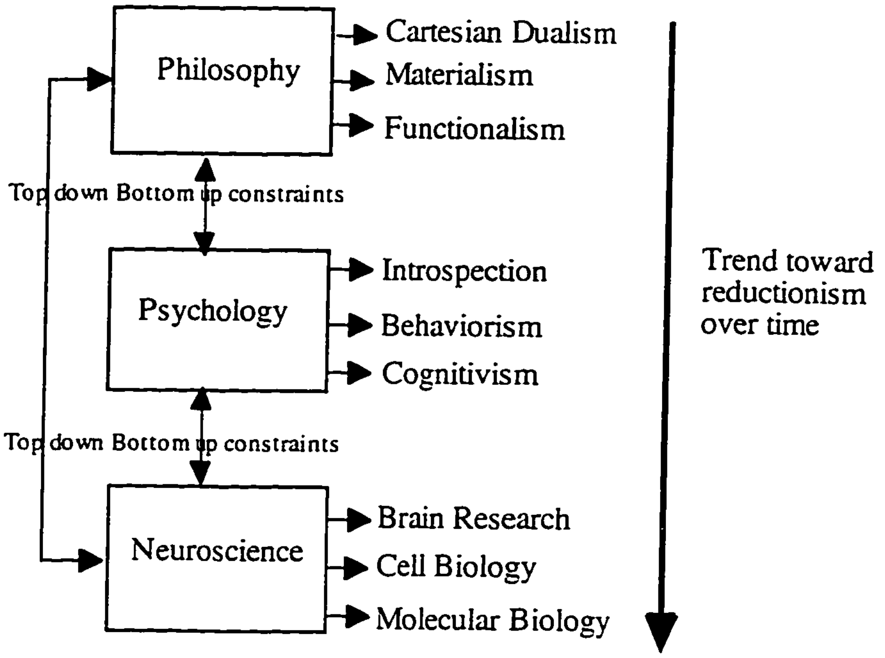


Figure 3: Three levels of explanations of the mind viewed as a reductionist continuum.

2.3.1.1 Philosophy

The contributions of philosophy represent the starting point from which lower level inquiries in the areas of psychology and neuroscience can be derived and influenced. Philosophy is deemed to be the most general level at which to discuss theories of the mind because it relies less on the formal scientific method to substantiate many of its arguments. On the other hand, philosophy provides a framework, often in the context of logic, from within which certain avenues of research may be worth pursuing.

Many different kinds of philosophical theories have been proposed, with respect to the mind, such as Cartesian dualism, materialism, and functionalism (Churchland, 1986b; Lavine, 1984). Early philosophical ideas of the mind were grounded in epistemology and typically resulted in subjective qualitative investigations (Gardner, 1985). René Descartes postulated a theory of a spilt between mind and body, that is referred to as Cartesian psychophysical dualism. According to Lavine (1984):

> Cartesian psychophysical dualism may be defined as the doctrine that reality consists of two kinds of substances, mental and physical, and that the one kind of substance can never be shown to be a form of, or reduced to, the other. So for psychophysical dualism, mind can never be shown to be

> derived from, or a form of, or a function of, or reducible to, matter. (p. 122)

In Descartes' view, the mind, unlike the body, was thought to be a rational entity that did not avail itself to decomposition in the same way as a physical entity. Descartes viewed the act of thinking and the representation of knowledge as a logical process that could be explained by intuition and deductive reasoning.

In contrast to Descartes many philosophers proposed an alternative philosophical stance referred to as "materialism." Moody (1993) describes materialism in the following way:

> ...there are not two categories of things in the world [mind and body] but only one: material, or physical, things. Everything that happens in the universe involves physical objects, forces, and processes, and nothing stands outside of that totality of physical interactions. For all physical phenomena, physical causes must be sought and can, in principle at least, be found. (p. 31)

Materialistic philosophy offers the possibility of understanding the processes of the mind with respect to how such processes are carried out in the physical substrate from which they are assumed to have arisen. Furthermore, without a materialistic philosophical viewpoint, most of present day psychological and neuroscience research would be deemed useless.

Materialism represents a class of theories, and a subclass called "functionalism", has direct implications for computational cognitive modeling (Memmi, 1990). This theory states that "mental states are just functional states of a complex system" (Moody, 1993, p. 43). Functional states have a cause and effect relationship between each other. Two systems (e.g., computer and human) can be viewed as functionally equivalent if they process the same input and produce the same output. Functionalism has proven to be an important philosophical perspective for cognitive modeling research using the symbolic techniques of artificial intelligence (Pylyshyn, 1989). Memmi (1990) describes the theoretical perspective of functionalism that underlies this symbolic modeling assumption:

> ...the conviction that the level of symbolic representations is the appropriate level of description. Representations and their manipulation are thus not only necessary but also sufficient for the explanation of cognitive processes.

> The way in which representations are implemented is basically irrelevant, and mental operations can be carried out equally well by computers or by brains. The usual analogy is that only software is important, not the hardware. Physical reductionism is considered uninteresting because it won't tell us anything worthwhile about cognitive functions. (p. 118)

The symbolic representations referred to by Memmi and the causal relationships that can occur between such representations is discussed in more detail in the section concerned with symbolic artificial intelligence. Suffice it to say these symbolic representations are realized in the rules of formal logic that are often used by philosophers to describe a reasoning process.

The three theories, Cartesian Dualism, Materialism, and Functionalism can be organized in keeping with the reductionist trend as illustrated in Figure 3. For example, the use of propositional logic in conjunction with functionalist theory can be viewed as part of this reductionist trend to decompose the human reasoning processes into a formal system, with the prospect of making these reasoning processes explicit. The use of formal logic makes a description of reasoning specific and reproducible (Haugeland, 1985).

2.3.1.2 Psychology

The psychological level of explanation is essentially an outgrowth of the philosophical level (Hothersall, 1990). Psychology attempts to formalize the investigative procedures used to study mental processes by the inclusion of scientific methods to control the experimental approaches. A formal scientific approach to research is by its very nature reductionist because it constrains the investigative parameters and requires that procedural steps be identified such that the experimentation process is reproducible.[5]

The reductionist trend in psychology can be classified chronologically into the following four general psychological theories: (a) introspection, (b) behaviorism (c), cognitive psychology, and (d) cognitive science. Each manifests itself as an outgrowth of its predecessor. Only the first three theories will be reviewed, since cognitive science was discussed previously.

[5] The chronological order of the scientific method is usually: (a) identification of the problem, (b) definition of the problem, (c) formulation of hypotheses, (d) projection of consequences, and (e) testing of hypotheses (Frankel & Wallen, 1993).

Introspection of mental processes is a self-observational approach to data collection which dominated early (pre 1900s) forms of mental investigation (Hothersall, 1990). Subjects involved in introspection experiments would be shown objects, such as a chair, and asked to report their immediate experiences with regard to height, colour, and structure of the chair. The idea was that an introspective researcher could gain insight into how the subject's consciousness might be organized (Schunk, 1996). Introspective experimenters had to undergo rigorous training in how to conduct interviews and gather data. Although introspection generated a wealth of intuitive data on the mind's operations, models based on this approach were unreliable and of little pragmatic research value.

Behaviorism grew out of psychology's frustration with the introspective method. Watson (1913), one of the founders of behaviorism, wrote:

> Psychology as the behaviorist views it is a purely objective natural science. Its theoretical goal is the prediction and control of behavior. Introspection forms no essential part of its method nor is the scientific value of its data dependent upon the readiness with which they lend themselves to introspection in terms of consciousness. (p. 158)

Behaviorism represents a further attempt to formalize the field of psychology by using a more rigorous scientific approach to gather precise measurements of behavior. Early behaviorists employed the investigative method of "Stimulus-Response" (S-R) to collect experimental data. In its extreme form, behaviorism claimed "thought processes are really motor habits in the larynx" (Watson, 1913 , p. 174) and mental processes as being habits of language. What behaviorism lacked was an explanation of mental states and processes which accounted for these observed behaviors.

Cognitive psychology (information processing) emerged as a challenge to behaviorism and attempts to more accurately account for mental events, representations, and processes (Bruner, Goodnow, & Austin, 1956; Miller, Galanter, & Pribram, 1960; Newell & Simon, 1956). Behaviorism came to be viewed as a theory which lacks reasonable explanations for cognitive processes such as perception, attention, concept formation, and some forms of learning. The information processing metaphor for the mind became the dominant paradigm during the post W.W.II period and is currently the mainstay of psychological research programs at most universities. This theory is derived from the computer model of information processing and both the computer scientist and cognitive psychologist talk in terms of input, output, long term memory, working memory, encoding (representations), and so forth. Information—objects with meaning/content—are processed by the mental act of generating states (hypothesis) and moving through these states (evaluating these hypothesis). Essentially, a cognitive psychologist postulates a functional organization (system), usually outlined in the form of flow-charts, that attempts to explain how various mental processes interact.

Much of cognitive psychology research is geared toward quantitative experimental design and thus purports to measure some psychological phenomena, (e.g., how many pieces of information can a person hold in working memory). In general, experimental methods of cognitive psychology can be thought of as a coarse grain approach in the study of observed behaviors. For example, Piaget's classical experiments on the concept of conservation suggest that children appear to progress through four distinct developmental stages. More recently, information processing theories of child development have focused on such things as memory limitations, representation of information, and transformation of representations (Siegler, 1991). Furthermore, there appears to be a good deal of empirical evidence to support the notion of both serial and parallel processing of information (Massaro & Cowan, 1993).

The three psychological theories of introspection, behaviorism, and cognitivism are shown in Figure 3 as being organized along a top-down reductionist continuum. Each psychological theory postulates models of observable behavior. Following the reductionism trend, these models become increasingly more explicit with respect to their descriptive power. In other words, such models may offer a greater level of predictive and explanatory power with respect to the observable behavior they attempt to describe.

2.3.1.3 Neuroscience

Research indicates that the brain is an extremely complex interconnected structure estimated to contain between 10^{10} to 10^{12} individual neural processing units (Sejnowski & Churchland, 1989). Broadly speaking there appear to be three major properties of brain structure, (a) regional or functional components (modules) within which there are both local interactions of elements within a region or function and global interactions among elements between regions or functions, (b) learning involves structural neural changes, (c) there are genetically pre-specified structures and operations which are relatively invariant over time (Churchland & Sejnowski, 1992; Crick & Asanuma, 1986; Fischler & Firschein, 1987; Kolb & Whishaw, 1990; Luria, 1980; Posner, Peterson, Fox, & Raichle, 1988).

There are many areas of neuroscience[6] research such as brain measurements (PET-scans), cellular measurements (single cell activity levels), and molecular measurements (electrochemical synaptic connections). One promising area of neuroscience research is that of lesioning neural structures (Sejnowski & Churchland, 1989). Lashley (1929), was one of the first to use the techniques of surgical lesioning on the neural tissue in rats as means of understanding their behavior. Studies of brain injured patients are illustrative of the notion that damage to the central nervous system can supply valuable behavioral data on mental operations (Shallice, 1988). More recently, neuroscientists have attempted to understand the overall processes of the brain by the use of various non-intrusive techniques, among them positron emission tomography (PET-scans). Interestingly these PET-scans show that functionally distinguishable subprocesses (e.g., lexical tasks) are separately localized in the brain (Posner, et al., 1988).

In general, work in neuroscience represents a fine grain approach to the study of structural change. For example, early researchers like Hebb (1949) argued that learning involved physical changes at the neural level. Greenough, Black, and Wallace (1987) have proposed that mammalian brain development relies on two plasticity components: (a) experience-expectant (the brain is prewired to expect certain information storage to occur during specific points in the developmental process), and (b) experience-dependent (the generation of new synaptic connections in response to the occurrence of environmental events that need to be remembered). Understandably, many neuroscientists adhere to the purely reductionist approach to understanding a psychological phenomenon. As Black (1991) states:

> ... the scientific concept, psychologic modularity, appears to have a molecular, cellular, and systems reality. Subsequently we examine whether psychologic organization and functions can be driven "down" to the physical level. Indeed, psychology is not separable from the physical. Psychologic modular organization arises from physical modular organization. To understand mechanisms governing psychologic modularity, we must understand the physical modularity upon which psychology is based. (p.21)

[6] The term neuroscience refers the class of research areas which attempts to identify the link between physical brain or neural activity and observed behavior. Subclasses include research areas such as neuropsychology, and cognitive neuroscience.

Neuroscience is considered, for this discussion, to be the lowest level of reductionism. At this level the explanation of the processes of the mind are subjected to rigorous scientific investigation using techniques from chemistry, biology and physics. Such theoretical investigations contribute a wealth of data on the structure and function of the physical systems assumed responsible for mental processes such as visual perception (Fischler & Firschein, 1987).

Figure 3 also shows connecting arrows pointing both upward and downward between the boxes which enclose each of the disciplines. These connecting arrows represent the flow of knowledge between each of the three disciplines. For example, knowledge acquired by researchers at the neuroscience level can flow upward to the philosophical level and vice-versa. Knowledge obtained from a certain discipline may or may not have an effect on research at another level. Some researchers may choose to ignore research findings from other disciplines because they find them irrelevant or confounding. On the other hand, knowledge from one level may have significant influence on research at other levels. As a result, the second perspective that Figure 3 conveys is that research at different levels co-evolves; no one discipline is immune from the influences of the other. Modifications to existing disciplines can result in new theories being created and older ones disappearing. According to Churchland and Sejnowski (1992) "the hallmark of co-evolution of theories is that research at one level provides correction, constraints, and inspiration for research at higher levels and at lower levels" (p. 11). To formulate an explanation of the mind is difficult enough and disregarding results from other disciplines only makes such an explanation less robust and more easily falsifiable. The arrows linking the three disciplines in Figure 3 are labeled "Top-Down Bottom-Up Constraints" to indicate the constraining space of explanations each discipline imposes on the other.

2.3.2 Where does cognitive science fit?

As Miller, Galanter and Pribram (1960) point out, it is the dash between S-R that required further explanation, and cognitive psychology is essentially an attempt to posit such an explanation. Cognitive psychology operates under two basic assumptions. First, thought involves the concept of mental states. Second, the mind can be viewed as an information processing system that actively processes these mental states. Figure 4 shows a general concept map to which a cognitive psychologist would likely subscribe. Information, in the form of external events, acts as input. This input is processed by the information processing "black box" which produces the resulting output. The black box can be viewed as holding processes such as short term memory, long term memory, working memory, knowledge representation schemes (procedural and declarative knowledge), and executive control. Other processes assumed to be executed in the black box include attention, perception, and language development. Black box models of cognitive processes are usually expressed in terms of flow charts designed to describe how the processing actually occurs.

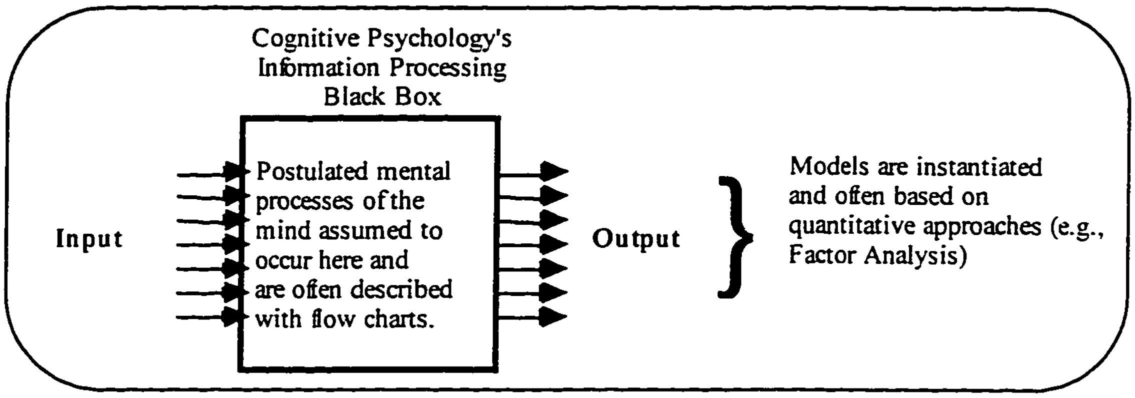

Cognitive Psychology's
Information Processing
Black Box
Input
Postulated mental processes of the mind assumed to occur here and are often described with flow charts.
Output
Models are instantiated and often based on quantitative approaches (e.g., Factor Analysis)

Figure 4: General cognitive psychology paradigm

A typical cognitive psychology experiment collects data using measurement tools purporting to assess the underlying cognitive processes in question (e.g., visual search reaction time). Based on a data analysis two things usually result: (a) modify or support an existing model of processes in the black box, (b) postulate a new black box model. The goal of such models is to provide cognitive psychologists with an element of predictive power in their explanation of observable behavior. For example, results from a study measuring cognitive processing skills in mathematics may allow an educational psychologist to formulate an intervention strategy (Carr & Jessup, 1995). The effectiveness of the intervention strategy is again the subject of another cognitive psychology experiment.

In most cognitive psychology theories an explanation regarding how cognitive processes function usually stops at the black box level. The underlying cognitive processes are thus made explicit only in the flow-chart sense. As Quinlan (1991) states:

> ...cognitive psychology [has] flourished as a discipline centred on the flow-diagram conceptions of internal representations and processes. Work in mainstream cognitive psychology typically dwelt on a limited aspect of adult performance and in a rush to posit the next boxes-and-arrows model of this aspect of cognition, many simply ignored problems about how the putative systems might have developed. (p. 38)

Figure 5 shows the classic and often referenced black box information processing model developed by Broadbent (1958). At the time, use of such a block diagram was considered a novel approach to summarizing an information processing theory (Johnson-Laird, 1988). According to Broadbent, information is seen as flowing through channels. An overload of information, such as too many voices being heard at the same time, results in the use of a selective filter that focuses attention. Although Broadbent's model has been revised based on further experimental findings (Treisman, 1964), the use of control engineering type flow-charts to specifically represent cognitive processes has remained unchanged to this day.

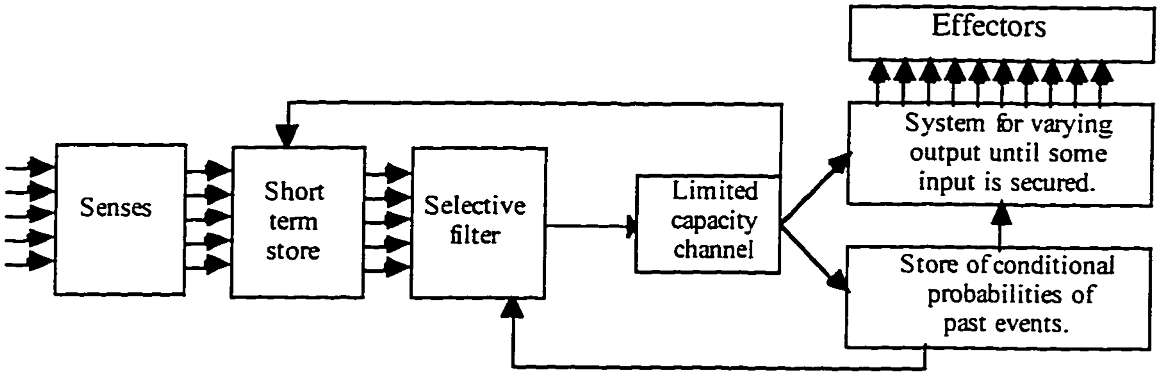

Effectors
System for varying output until some input is secured.
Senses
Short term store
Selective filter
Limited capacity channel
Store of conditional probabilities of past events.

Figure 5: Broadbent's information processing flow chart

Based on Broadbent's notion of sensory (visual, auditory) information processing Figure 6 shows a recent black box model developed by Das, Naglieri, and Kirby (1994). Das et al.'s model is used to explain their version of cognitive processing and is said to be a diagrammatic representation of the relationship between planning, attention, and simultaneous and successive processes (PASS). The PASS model is typical of the present day black-box models and has recently received attention in educational psychology circles (Languis & Miller, 1992). The Das et al. model is based on the results from factor analysis studies and takes into account the neuropsychological work of Luria (1980). Therefore, unlike Broadbent's model, the PASS model considers an additional bottom-up constraint by including evidence from the neuroscience level. This is an attempt by the PASS developers to argue for a higher degree of validity for their model.

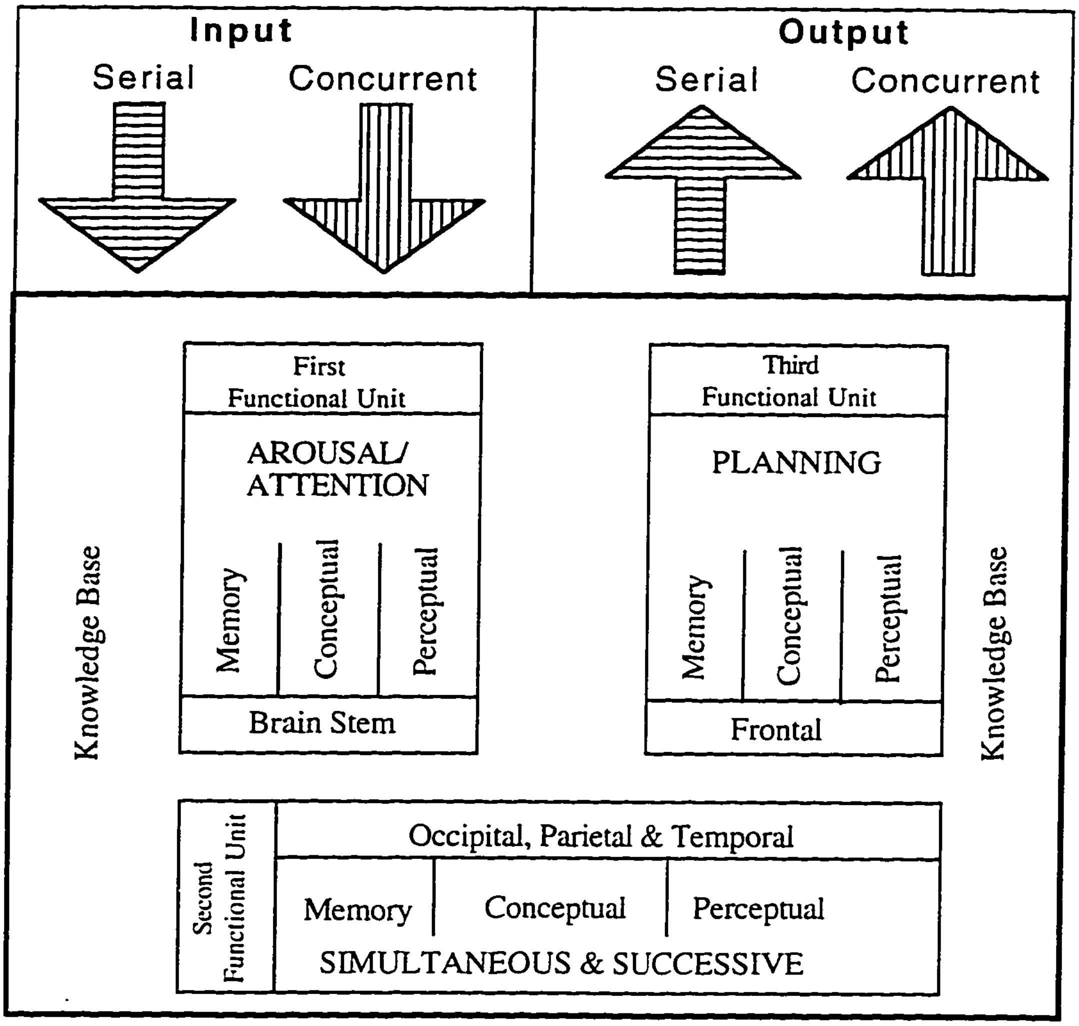


Figure 6: PASS model of information processing

Model building of the type suggested by Broadbent or Das et al. suggests, based on some experimental results, that it is highly likely that cognitive processes are present in the dash between S-R. Furthermore, flow-chart models of this type are an obvious attempt to further formalize our understanding by postulating relationships between specific cognitive processes. This formalization, as means of achieving more explanatory power, is part of the reductionist tendency of gaining greater insight into a phenomenon by breaking it into further sub-components.

Are these black box models a satisfactory explanation of the cognitive processes resident in the mind? For example, with respect to the Das et al. (1994) model, how do hypothetical constructs such as "simultaneous" and "successive" encoding of knowledge actually work? Over 20 years ago, Newell (1973) asked similar questions when commenting on a series of papers given by some of the leading cognitive psychologists of that time (Chase, 1973). Newell stated, "this phenomenon of simultaneous versus sequential grouping has occasioned some hundreds of papers over the intervening years [1958-1973], in an attempt to clarify the issues (was it channel switching or not?)" (p. 290). Furthermore, Newell was also frustrated with the flow-chart approach to describing the underlying cognitive processes. As an alternative he advocated the notion of "complete processing models" in the form of computer simulations, whereby "theory, embodied in the simulation, actually carries out the experimental task" (p. 301). Newell goes on to say:

> As I noted earlier, the attempts in some of the other papers to move toward a process model by giving a flow diagram... seem to me not to be tight enough. Too much is left unspecified and unconstrained... [and] these flow diagrams are not sufficient to perform their tasks. That flow diagrams may leave something to be desired as a scheme for cumulating knowledge might be inferred from a comparison of Donald Broadbent's two books (1958 and 1971), both of which contain flow diagrams representing what is known (at each respective date) about short-term memory and the immediate processor. (Newell, 1973 , p. 301)

Newell's notion of using computer modeling as a means to build more precise models of mental processes was supported by several other researchers (Hunt, 1962; Minsky, 1968; Nilsson, 1971; Schank, 1973). Making flow-chart models more accountable in the implementation sense (computer program) has become an essential theme in cognitive science. The computer is viewed as a tool that allows researchers to construct cognitive models, where such models are often constrained by empirical results from cognitive psychology. Therefore, computational cognitive modeling can be viewed as an attempt to break open the information processing black box and examine the cognitive processes from a computational perspective. It is this black box accountability process which forms the foundation of a cognitive scientist's work.

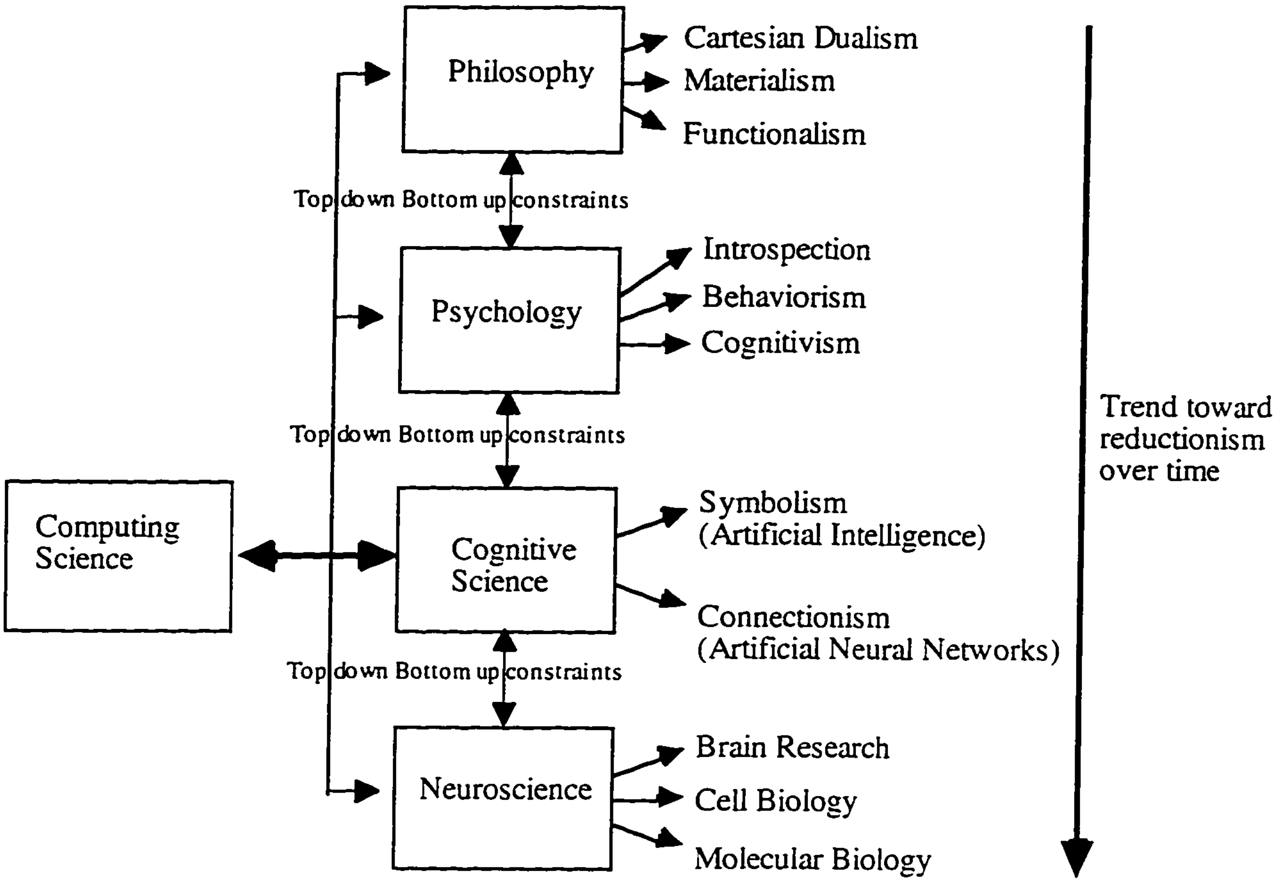

Philosophy
Cartesian Dualism
Materialism
Functionalism
Top down Bottom up constraints
Psychology
Introspection
Behaviorism
Cognitivism
Top down Bottom up constraints
Computing Science
Cognitive Science
Symbolism (Artificial Intelligence)
Connectionism (Artificial Neural Networks)
Top down Bottom up constraints
Neuroscience
Brain Research
Cell Biology
Molecular Biology
Trend toward reductionism over time

Figure 7: Expanded "levels of explanation" of the mind

Looking back on the definition of a cognitive scientist that was given earlier, the central argument is that a cognitive scientist builds computational models of cognitive processes. To accept this research approach, one must also accept Turing's (1936) notion of computability, which will be explained in some detail in section four. If one accepts the notion that the processes of the mind can be realized computationally the next step is to seek the appropriate level at which this explanation should reside.

Figure 7 is a modified version of Figure 3 presented earlier. Two more boxes have been added; one is labeled "Cognitive Science" and placed between neuroscience and psychology. The arrows pointing right connect the two most prevalent associated theories "Symbolism" and "Connectionism." These theories represent further sub-levels of explanation within cognitive science. Thus notions of levels of explanation are equally as important within cognitive science as they are within philosophy, psychology, or neuroscience. The top-down and bottom-up arrows, as before, show the movement of research knowledge between major research disciplines. Another box containing the label "Computer Science" appears to the left of the cognitive science box.

As shown by the position of cognitive science in Figure 7, it is deemed to be a more reductionistic approach to understanding the mind than psychology. As Sejnowski and Churchland (1989) point out:

> Assuming that there are a number of levels of organization in nervous systems, such that cognitive science specifically addresses higher levels whereas neuroscience typically addresses lower levels, we can acknowledge this joint effort by saying that the goal is to figure out how the mind-brain works. In this sense, an ultimate goal is the reductive integration of the psychological and neurobiological sciences, and thus cognitive neuroscience is a genuinely interdisciplinary undertaking (LeDoux and Hist 1986). Reduction here does not entail elimination, any more than the reduction of chemistry to physics entails the elimination of chemical principles. On the contrary, an integrative reduction between theories at different levels can provide insights that enrich the principles at both levels (Churchland 1986). (p. 344)

The reasons for placing computer science to the left of cognitive science are twofold. First, the research themes in computer science rarely have anything to do with understanding how the mind works. In fact, most computer scientists, with the exception of some in Artificial Intelligence, care little about understanding the mind. The second and main reason is the impact that computer science has had on investigative procedures into understanding mental processes. This impact can be felt within all four disciplines: philosophy, psychology, cognitive science, and neuroscience. In this respect computer science can be seen as a type of 'back seat driver' in the "mind investigation" game. For example, few of the scientific experiments in neuroscience could be accomplished without the aid of computer technology. PET-scans or analyzing data from a single cell experiment would be unthinkable without the use of a computer and the associated programs. In the field of philosophy the impact of computer science has been tremendous (Dreyfus, 1979; Fetzer, 1991; Moody, 1993). Winograd and Flores (1987) call into question the validity of any computational process of cognition based on the traditional philosophical stances of hermeneutics (the study of interpretation) and phenomenology (the philosophical examination of the foundations of actions).

Obviously psychology has been dramatically altered by developments in computer science. But it is the link between cognitive science and computer science which is the strongest. According to Pylyshyn (1989), a "principle characteristic that distinguishes cognitive science from more traditional studies of cognition within psychology is the extent to which it has been influenced both by the ideas and techniques of computing" (p. 51). Without the general notion of computation of an effective procedure (a computer program) it would be virtually impossible to talk about a field of cognitive science.

The concept of levels of explanation will be revisited in the context of symbolic and connectionist cognitive-modeling. Selecting the correct level to model a process is extremely important and will have direct impact on the explanatory nature of the model (Dinsmore, 1992a).

2.3.3 Models in general

Two important but often overlooked questions should be asked. The first is, why is it important to build models? The second is, at what level should one choose to model something? Churchland and Sejnowski (1992) address these very questions. With respect to the first question, models help to organize data and to postulate an explanation of the phenomenon being investigated. Furthermore, a quantitative model is appealing because assumptions can be rigorously analyzed as to their truth value. More importantly, models will increase believability as they survive tough experimental tests (Churchland, 1986a). This notion of believability is extremely important "in the pioneering days of a discipline, when data are relatively sparse, progress is closely tied to ruling out a class of models and hypotheses" (Churchland & Sejnowski, 1992 , p. 6).

The second question regarding what level one should choose to model something, is somewhat more difficult to answer. Deciding on the appropriate level of explanation depends largely on the goals one is trying to achieve. If you are a philosopher it would be highly unlikely that you would attempt to build models at the level of neuroscience. Likewise, most neuroscientists would hesitate to argue philosophical models with philosophers. Mendel's early ideas about pea plant inheritance might be considered a high level explanation of heredity. On the other hand, Watson and Crick's research on the structure of DNA could be considered a lower level explanation of heredity. Both are attempts to help describe the process of inheritance using entirely different approaches. These researchers were constrained by the limits of their knowledge, techniques, and the tools available during their investigations. What appears to be fundamentally important is that research at higher levels should not contradict what is known at a lower level (Sejnowski & Churchland, 1989). Mendel's theories were developed long before Watson and Crick discovered the structure of DNA. In a real sense Mendel's theory was at the mercy of Watson and Crick's discovery. Had something totally different been discovered which contradicted Mendel's observational research, our ideas regarding Mendelian genetics would have required reassessment.

Figure 7 outlined different levels of explanations which take the form of theories. If the ability to explain a set of phenomena means that one can describe, predict and control such phenomena with a high degree of certainty and accuracy (Borg & Gall, 1989), then one would assume theories of mind are backed by an array of models, all of which attempt to do just that. For example, how should we approach building a model which makes explicit these mysterious mechanisms embodied in Piaget's theory of a dynamic system? As corollaries to this: (a) What possible constraints does the model need to account for?, (b) Where should these constraints be in the model?, and (c) How should these constraints function in the model?

Recently, a number of individuals have advocated dynamic-modeling as a paradigm for investigating cognitive developmental change (Howe & Rabinowitz, 1994; Thelen & Smith, 1994). The advocates of dynamic-modeling describe this approach with terms like chaos theory, non-linear dynamics, and self-organization (Barton, 1994). The form these models often take are non-linear mathematical equations. A model at this level is analogous to the general linear modeling system employed in statistics and more familiar to educational researchers. The difference is that non-linear equations offer a more 'powerful' representational approach with respect to classification and should in general out perform linear statistical regression with respect to prediction (Hecht-Nielsen, 1990).

In the context of the PASS information processing model postulated by Das et al. (1994), suppose the linear modeling (factor analysis) is replaced with a more robust mathematical modeling technique employed in dynamic-modeling. What is gained? Has the level of explanation changed? The predictive power should be at least equal to or greater than before. In other words, a change in the mathematical analysis of the data may provide more support for the PASS model. On the other hand, the level of explanation of the underlying phenomena has in one small sense changed, but in one large sense not changed. The act of switching from linear mathemathics to non-linear mathemathics can be viewed as a step toward reductionism (this is the change). Simply put, the nonlinear approach is more exact (granular). What has not changed is our understanding of the processes (simultaneous and successive) as outlined by the PASS model flow-charts. How these processes might actually function still remains obscure. The point is that nonlinear dynamic-modeling must be embedded at the level of explanation of cognitive science, in other words, part of a computational cognitive model.

2.4 The mind as a computational entity

> Theories of the mind should not take so much for granted that their content is obscure. A useful constraint is to express them in a computable form, because the mathematics of computation was originally developed to show what could be computed starting from principles that are entirely transparent (Johnson-Laird, 1988 : p. 37).

Johnson-Laird's argument suggest that any theory of the mind should be expressed in computational form because it makes the theory explicit. Why should anyone accept his conclusion? This section will outline a number of early arguments that help to provide support for Johnson-Laird's statement.

2.4.1 Five early developments that influenced cognitive science

During the years surrounding World War II the human mind began to be viewed as an information processing system. The basis for this view is generally attributed to the development of computer technology in both Great Britain and United States during the war. The link between humans as processors of information and the mind as a computational device is obviously quite close, as Kenneth Craik (1943) pointed out in his book the Nature of Explanation:

> My Hypothesis then is that thought models, or parallels, reality—that its essential feature is not "the mind," "the self," "sense-data," nor propositions but symbolism, and that this symbolism is largely of the same kind as that which is familiar to us in mechanical devices which aid thought and calculation. (p. 57)

A number of key events occurred to help solidify the view that the mind could be conceived as having a computational structure. The following discussion will describe five important developments often cited in the literature. These are by no means the only significant events but in hindsight it is easy to see their impact. Figure 8 gives a timeline of events and individuals primarily responsible for them.

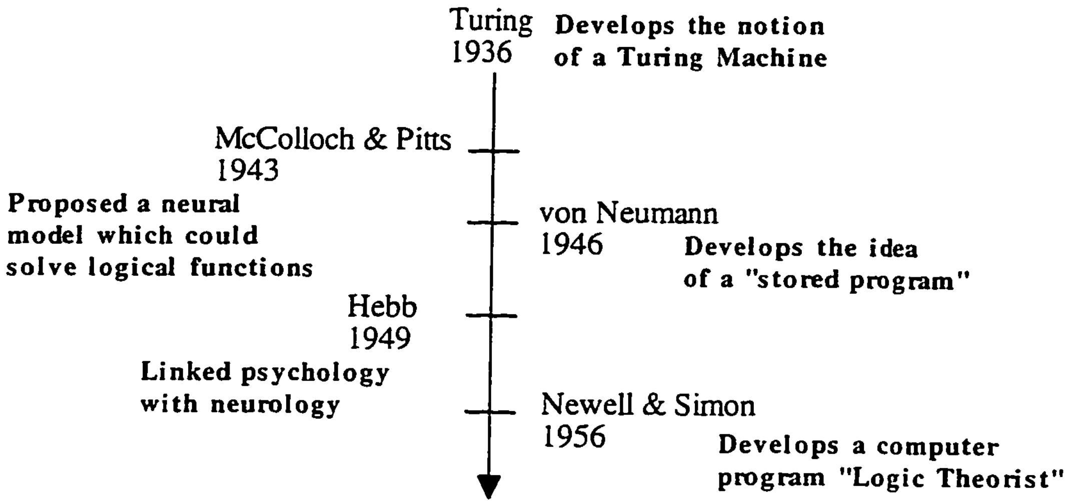

Turing 1936
Develops the notion of a Turing Machine
McColloch & Pitts 1943
Proposed a neural model which could solve logical functions
von Neumann 1946
Develops the idea of a "stored program"
Hebb 1949
Linked psychology with neurology
Newell & Simon 1956
Develops a computer program "Logic Theorist"

Figure 8: Chronology of developments prior to cognitive science

2.4.1.1 Turing machine

The early work of a young British mathematician named Alan Turing (1936) is primarily responsible for our ideas of what tasks can and cannot be accomplished by a computing device (Barwise & Etchemendy, 1993; Fischler & Firschein, 1987). All subsequent research on the relationship between computing and mental processes can be traced to Turing's original ideas on computability.

In 1936 Turing postulated the idea of a "Turing Machine," which is functionally equivalent to any modern day computer. Any process which can be called an effective procedure can be realized by his machine. An effective procedure is a set of rules (e.g., a computer program) which tells a device precisely what operations to perform in a stepwise fashion (Fischler & Firschein, 1987).

More precisely, an effective procedure can be viewed as a computational task. A computation is the evaluation of some function y = f(x), where x represents the input of data and y represents (set equal to) the function's desired output. Furthermore, x and y can be defined to represent all sorts of things like numbers, words, symbols, objects, or events. The function f(x) can be further decomposed into a sequence of steps. Any computable function is one that can be computed by a program in a finite number of steps, as a result one could refer to the function as being calculably deterministic. Therefore, if f(x) is a computable function (effective procedure) it can be computed in a finite number of elementary operations. A Turing machine can execute any computable function.

The Turing Machine contains two essential elements of any computer: (a) memory, and (b) processor (see Figure 9). The Turing machine consists of a one-dimensional tape which is of infinite length (memory). The tape is divided into square boxes. The machine has a read-write head so it can only read or write into each square. The Turing machine can carry out four basic operations: (a) it can erase a 0 and replace it with 1, (b) it can erase a 1 and replace it with a 0, (c) the read-write head can move one square left, and (d) the read-write head can move one square right. This version of the Turing machine uses a binary symbol structure as its means of representing information. What is important here is the notion that simple symbols (1 and 0) can be used to form a representation of a given concept. The letters of the alphabet could be considered an analogous type of representational symbol system but with more symbols (Barwise & Etchemendy, 1993).

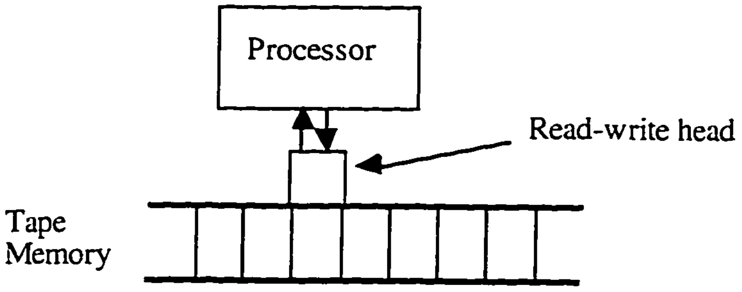


Figure 9: Simple Turing Machine

The Turing machine's instructions are controlled by a "state table" stored in rows. Each row contains five elements which comprise a single instruction. The instructions can be viewed as a condition action relationship and can be generalized as a rule: IF condition THEN action. The condition is determined by two things, the state of the machine and the symbol on the tape. The action is one of the four basic operations described above. Turing showed that a state table could be created which would solve any problem such as addition, subtraction, multiplication, and division. More complex problems could be represented in a simplified form. "Think of it this way: If a given task can be specified in a finite set of completely unambiguous instructions—also known as an algorithm—then a Turing machine can do it" (Moody, 1993 : p. 63). In other words, this simple machine could execute a program which is equivalent to that of any of today's computers.

The entire program (set of instructions) of the Turing machine can be represented by a single binary number. Thus a complete description of any Turing machine can be provided in one number. Turing proved it possible to construct a "Universal Machine" that could decode the binary number and thus simulate the operation of a Turing machine. A modern day digital computer is such a universal machine and can simulate any Turing machine. By the same token anything that can be computed by a digital computer can be computed by a Turing machine. Turing postulated that a simple Turing machine can compute any computable function, but not every function is computable. Gödel (1931) showed that there was no test to determine when a computable function would halt or whether it would continue computing forever. Furthermore, certain computable problems are classed as "difficult" in the sense that one would require an infinite amount of time and space (memory) to solve them. These types of problems are referred to as "computationally intractable" because their solution time grows exponentially relative to the size of the problem (Garey & Johnson, 1979).

A major notion in cognitive science is that the mind is a symbolic processing system and that a Turing machine can also be represented as a symbolic system. The postulated link between the Turing machine and the human mind is that they both represent symbolic processing systems (Simon & Kaplan, 1989).

The importance of Turing's work had a significant effect on the researchers of 1940s and 1950s. Gardner (1985) makes the following statement with respect to Putnam's (1960) interpretation regarding the significance of Turing's work:

> ...the development of Turing-machine notions and the invention of the computer helped to solve—or to dissolve—the classical mind-body problem. It was apparent that different programs, on the same or on different computers, could carry out structurally identical problem-solving operations. Thus, the logical operations themselves (the "software") could be described quite apart from the particular "hardware" on which they happened to be implemented. Put more technically, the "logical description" of a Turing machine includes no specification of its physical embodiment.
>
> The analogy to the human system and to human thought processes was clear. The human brain (or "bodily states") corresponded to the computational hardware; patterns of thinking or problem solving ("mental states") could be described entirely separately from the particular constitution of the human nervous system. Moreover, human beings, no less than computers, harbored programs; and the same symbolic language could be invoked to describe programs in both entities. (p. 31)

2.4.1.2 Linking neural structure and computation: McColloch and Pitts

The Turing machine as described above was a fascinating theoretical device which inspired the development of the digital computer some years later. While Turing's focus was strictly one of theoretical computation, McColloch and Pitts (1943) were interested in the link between the neural structure of the brain and computational theory. McColloch and Pitts wondered how simple logical operations could be executed by a neuron and how entire logical processes could be realized in networks of neurons. They developed a representational scheme for showing how propositional logic (where statements are either TRUE or FALSE) could be realized at the neural level. Although the logical operations they investigated may appear somewhat simple (Boolean AND and OR), it was their analogy to how such logical functions could be mapped to nerve cell firings which proved significant (see Figure 10). McColloch and Pitts' believed higher level processes could emerge out of primitive logical events. Their ideas about neuron functionality are correlated, as closely as possible, to their theoretical computational model of a neuron. Quinlan (1991) summarizes McColloch and Pitts ideas regarding the properties this computational model of the neuron should have: (a) the activity level of the neuron should be all or none, (b) in order to excite the neuron there must be a fixed number of synapses excited during a critical time period, (c) the only significant delay is synaptic delay, (d) an inhibitory synapse that is excited will prevent excitation of a neuron at that time, and (e) there is no structural change among the neurons' interconnections.

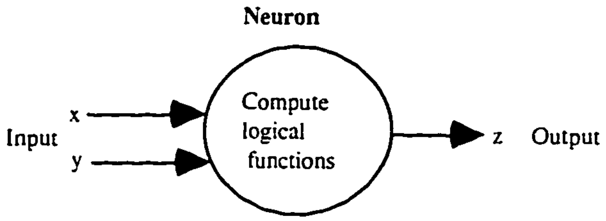


Figure 10: McColloch & Pitts' neuron

The all or none threshold level approach is ideal to model propositional logic where the binary state is either true (activated) or false (not activated). McColloch and Pitts were able to show that 14 of a possible 16 two-value propositional functions could be realized with their simple neural structure. Their work would prove to have a tremendous impact on psychologists interested in studying the mind in a computational form. Some years later, Miller, Galanter and Pribram (1960) noted the following:

> ...McColloch and Pitts had invented a formal representation of neural nets and used it to establish that any function that could be described logically, strictly, and unambiguously in a finite number of words could be realized

> by such nets. That is to say, they showed that their neural nets comprised a Turing machine. This formalization made possible some very sophisticated analyses of neurological functions and properties even before they were simulated by computers. The speculations about neural nets were widely publicized and seem to have had a stimulating effect on neurology and neurophysiology. We have every reason to expect great strides forward in this field. (p. 49)

Whereas Turing was able to theoretically show the power of computational devices, McColloch and Pitts were able to demonstrate how the brain, composed of neurons, could also computationally realize logical functions (Gardner, 1985).

2.4.1.3 von Neumann

Both the work of Turing and that of McColloch and Pitts had a significant effect on John von Neumann's (1963) work. In 1945 von Neumann proposed a computer design which incorporated a stored program. A stored program is one which resides within the computer's own internal memory. The computer could be visualized as performing different instructions without having to be rewired by hand, which was the common method of inputting instructions in early computing devices (Williams, 1985). The importance of this is easy to overlook today but for the first time a computer could be conceived as executing its own program. It now became possible, to some degree, to separate the programming of the computer from computer physical operations. This led to numerous breakthroughs in the development of so called "high level" computer languages. According to Gardner (1985), it was not clear whether von Neumann realized how powerful such systems would be in helping to solve difficult problems. Today's conventional computers are based on the von Neumann architecture.

The instruction set of a typical von Neumann computer included: (a) arithmetic operations (addition, subtraction, multiplication, and division); (b) logical operations (AND, OR, NOT); (c) comparison operations (e.g., less than or equal to, greater than or equal to); (d) branching operations (jump from one instructional sequence to another); and (e) data-shifting operations (move data from one set of memory locations to another). These basic sets of instructions operate on binary representations in a rule processing manner and give the computer its symbolic rule processing character.

Researchers now had two ways to realize the possibility of a Turing machine: (a) a digital computer of the type proposed by von Neumann which could carry out a Turing's 'effective procedure', and (b) the neural system of the brain in the form of all-or-none logic neurons described by McColloch and Pitts.

2.4.1.4 Linking neural structure with learning: Hebb

Unlike McColloch and Pitts who were interested in the relationship between logic and neural structure, Donald Hebb (1949) was interested in the relationship between psychological processes and neural structure. Hebb does make reference to some of McColloch's work on visual-area cells but not to cells as logical neurons. Hebb was primarily interested in accounting for memory in terms of structural and metabolic changes at the junctions between cells. More importantly, Hebb emphasized that simultaneous and sequential associations are extremely important in how neural cell assemblies develop. Hebb proposed the following law which was quickly adopted by modern day researchers working on connectionist models: "When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth process or metabolic changes takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased" (Hebb, 1949 , p. 62).

Hebb believed that "cell-assemblies" were the foundation for more complicated associative networks. According to Hebb two important things happened to the brain as a function of time: (a) the structure changed during exposure to external events in the world, and (b) the learning capacity changed. Therefore, the brain was viewed at the neural level as being a dynamic entity where earlier learning affects later learning.

Hebb's ideas are important for two reasons. First he emphasized that the structure and function of the brain, at a neural level, provided important insight into how the mind might actually work. Second, the link between development and learning, at the neural level, are major factors to be considered when modeling mental processes. Although Hebb really didn't view the mind as a computational device, the obvious link between his work and McColloch and Pitts would later prove to be a critical piece of information for connectionist cognitive modelers' (Caudill & Butler, 1990).

2.4.1.5 Newell and Simon

According to Gardner (1985) a conference held at the Massachusetts Institute of Technology in the summer of 1956 was a watershed in the development of cognitive science. It was at this conference that the theories of "information processing psychology" (cognitive psychology) began to take shape. Gardner (1985, p. 29) goes on to reference a quote from Newell and Simon some sixteen years later:

> One can date the change roughly from 1956: in psychology, by the appearance of Bruner, Goodnow, and Austin's Study of Thinking and George Miller's "The magical number seven"; in linguistics, by Noam Chomsky's "Three models of language"; and in computer science, by our own paper on the Logical Theory Machine. (Newell & Simon, 1972 , p. 4)

The "Logical Theory Machine" was essentially a program known as "Logic Theorist." The goal of this program was to discover proofs to theorems in propositional logic (Newell, Shaw, & Simon, 1959). In their description of the Logic Theorist, Newell and Simon (1972) discuss how they employed axioms, definitions, and rules described in the Principia Mathematica of Whitehead and Russell (1950). The program would execute an axiomatic inference type of reasoning (deduction) and could discover a number of proofs for many of Whitehead's and Russell's theorems. What is important about their work was that Newell and Simon were able to use symbols to represent elementary facts (axioms), and rules (inference procedures) to represent relationships between these facts. Computers could be shown to: (a) follow rules, (b) deduce how facts about the world affect each other, and (c) what results when facts change. In other words, computers could be viewed as simulating logical thinking (Dreyfus & Dreyfus, 1985 , p. 53).

The use of a computer program to automate theorem-proving is the first "real" (in the computer program sense) attempt to build a computational model of a reasoning process. This would have a tremendous impact on the field of psychology because it demonstrated that mental processes could be explicitly represented. With respect to Newell and Simon's work, Miller, Galanter, and Pribram (1960) make the following comment:

> It is impressive to see, and to experience, the increase in confidence that comes from the concrete actualization of an abstract idea—the kind of confidence a reflex theorist must have felt in the 1930's when he saw a machine that could be conditioned like a dog. Today, however, that confidence is no longer reserved exclusively for reflex theorists. Perhaps some of the more fanciful conjectures of the "mentalist" should now be seriously reconsidered. Psychologists have been issued a new license to conjecture. What will they do with it? Will the new ideas be incorporated into the existing theory? Or will it be easier to begin afresh?
>
> A major impetus behind the writing of this book [Plans and the Structure of Behavior] has been the conviction that these new ideas are

> compatible with, and provide extensions of, familiar and established psychological principles. (p. 56)

The simulation of logical thinking is the central theme in the development of "symbolic artificial intelligence" which is discussed in the next section.

Another significant development directly linked to Newell and Simon's work was the use of "think-aloud" protocols to investigate human subjects' ability to solve logic problems (Moore & Anderson, 1954). Researchers thus began to compare the results of human experimental data on logical problem solving with the approach used in the Logic Theorist. The link between human reasoning and machine reasoning was undeniable and formed the first clear case where the human brain could be envisioned as an information processing system similar to that of a machine (Newell, Shaw, & Simon, 1958).

2.4.2 Computational cognitive modeling

This section began with a quote from Johnson-Laird regarding the use of computational models in the study of the mind. To support Johnson-Laird's quote, the work of several early researchers, which helped to shape the foundation for the view of the mind as a computational entity, was reviewed. To provide a detailed account of the intervening years between the work of Newell and Simon (1958), and Johnson-Laird's (1988) quote would require describing a sequence of computer programs and hardware devices which attempt to model cognitive processes. A somewhat better approach is to examine the two most prevalent theories of cognitive modeling: (a) symbolism (which can be envisioned as following the research trend of Turing to von Neumann to Newell and Simon), and (b) connectionism (which can be envisioned as following the research trend of Turing to McColloch and Pitts to Hebb).

The connectionist approach assumes that cognitive processes can be simulated, on a computer, by the interaction among abstract neuron-like elements. Rumelhart (1989) states the following with respect to the connectionist research paradigm:

> Our strategy has thus become one of offering a general and abstract model of the computational architecture of brains, to develop algorithms and procedures well suited to this architecture, to simulate these procedures and architecture on a computer, and to explore them as hypotheses about the nature of the human information-processing system. We say that such models are neurally inspired, and we call computation on such a system brain-style computation (p. 134).

The intent of connectionist modelers is to replace the traditional computer metaphor (serial processing system) view of information processing psychology with the brain metaphor (parallel processing system). Both the computer metaphor and the brain metaphor can be thought of as different levels at which to seek explanations for cognitive processes. In other words, they represent different cognitive-modeling approaches (Dinsmore, 1992b).

The computer metaphor level is often referred to as "symbolic cognitive modeling." Symbolic cognitive models are usually constructed as hierarchical rule-based computer programs (Newell & Simon, 1972).[7] The program manipulates strings of symbols (input) in a serial fashion to produce output. The symbols are discrete objects like the letter "A" (i.e., there is no notion of an approximate representation for the letter A). The rules, which manipulate symbols, are almost always preprogrammed by hand as opposed to being learned. Furthermore, the symbolic level of explanation presents a definite distinction between the hardware and software. In other words, symbolic representations allow cognitive processes to be abstracted to a functional level independent of the hardware on which they are implemented. This means that the biological neural architecture, from which higher level cognitive processes must emerge, is of little or no concern for symbolic cognitive modeling (Simon, 1992).

It seems somewhat fitting to end this section with a final comment from Johnson-Laird (1988), with whom the section began:

> Students of the mind do not always know that they do not know what they are talking about. The surest way to find out is to try to devise a computer program that models the theory. A working computer model places a minimal reliance on intuition: the theory it embodies may be false, but at least it is coherent, and does not assume too much. Computer programs model the interactions of fundamental particles, the mechanisms of molecular biology, and the economy of the country. (p. 52).

[7] Traditionally the field of artificial intelligence is based on the use of symbolic level explanations of intelligent phenomena, e.g., a computer program which plays chess. As a result, educational researchers, applying artificial intelligence techniques to computer-based instructional systems, have almost entirely relied on the symbolic cognitive-modeling techniques (Ohlsson, 1988; Wenger, 1987).

2.5 Symbolism

The traditional approach to modeling cognition by computer simulations is referred to as symbolic artificial intelligence (symbolism). Two main tenets are central to this modeling approach: (a) the notion of symbols as a form of representation, and (b) the computer architecture as a metaphor for the brain. Essentially a symbol is something which stands for something else. For example, 3 is a symbol which represents the numeric concept of the quantity 3, while dog is a complex symbol which represents a physical class of objects. The basic idea is that a symbol designates something. According to Newell and Simon (1976) "symbols lie at the root of intelligent action, which is, of course, the primary topic of artificial intelligence" (p. 114). In Newell and Simon's (1972) view, it is the computer's ability to interpret symbols and symbol structures that allows for the instantiation of intelligent behavior, human problem solving to be specific, on a physical computing device. Furthermore, they postulate the abstract idea of a physical symbol system which is somewhat analogous to physical computer architecture or von Neumann architecture. A physical symbol system consists of the following components: (a) the means to store symbols and symbol structures (expressions); (b) a set of processes to manipulate the symbol structures located in memory; (c) a means of controlling the interpretation for a given symbol structure, (usually in a sequential manner); and (d) a way of dealing with input and generating output.

A system which can understand a given symbolic representation can be considered a physical symbol system. Words on paper would not be a physical symbol system even though they are syntactically made up of symbols and can be interpreted by the reader as having semantic content. Readers are considered to be the physical symbol system because they supply an interpretation to the input of words (symbols).

Two problems exist in postulating how the computer would carry out a similar interpretation process. First, the computer must have a set of rules that would operate on the symbol structures contained on the paper, i.e., words and sentences. This set of rules would be the program which defines how the computer would behave in the context of reading this paper. Where does such a program come from? According to Newell and Simon (1972) "as inductive scientists, we must discover this program in order to describe a human solving a problem" (p. 31). Second, the computer must have a set of primitive rules to begin the interpretation process. The complete set does not have to be present and the primitive set can be enhanced as a function of the program the computer actually interprets.

In keeping with this theme, Newell and Simon (1976, p. 116) postulate The Physical Symbol System Hypothesis, that simply states, "a physical symbol system has the necessary and sufficient means for general intelligent action" (p. 116). What they believe is that any system which demonstrates intelligent behavior can be shown, at some level of analysis, to be a physical symbol system.

Overall the two most prevalent symbolic modeling approaches are: (a) logic based (i.e., represent and manipulate symbolic structures using formal logic syntax and logical inference procedures); and (b) rule based or production systems (i.e., represent and manipulate symbolic structures by satisfying condition-action rules). The two approaches are closely related. In logic based systems, propositions (true or false) are stated with respect to relationships in the real world, such as "New York is a dangerous place." This proposition can be given symbolic designations and inference procedures (e.g., deductive reasoning) could be applied to manipulate symbolic structures so that truth relationships are preserved (Dinsmore, 1992b). The following deductive reasoning approach could be implemented given this simple symbolic structure:

```
city(New York)
city(X) -> dangerous(X)
```

thus by deductive reasoning X is assigned the value New York and it follows that:

```
dangerous(New York)
```

meaning, New York is dangerous. This piece of knowledge could then be added to the overall knowledge base of the system.

A production system consists of a set of rules (symbolic structures) called productions which are reducible to logical relationships. Productions consist of a condition-action relationship that defines a piece of problem-solving knowledge. The total set of rules in a production system is referred to as the system's rule base. Production systems usually contain an initial set of facts (truths, givens). The total set of facts in a production system is referred to as the system's fact base. In somewhat more familiar terms, the rule base can be considered as procedural knowledge and the fact base as declarative knowledge. The total set of rules and facts form a substantial part of the production system's knowledge base (i.e., its knowledge representation scheme). The production system works by interpreting rules sequentially whereby the rule interpretation process is viewed as the inference engine which drives the system towards a solution state. If the condition matches some symbolic structure an action will happen that manipulates a symbolic structure. For example, using the following condition-action rule:

Condition *Action*
IF city(X) THEN dangerous(X)

would add the dangerous(X) fact to the system's knowledge base, where X can be instantiated to any appropriate symbolic type. For example, X can be from the class of city types and the symbol structure "New York" would represent a valid instantiation for variable X.

2.5.1 Symbolic level of explanation

In the classical view of symbolism both minds and computers operate on representations that are symbolic. Pylyshyn (1989 : p. 57) summarizes the classical view as having the following three distinct levels of organization:

1. Knowledge level: How knowledge is used by either human or computer to achieve goals.

2. The symbol level: The structure of semantic knowledge in its symbolic form.

3. Physical level: For the system to function it must be realized in a physical form either biological or as a computer.

These levels define the cognitive architecture of the system and closely resemble Marr's (1982) three levels discussed in the section on connectionism.

The relationship between the knowledge level and symbol level is very important. It may appear that the medium of symbolic computation is a practical approach to knowledge representation given the typical von Neumann computer. Unfortunately the symbolic level of analysis doesn't describe what symbolic structures should be used. More importantly, the combination of symbolic structures together with logic may appear to be a good way to represent and reason about certain aspects of the world, but neither symbolic structures nor logic describe how to organize one's knowledge or how to reason when solving a problem (Jackson, 1990 : p. 131). In Newell's (1982) classic paper entitled the Knowledge Level he proposed the following hypothesis:

> The Knowledge Level Hypothesis. There exists a distinct computer systems level, lying immediately above the symbol level, which is characterized by knowledge as the medium and the principle of rationality as the law of behavior. (p. 99)

Here the concept of rationality refers to the idea that intelligent systems in pursuit of satisfying goals will take known actions that help to attain such goals. The idea is that "knowledge" is the vehicle for moving through various states to achieve this goal. In this sense knowledge is characterized by what it does and not how it is represented. Thus the knowledge level provides the capability to constrain the space of possible solutions within the knowledge representation reasoning framework, such that a system can achieve its goal. The relationship between the knowledge level and the symbol level is best stated in the following quote by Newell (1982):

> Knowledge, the medium at the knowledge level, corresponds at the symbol level to data structures plus the processes that extract from these structures the knowledge they contain. To 'extract knowledge' is to participate with other symbolic processes in executing actions, just to the extent that the knowledge leads to the selection of these actions at the knowledge level. The total body of knowledge corresponds, then, to the sum of the memory structure devoted to such data and processes. (p.113)

In other words, knowledge, represented by symbols, is manipulated during the reasoning process of problem solving such that a symbolic solution is formed. Thus the reasoning process is embodied in the symbols and their manipulation during problem solving.

Finally, symbolic modelers view the actual neural level of analysis of cognition as providing little insight into how models should be built. As Simon (1992) points out, "explanation of cognitive processes at the information processing (symbolic) level is largely independent of explanation at the physiological (neurological) level that shows how the processes are implemented" (p. 153). The idea is that higher level theories of cognition represent parsimonious explanations of lower level neural details. In other words, it is still possible to carry out a reduction of abstract symbolic explanations to the neural level of explanation. According to Simon (1992) this means that research on cognition "does not have to stand still with breathless expectation until neurophysiology completes its work... it can proceed with its task of explaining thought processes at the level of symbol systems" (p. 153).

2.5.2 General features of the symbolic approach

A number of recent theories about cognition that are grounded in symbolic modeling are candidates for what Newell calls "unified theories of cognition," for example, Newell's SOAR (1990) or Anderson's ACT* (1990). Such theories could fill volumes with details and discourage even a veteran researcher from further investigation. Bower and Hilgard (1981) cite the following quotation by W. R. Reitman (1964) which is still applicable some 30 years later for those who want to explore the details of either Newell's or Anderson's modeling approach:

> The description of a recent version of the Newell, Shaw, and Simon General Problem-Solving program (GPS) runs to more than 100 pages and even so covers only the main details of the system. Furthermore, the discussion assumes a knowledge of Information Processing Language V (IPL-V), the computer language in which it is written. Finally, the appendix, which simply names the routines and structures employed, takes another 25 pages. Unless one is familiar with similar systems, a thorough grasp of the dynamic properties of so complex a model almost certainly presupposes experience with the running program and its output. (p. 4)

As a result, understanding a given cognitive modeling approach depends on a given problem context, knowledge domain, representation scheme (e.g., semantic net, production system, or logical inference) and to some degree the computer language used to implement the model.

The following two assumptions are made in order to simplify the explanation of the symbolic approach: (a) the context is that of solving a problem, and (b) a rule-based approach is employed as the primary vehicle for knowledge representation. One popular approach advocated by Newell and Simon (1972) is called the means-ends analysis to problem-solving. The central idea is that any problem solving task involves a start state and a goal state. Every step, from start to goal, involves reducing the distance between the start and the goal (see Figure 11). The approach to problem solving is then one of moving through a series of sub-goals, often called a search space in the artificial intelligence literature, to reach a goal.

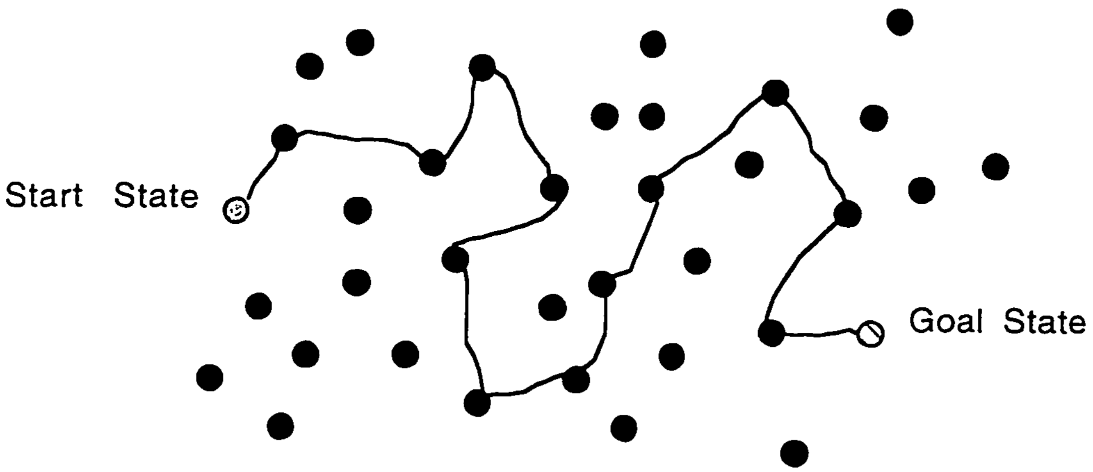


Figure 11: Typical search space

Problem-solving involves the application of knowledge to prune the size of the search space down to a manageable size and to construct a path from the start to the goal. For example, imagine a robot that must navigate through a maze filled with obstacles. The robot begins at one end of the maze with the goal of getting out. Each step in the search for the exit can be viewed as representing a state in the robot's search space. The possible number of states will depend on many factors, like the size of the maze, number of obstacles, complexity of the maze and so forth. The means-ends approach to solving this problem would suggest that with every step the distance between the 'start state' and the 'goal state' would be reduced. This would involve the robot having a comparison measure such as the current distance from its position to the goal. A simple measure of distance is a straight line between two points. The robot would, at each step in the search space compare the new current distance to that of the old current distance to see if it had been reduced.

In essence the robot's state space search strategy is very simple and based on three ideas: (a) starting state; (b) termination test, that is, has the goal been reached; and (c) set of operations that can be applied to move from state to state. One of the key factors that can greatly influence the effort required to solve the state space search problem is the selection of a good representation. Suppose we have the following problem; a child is in a room containing a chair and a bunch of cookies hanging from the ceiling just out of reach. Given these simple constraints what set of actions would the child take to get the cookies?[8] The answer appears to be relatively simple, the child walks over to the chair, pushes it under the cookies, climbs on the chair, and grabs the cookies. There is a start state (the child is in some position in the room) and a goal state (the child has the cookies). The state space is any other possible states between the start state and the goal state. What would these other possible states be? The child could at any time walk anywhere in the room, that alone would generate an infinite number of states in the search. To analyze this state space search problem we require a more formal representation of the problem. A simple method would be to constrain the number of possible actions the child could carry out. For example, the following four actions would be sufficient: (a) move to another room location {MoveTo}, (b) push the chair to a new location in the room {PushChair}, (c) climb on top of the chair {ClimbChair}, (d) grab the cookies {GrabCookies}. The bracketed terms formally define the operators for each action.

At the start state, assuming the child is not standing by the chair, the set of possible operators the child can apply consists of one {MoveTo}. By applying the MoveTo operator the child can change her physical location in the room to another position. One could imagine there could be an infinite number of positions in a room. Assuming the child uses the MoveTo operator and goes to the chair, two other operators are added to the set of possible operators {MoveTo, PushChair, ClimbChair}. The child can now move to another location (apply the MoveTo operator) or push the chair to a new location (apply the PushChair operator) or climb up on the chair (apply the ClimbChair operator). If the child pushes the chair underneath the cookies then all operators are available {MoveTo, PushChair, ClimbChair, GrabCookies}. The child is now in a position to climb the chair and grab the cookies. Figure 12 shows the simple state space required to solve this problem given our constrained representation of operations.

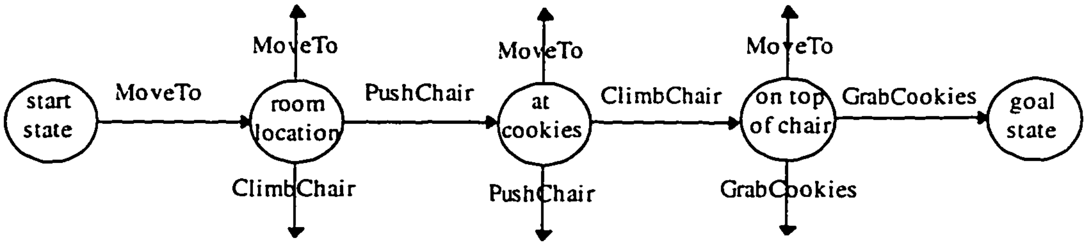


Figure 12: State space search for the child and cookie problem

[8] This problem is very similar to the "Monkey and Banana" problem see Nilsson (1971).

The following state table (see Table 1) could be used to map out the possible states.

Table 1: State table of child and cookies problem

State	Operator applied	Location of child	Location of chair	On top of chair	Has cookies
0	MoveTo	a room location	a room location	No	No
1	MoveTo	a room location	a room location	No	No
...	...	...	...	...	...
n	GrabCookies	a room location	a room location	Yes	Yes

Writing the computer model of this problem simply involves specifying the rules required to get the cookies. The important point here is that knowledge required to specify the symbolic structure of the rules and their execution, would constitute the problem solving algorithm. Such knowledge is said to reside at the knowledge level (Newell, 1982).

2.5.3 Heuristics

A variety of methods have been employed to search a given state space for possible solution states. For example, in the child and cookies problem it appears easy to see, given the problem context (a room, a child, a chair, cookies hanging from ceiling), that the solution is obvious. That is, if the child knows that pushing the chair underneath the cookies creates a situation whereby she can stand on the chair and grab the cookies then the solution is straightforward. Suppose, on the other hand, the child is blind. She may still possess the knowledge that pushing the chair underneath the cookies would allow her to reach the cookies, but to accomplish this she first has to locate (search) the chair, then push it underneath the cookies (search), then stand up on the chair (again this is a search because for every new position she pushes the chair she stands up to feel if the cookies are there) and finally grab for the cookies. The number of possible states in the search space could grow very large before the child got the cookies. What the child requires is some "rule of thumb" (i.e., a heuristic for cutting down the search space).

Returning to the problem of the robot navigating the maze of obstacles, assume the robot is blind. How could it be guided to search more efficiently? One approach is to give the robot a heuristic to apply to help prune the number of states. For example, a very simple heuristic called hill-climbing involves applying a function to determine how well the search process is going (Luger & Stubblefield, 1989). This would involve generating a state(s) in the search (the set of positions the robot could move to based on a given location in the room) and then evaluating the positions with respect to: (a) if the position is the goal state then quit; or (b) if any state in the set of possible states is better than the current state, select that state (i.e., move to that position). The problem with hill climbing is that one may encounter a local maxima on route to the goal. At a local maximum the robot has moved to a point in the maze where no further positions exist which are better than its current position and the goal state has not been reached. To solve this problem a list must be kept of the previous states (memory) which the robot could return to and resume a different search path. The best approach would be for the robot to return to a state which it had not visited but that represents the next best state. In the case of the blind child trying to find the cookies, one possible heuristic might be "directional sense of smell." This might help her to find the cookies but finding the chair will still prove difficult. If no heuristic is available the child will be left with little alternative but to carry out an exhaustive search of the room.

2.5.4 Expert systems

State space search is a valuable conceptual device of symbolic modeling. Much of the research work in artificial intelligence really amounts to developing better search strategies in specific problem domains like medical diagnosis. The area of expert systems research involves making domain knowledge explicit in an attempt to prune the number of states during search (Jackson, 1990; Luger & Stubblefield, 1989).

Expert systems are required, using their own knowledge or in conjunction with input from the user, to solve either part or all of a problem. Jacob, Gaultney and Salvendy (1986) identify two important questions when considering any expert system for use in problem solving: (a) How close does the system's decision-making (reasoning) outcomes approximate those of the expert? (b) How reliable is the knowledge in the knowledge base?

With respect to the first question, often there is a vast amount of knowledge the expert must assess before coming to a conclusion. The expert can often quickly eliminate irrelevant information and reduce the problem "search space" to the most relevant set of alternatives. This process of efficiently pruning the problem search space has proven to be a difficult task for expert system designers. For example, early systems like MYCIN (Cendrowska & Bramer, 1984) used a combination of heuristics or certainty factors (mathematical calculations), and both rule and premise ordering to solve the search space pruning problem (Clancey, 1983). Expert systems like R1 (McDermott, 1982) prune the search space with each step toward the goal, but no backtracking is allowed; as a result, the path from initial state to goal may not be optimal. R1 also requires that all information must be present at the initial state for it to generate a solution path. The DENDRAL system (Buchanan & Feigenbaum, 1978) used a weak method, "generate-and-test", along with input from the user, to prune the search space. Davis (1980) proposed a difficult to implement, "meta-knowledge" reasoning strategy to deal with problems containing vast amounts of knowledge sources. According to Jacob et al. (1986), "these efforts to achieve decision-making efficiency in expert systems are similar to selective perception which humans use to evaluate subsets of the large volume of data " (p. 136). This implies that the reasoning strategy employed by many expert systems may be subject to the same biases associated with deductive reasoning (e.g., confirmation bias). As Evans (1989) states, "the major cause of bias in human reasoning and judgment lies in factors which induce people to process the problem information in a selective manner" (p. 19).

The second question indicates that decisions from an expert system can only be as good as the knowledge base from which they are derived. Essentially, this means both the procedural and declarative knowledge must be "complete" and "unbiased" (Jacob, et al., 1986). The creation of complete and unbiased knowledge is difficult given that scientists often construct different points of view, even when exposed to the same evidence (Carrier & Wallace, 1989). Furthermore, in designing an expert system, the knowledge engineer (KE) must acquire the "expertise" from the subject matter expert. If the KE fails to identify certain biases of the expert, then these biases will be incorporated into the system. Obviously, there is a probability these biases may degrade the performance level of the expert system and thus result in error (Silverman, 1992).

2.5.5 Bias in human reasoning

One problem with modeling cognition based on logic is that humans exhibit a good deal of bias during problem-solving tasks. For example, Wason's (1966) selection reasoning task is a classic demonstration of how errors in human's deductive reasoning strategy may arise due to confirmation bias—which is the tendency to seek evidence to confirm one's theories (Evans, 1989). Although the task appears at the outset to be simple, the majority of subjects propose an incorrect solution. In this problem task subjects are shown a set of 4 cards lying on a table; each card has a capital letter on one side and a number on the other. Two cards are lying with their letters face up (letters A and D), while the other two are lying with their numbers face up (numbers 3 and 7). The subjects are then told that the experimenter has the following rule in mind: "IF there is an A on one side of the card, THEN there is a 3 on the other side of the card." The subjects are asked to determine which cards need to be turned over to find out whether the rule is true or false. Wason (1966) found that most subjects would say either the A card alone, or both A and 3 cards. The correct answer is A and 7 or all four cards, depending on the interpretation of the problem. Figure 13 shows the logical combination associated with each card where the rule is of the general form "If p then q."

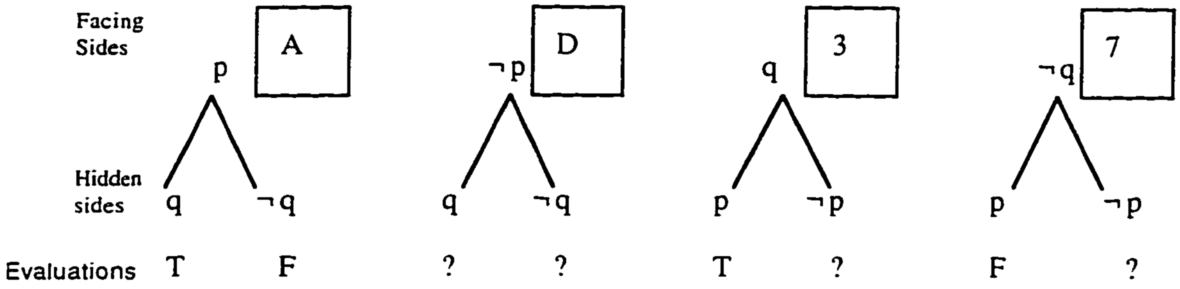


Figure 13: The logical structure and possible consequences of IF p THEN q are associated with the face up cards A, D, 3, and 7 (boxes). Adapted from Evans (1989) and where T (true), F (false), ? (irrelevant), and ¬ (not).

In this form p represents A and ¬p represents anything other than A (i.e., not A) and q represents 3 and ¬q represents any number other than 3 (i.e., not 3). To solve the task one must realize that the rule will be false if A is with a number other than 3 and that it is necessary to show any condition which might make this condition false. Notice in Figure 13 that only two cards lead to a false result p (card showing A) and ¬q (card showing 7). Wason (1966) concluded that subjects are exhibiting confirmation bias because their errors suggest they were finding evidence which "confirms" rather than "disconfirms" the rule. Figure 13 clearly shows that card A and card 3 (i.e., p and q) lead to true, while any card combination containing ¬p is irrelevant (?).

2.6 Connectionism

In recent years there has been an explosion in the use of cognitive modeling techniques employing connectionist architecture. Schneider (1987) suggests that connectionism may represent a paradigm shift for psychological researchers, similar to the move from behaviorism to cognitive psychology. Connectionism goes by several names: neural networks, parallel distributed processing (PDP), and auto-associative memory. Connectionist modeling views information processing as occurring through the interaction of a large number of simple processing units linked together in a network architecture (Quinlan, 1991). These links between processing units contain weights which specify the strength of the connection and represent the system's stored knowledge. Knowledge can be viewed as distributed throughout the network in the form of weights. The network typically learns by processing examplars (input) paired with desired output and modifying the weights through the use of an error-correcting algorithm. Essentially the network, once trained, can provide the desired output for a given input.

From a general perspective connectionism offers the possibility of modeling cognitive processes based on the notion of "biological plausibility" (Rumelhart, 1989). Biological plausibility refers to using computational models that have brain-like neural network architecture. In keeping with this brain-like metaphor, processing units in a connectionist network correspond to cell neurons. The electrochemical activation potential in a neural cell corresponds to the mathematical activation value of a connectionist processing unit. The connections between processing units correspond to synapses and connection weights to synaptic strengths. Underlying this formalism the brain is perceived as a dynamic functioning system based on locally-associated cells, each of which is assigned a specific property (Luria, 1980). This means that every behavioral or cognitive process is the result of a coordination of many different components in localized regions of the brain all dynamically interacting.

2.6.1 Connectionist level of explanation

2.6.1.1 Marr's view

According to Marr (1982), any levels of explanation of a complex information processing system (IPS) should be related so when viewed holistically they comprise a cohesive structure. Thus, Marr's is a general approach applicable to either symbolic (Pylyshyn, 1989) or connectionist modeling (Dawson, in Press). Marr proposed three levels of analysis a cognitive scientist could employ when attempting to formulate an explanation of an information processing problem (see Table 2).

Table 2: Marr's Computational Levels of Analysis

Computational theory	Representation and algorithm	Hardware implementation
What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?	How can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?	How can the representation and algorithm be realized physically?

Two key concepts are important for understanding the three levels: (a) representation, and (b) process. According to Marr, "a representation is a formal system for making explicit certain entities or types of information, together with a specification of how the system does this" (p. 20). For example, the alphabet is a representation (symbols) and the specification (rules) for putting letters together into words comprises a formal system. Marr restricts the meaning of process to "machines that carry out information processing tasks" (p. 22). The description of process involves the relationship between the three levels of analysis. One begins by analyzing the problem from a computational perspective, meaning a broad outline of what and why the IPS is carrying out a given task. The second level involves the how level of analysis. Essentially this means, given some representation like a dot pattern that represents the letter A, what algorithm might be used to recognize this pattern? The last level is simply, on what platform (physical instantiation) should this algorithm be implemented?

Marr's perspective is a top down approach to problem decomposition often used by computer scientists, that is, break the problem into smaller and smaller sub-problems. That being the case Marr believes that explanations for certain phenomena may only be available at one or two of the levels. For example, findings from neuroscience are directly tied to the physical properties that occur at the cellular level.

There are various interpretations to Marr's theoretical approach to understanding an IPS. Marr himself held the perspective that the computational level was the most important in understanding such a system. In other words, "an algorithm is likely to be understood more readily by understanding the nature of the problem being solved than by examining the mechanism (and the hardware) in which it is embodied" (p. 27). In contrast the following statement made by the connectionist modelers Rumelhart and McClelland (1986c) emphasizes the importance of the implementation level in modeling:

> We believe that PDP models are generally stated at the algorithmic level and are primarily aimed at specifying the representation of information and the processes or procedures involved in cognition. Furthermore, we agree with Marr's assertions that "each of these levels of description will have their place" and that they are "logically and causally related." Thus, no particular level of description is independent of the others. There is an implicit computational theory in PDP models as well as an appeal to certain implementational (physiological) considerations. We believe this to be appropriate. It is clear that different algorithms are more naturally

> implemented on different types of hardware and, therefore, information about the implementation can inform our hypotheses at the algorithmic level. [underlining added] (p. 122)

Clearly, Marr's top-down approach to understanding information processing functions may not be the starting point for uncovering representations and algorithms. Many recent connectionist modelers have doubts as to whether Marr's levels of analysis is appropriate given the brain levels of organization (Churchland & Sejnowski, 1989). That being said, any algorithm must still be capable of realizing the what and why of the computational level.

2.6.1.2 Smolensky's view

Unlike Marr, who approaches issues of level of analysis from a general IPS perspective, Smolensky (1988) is more specific. Smolensky suggests that cognitive modeling can be viewed at three levels: (a) symbolic, (b) subsymbolic (connectionism), (c) neural level (brain). The subsymbolic level is viewed as being much closer to the neural level than is the symbolic level. In Smolensky's view the symbolic approach represents too high a level of abstraction on which to model cognition. The basis for his argument stems from the fact that symbolic models "have taken structure from the syntax of formal languages, and their content from the semantics of natural language" (p. 4). Knowledge (rules and facts), formalized in linguistic structures and represented by symbols, is viewed as cultural knowledge. At one level humans are considered to be conscious rule interpreters that sequentially process this cultural knowledge. "In short, when people (e.g., novices) consciously and sequentially follow rules (such as those they have been taught), their cognitive processing is naturally modeled as the sequential interpretation of a linguistically formalized procedure" (p. 4). Smolensky argues that one problem with formalizing knowledge symbolically is that it offers a poor representation for an individual's "intuitive" knowledge. Facts and rules are thought of as explicit knowledge, while intuitive knowledge is viewed as implicit. For example, novice versus expert problem solving is not just a matter of the experts having more rules and facts (explicit knowledge) at their disposal. If that were the case, creating a reasonably good Expert System would be far less difficult than it is (Jackson, 1990).

Smolensky suggests that the subsymbolic level is a more appropriate way to model intuitive thought because it is based on sets of dynamic continuous variables (weights) and nonlinear equations (activation equations). Subsymbolic systems are thus continuous whereas symbolic systems are viewed as discrete systems. The nonlinear approach to knowledge representation appears to capture a part of the "nebulous" reasoning space between two discrete rules. As a result, subsymbolic systems subsume the symbolic, that is, a symbolic system is reducible (within a degree of approximation) to a subsymbolic system.

The relationship between the neural level and the subsymbolic level is addressed by Smolensky in the following two ways: (a) syntactically, (b) semantically. From a semantic perspective there is little that can presently be said because so little is known about the relationship between higher cognitive functions and neural structure. In this sense, relatively precise subsymbolic modeling (i.e., closer one-to-one mappings between brain and the model) of higher level cognitive functions (e.g., language) is still far from being a reality. Therefore, if one views the subsymbolic model's architecture (syntactic description) as "well-suited" for describing general neural processes, than "the best subsymbolic models of a cognitive process should one day be shown to be some reasonable higher-level approximation to the neural system supporting that process" (p. 10). This is much like Mendel's explanation of pea plant inheritance being a reasonable approximation of the inheritance model postulated much later by Watson and Crick. Smolensky provides a general comparison between the brain structure and the subsymbolic paradigm (see Table 3). The "+" symbol indicates the two systems closely approximate one another, the "-" symbol indicates a weak approximation.

Table 3: Brain vs. Connectionist model

Cerebral cortex		**Connectionist dynamical systems**
State defined by continuous numerical variables (potentials, synaptic areas,...)	+	State defined by continuous numerical variables (activations, connection strengths)
State variables change continuously in time	+	State variables change continuously in time
Interneuron interaction parameters changeable; seat of knowledge	+	Interunit interaction parameters changeable; seat of knowledge
Huge number of state variables	+	Large number of state variables
High interactional complexity (highly nonhomogeneous interactions)	+	High interactional complexity (highly nonhomogeneous interactions)
Neurons located in 3-d space have dense connectivity to nearby neurons; have geometrically mapped connectivity to distant neurons	- - -	Units have no spatial location uniformly dense connections
Synapses located in 3-d space	-	Connections have no spatial location
Distal projections between areas have intricate topology	-	Distal projections between nodes pools have simple topology
Distal interactions mediated by discrete signals	-	All interactions nondiscrete
Intricate signal integration at single neuron	-	Signal integration is linear
Numerous signal types	-	Single signal type

In sum, Smolensky helps to clarify the relationship between traditional AI symbolic techniques, connectionist modeling techniques, and the brain's actual implementation. Essentially this provides a degree of justification relative to the position occupied by the connectionist level of explanation in Figure 7.

2.6.1.3 Churchland's and Sejnowski's view

> It would be very convenient if we could understand the nature of cognition without understanding the nature of the brain itself. Unfortunately, it is very difficult if not impossible to theorize effectively on these matters in the absence of neurobiological constraints (Sejnowski & Churchland, 1989, p. 343).

Humans often build models based on observations of real world phenomena and in the process often come to a better understanding of the phenomena. A classic example is the airplane that models the flying capabilities of creatures like birds. Humans can't fly like birds but can observe birds in action, thus they realize that flight is possible, and then attempt to build a flying device. In order to fly, a bird must deal with the constraints of nature, for example gravity and air friction. Early attempts to build flying machines produced devices which had no more chance of flying than humans would if they just started flapping their arms. What was needed was the right level of representation (explanation) which satisfies enough (minimum set) of the constraints to get the flying device off the ground. An understanding of these constraints can often be expressed in terms of mathematical equations that help to predict behavior (e.g., Bernoulli's equations that are often used to describe the flow of air over a wing, or Newtonian laws of gravity), and allow for the designing, building, testing, and refining of models.

Computational models that employ the brain type architecture as a metaphor are crude attempts to model cognitive processes by incorporating bottom-up constraints from the field of neuroscience. As Churchland and Sejnowski (1992) state, "emergent properties [such as vision or cognition] are high-level effects that depend on lower-level phenomena" (p. 2). Furthermore, "unless our theorizing is geared to mesh with the neurobiological data, we risk wasting our time exploring some impossibly remote, if temporarily fashionable, corner of computational space" (Churchland & Sejnowski, 1989, p. 45).

To emphasize this point Churchland and Sejnowski (1992) provide an excellent example from the work of Selverston and Moulins (1987) who did extensive research on the stomatogastric ganglion of the spiny lobster. They thoroughly studied all 28 network neurons responsible for muscular control of the teeth of the gastric mill which grinds food during digestion, and compiled a detailed description of the electrochemical and connectivity patterns of the 28 neurons. Stomatogastric ganglion research is well documented in neurobiology, and if the purely neuroscience approach to understanding a phenomenon works anywhere, one would expect it to work here (Churchland & Sejnowski, 1992).

The spiny lobster's network of neurons produces a muscular rhythmic pattern during the food grinding process. What appears to be lacking is a complete understanding of how the network functions holistically to generate the rhythmic pattern. In other words, somehow all the neurons in the network function together to produce a rhythm or change a rhythm under various biochemical states. So while overall properties of the network may rely on specific properties of individual neurons it appears to be combinations of these properties that account for the complex dynamics of the system. "What the stomatogastric ganglion seems to be telling us is that we need to figure out the interactive principles governing the system, and that although interactive hypotheses should be constrained by microlevel data, their job is to characterize higher level features" (Churchland & Sejnowski, 1992, p. 5). Churchland and Sejnowski conclude as Marr did that microlevel data are necessary to understand the system but not sufficient. As Marr's (1982) states:

> There must exist an additional level of understanding at which the character of the information processing tasks carried out during perception are analyzed and understood in a way that is independent of the particular mechanisms and structures that implement them in our heads. This was what was missing—the analysis of the problem as an information processing task. Such analysis does not usurp an understanding at other levels—of neurons or of computer programs—but it is a necessary complement to them, since without it there can be no real understanding of the function of all those neurons. (p. 19)

Selverston (1988) postulates a computational neural network model as a way to explain the overall rhythmic patterns of the stomatogastric ganglion. Thus Selverston's conclusion is that the level of explanation appears to lie just above that of neuroscience, (i.e., at the cognitive science level in the form of a connectionist model). The details obtained by neuroscience researchers help in the construction of more biologically plausible computational models. Without these low level details an explanation of the phenomenon may be impossible to construct in an abstract representational form.

One of the main tenets of Churchland and Sejnowski's work is that a good model will be constrained by research results from neuroscience as well as those from cognitive psychology. "Consequently, cross-disciplinary research, combining constraints from psychology, neurology, neurophysiology, linguistics, and computer modeling, is the best hope for the co-evolution that could ultimately yield a unified, integrated science of the mind-brain" (Churchland & Sejnowski, 1989, p. 45). Thus connectionist models should incorporate both top-down psychological and bottom-up neuroscience constraints. The importance of top-down constraints, in the form of data gathered from observable psychological phenomena, should not be overlooked. Research at a higher level of explanation in experimental psychology is significant because it can determine what needs to be explained at lower levels. With respect to modeling Churchland and Sejnowski (1992) provide the following words of advice for connectionist modelers:

> Every level needs models that are simplified with respect to levels below it, but modeling can proceed very well in parallel, at many levels of organization at the same time, where sensible and reasonable decisions are made about what detail to include and what to ignore. There is no decision procedure for deciding what to include, though extensive knowledge of the nervous system together with patience and imagination are advantageous. The best directive we could come up with is the woefully vague rule of thumb: make the model simple enough to reveal what is important, but rich enough to include whatever is relevant to the measurements needed. (p. 137)

2.6.2 Connectionist systems details

There are a wide variety of architectures and learning techniques used by connectionist modelers (Gallant, 1993; Hecht-Nielsen, 1990; Lippmann, 1987; Rumelhart & McClelland, 1986b). The purpose of this section is to provide a brief overview of the general connectionist framework. What follows is a basic description of the most typical type of network architecture, and most often cited network learning procedures.

2.6.2.1 Basic description of connectionist systems

A network consists of a set of autonomous processing cells that are joined by directed arcs. This pattern of connectivity can be thought of as the network topology. Figure 14 shows a very simple network consisting of two cells.

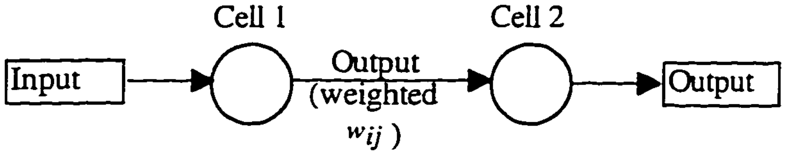


Figure 14: Simple two-cell connectionist network

Between each of the cells there is a weight value w_{ij} that indicates the influence (effect) cell j (cell 1) will have on cell i (cell 2). Thus, $w_{21} = 3$, would mean that cell 1 has a weighted influence of 3 on cell 2. A positive weight value indicates reinforcement between two cells while a negative weight value will indicate inhibition. Initially, network weights are assigned randomly. These weights are adjusted through a process of training using a learning rule. The final values of all the weights in the network will determine how the network will behave; as a result, these final weights values are analogous to a computer program. For example, given an untrained network and a set of weights from a trained network, and assuming the networks both have the same configuration, we could assign the weights from the trained network to the untrained network and have a fully operational connectionist network. This would be exactly the same as writing a computer program on one machine and executing it on a different machine. In psychological terms the weights of the network represent a learned knowledge base. Thus network weights and knowledge representation can be viewed as being synonymous.

Networks usually consist of at least two or more layers of cells with one layer being the input layer and another the output layer (see Figure 15). The output layer is the source of output for the entire network and all lower layers are linked in some way to the cells of the output layer. Cells between the input and output layers are called hidden layer cells because they are hidden from the external world of input and output.[9] In a massively

[9] To be consistent the input layer is counted as a layer, thus the network in Figure 14 consist of three layers.

parallel network, as is Figure 15, all cells at one layer are connected to all other cells at the next highest layer.

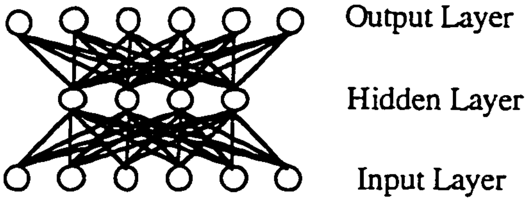


Figure 15: Typical three-layer network

Information the network receives is represented as numerical data, usually in the form of binary numbers (zeros and ones), and enters the network via the input layer cells. This begins a chain of events which spreads (propagates) throughout the network. The end result is a set of numerical values produced by the output cell(s). A network where all cells are linked together such that the spreading effect occurs in one direction (acylic fashion from input cells to output cells) is usually referred to as a feed-forward network.[10]

Generally a typical cell in a connectionist network can be viewed in two parts. Figure 16 shows a generalized representation of a connectionist cell, in which the lower part of the cell receives input from other cell(s). The input received from all lower level connecting cells is combined at the input part of the cell. The output part of the cell takes the combined input and transforms it in such a way to produce an activation value (output) for the cell. This activation value is then passed forward to all cells connected at the next higher level. At each cell at the next higher level the activation value is uniquely weighted depending upon the cell to which the connection is made.

[10] Another type of network design that has received a great deal of attention is called a recurrent network. In these networks the cells can feed back into themselves or back to cells in a previous layer. These types of networks are usually employed to represent temporal relationships between patterns (Elman, 1990).

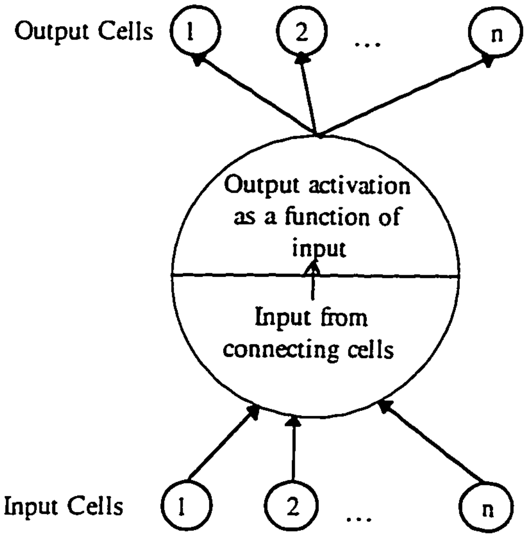

Output Cells
1
2
...
n
Output activation as a function of input
Input from connecting cells
Input Cells
1
2
...
n

Figure 16. Typical processing cell in a connectionist network connected between input and output cells.

In sum, Rumelhart and McClelland (1986) define a connectionist network to contain the following eight properties: (a) a set of processing cells; (b) a state of activation where each processing cell changes its state through time, (c) an output function for each cell; (d) there is a pattern of connectivity among cells where each connection is weighted; (e) a rule is required to propagate values through the network; (f) an activation rule is necessary to define a cell's activation state; (g) the weights between cells are modified as a function of a learning rule; and (h) the network must function in an environment meaning the set of inputs it must process. This set of properties is used widely by other researchers in the connectionism field when discussing network architecture.

2.6.2.2 Network learning

The majority of network learning consists of making systematic changes to the weights in order to improve the network's performance. Improving the network performance means improving the response of the network to a class of inputs. Although there are a variety of ways in which networks can learn (Gallant, 1993; Hecht-Nielsen, 1990) this section will examine only three approaches to learning grounded in the ideas of associative learning. The first is often call Hebbian learning which is attributed to the work of Hebb (1949) and better known as Hebb's rule.

The second approach to learning is called the delta rule and was developed to overcome some of the limitations of Hebbian learning (Rosenblatt, 1958; Widrow & Hoff, 1960). The final approach to learning is the generalized delta rule or learning by error back propagation (Rumelhart, Hinton, & Williams, 1986). All three are linked to one another by the same line of mathematical reasoning, that being network output error minimization. In this sense all networks are looking for an optimal solution, if one exists, within a given degree of tolerance.

2.6.2.2.1 Pattern associators

One way to view weight adjustments is as changes to the memory structure. Information is stored in the weighted connections. In this context the relationship between weights and output can be viewed as a form of associative memory. Associative memories lie at the centre of much of the connectionist research efforts (Brown, Hulme, & Dalloz, 1996). An associative memory system would store information by associating it to or correlating it with other information in memory. Most models of associative memory can recall information based on limited noisy input. Humans exhibit features of associative type memory all the time. For example, with only a small piece of information many of us can reconstruct events from the past. This reconstructive nature of memory is an important feature of associative memory. Furthermore, associative memories are robust in the sense that their performance does not degrade appreciably if some of the interconnections between associations are damaged (Caudill & Butler, 1990). This robustness is often referred to by connectionists as graceful degradation.

Pattern associators are the simplest associative memories used by connectionist models. They consist of an input layer of cells and an output layer of cells. Linking these two layers is a set of modifiable weights. No hidden layer exists in this system. Figure 17 shows a typical pattern associator.

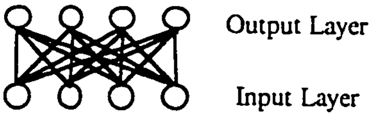


Figure 17: A typical pattern associator connectionist network

Normally the procedure is to take a set of input patterns (exemplars) and train the network to associate it with a corresponding set of output patterns. Learning in a system such as this requires weights to be updated so that when an input pattern is presented to the network it produces the correct output. It is important to recognize that the network does not contain internal replicas of the output for each given input. The only stored information (memory) the network contains is in the values of weights. More formally, McClelland et al. (1986) state that pattern associators are "models in which a pattern of activation over one set of units [input cells] can cause a pattern of activation over another set of units [output cells] without any intervening units to stand for either pattern as a whole" (p. 33). In other words, a pattern associator need not contain a specific cell that represent a concept (e.g., dog). In this sense memories are reconstructed by transforming input patterns as a function of the weighted connections (Quinlan, 1991). This is also true for three layer networks.

2.6.2.2.2 Hebb's rule

The main idea of the pattern associator is, given a set of input patterns and a set of target output patterns, the network can be trained to associate a given input to its target output. In this scenario there is a one-to-one relationship between the input and output patterns. If the network has properly learned the associations between input and output, it will be able to generate the correct output for each associated input. The simultaneous presentation to the network of input and desired output patterns (supervised learning) is a critical sequence of information in the training of the network.

Interestingly, one of the ways weighted connections can be learned was suggested by Hebb (1949). Hebb's law, discussed previously, was quickly adopted by connectionist researchers seeking a way to implement learning into the pattern associator. Hebbian learning or Hebb's rule, is probably the most often employed concept by connectionist modelers (Caudill & Butler, 1990). A typical connectionist version of Hebb's rule is: "When unit A and unit B are simultaneously excited, increase the strength of the connection between them." (McClelland & Rumelhart, 1986 , p. 36).

According to McClelland and Rumelhart (1989), if the connection is positive the change between cells is excitatory, and if the connection is negative, the change between cells is inhibitory. Notice this interpretation adds a significant component to Hebb's law (i.e., the concept of inhibition). Hebb ignored the inhibition process and his account of learning was based only on the neural growth resulting from the excitation between two cells. His explanation was a simple feed-forward spreading activation account of learning at the cell level. Furthermore, according to Quinlan (1991), Hebb's rule doesn't account for two constraints. First, modern neurophysiological findings suggest localized electrochemical changes within the cell are very important. Second, attempts to simulate Hebb's rule directly as stated on the computer have failed because weights between cells can rise to infinity. This poses a problem for a computational system that is trying to account for the learning process within finite limits.

There are a variety of ways that have been proposed to mathematically formalize Hebb's rule (McClelland & Rumelhart, 1989). The simplest one is that a change in the weight can be based on the product of input and output patterns along with the product of a learning parameter. This learning parameter is usually 1/n where n is the number of input cells. More formally, Hebb's rule can be stated by the simple mathematical expression:

$$\Delta w_{ij} = \varepsilon(input_i \times output_j),$$

where Δw_{ij} is the change in weight, i is the ith element of the input vector, j is the corresponding j element of the output vector, and ε is the learning parameter.

A well known limitation of adjusting weights in a pattern associator using Hebb's rule is that the input patterns must be orthogonal to each other. It can be shown that a system such as this can learn to correctly classify only those input vectors which are orthogonal to one another. From a mathematical perspective two vectors are said to be orthogonal if their dot product is 0. This means that all vectors in the input set must be at right angles to each other. Input patterns which are similar to other inputs (not at right angles) could not be classified into their own separate categories. From a statistical perspective, if two standardized variables X and Y are correlated their respective vectors must have an angular relationship less than or greater than 90°. On the other hand, if X and Y are completely uncorrelated their angular relationship must be exactly 90°.

2.6.2.2.3 Delta rule learning

An attempt to overcome the limitations of requiring that input patterns be orthogonal is embodied in delta rule learning. A constraint for delta-rule learning is that all input patterns (vectors) form a linearly independent set, meaning there is no vector in the set of input vectors which is a linearly weighted sum of the other vectors in the set. The idea is to adjust the weights so that they will reduce the difference or error between the output from the network and the target to be learned, in other words:

$$\text{error} = \text{target - output activation}$$

Such a system is guaranteed to find a set of weights that correctly classifies a linearly independent set of input patterns. The set of problems which can be solved by a network system using the delta rule are known as linearly separable problems. Essentially what this means is that there exists a plane or hyperplane which can be used to categorize a given input in the space of all possible inputs. For example, the two layer network using the delta rule cannot solve the Boolean function "exclusive or" (XOR) because there is no line which can divide the function space to produce the right answer. Figure 18 shows the simple Boolean functions AND, OR, and XOR (i.e., this or that but not both). Notice for the two functions AND and OR a line can be drawn to carry out the correct categorization.

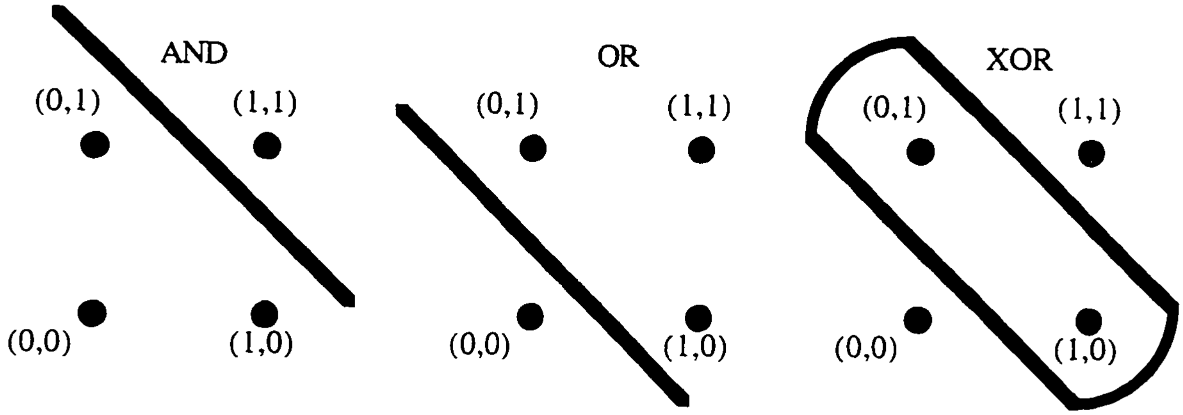


Figure 18: Representation of the separation of the Boolean functions (AND, OR, XOR).

2.6.2.2.4 Generalized delta rule

Minsky and Papert (1969) showed mathematically that one and two layer networks could never solve problems that were not linearly separable. In other words, non-linear problems like the XOR classification problem could not be solved in a two-layer network using the delta rule. In real life most problems are not linearly separable because there is no single line which can be drawn between many phenomena. Thus Minsky and Papert's solution asserted that networks could not be used for non-linear problems because they could not see any way in which weights of a network of three or more layers could be updated, the main reason being there was no way of knowing what the desired output should be for the middle layers of the network. As a result, updating the weights between the inputs and hidden layers was thought to be impossible.

The development of the generalized delta rule (or back propagation training algorithm) was important because it demonstrated that there is a procedure which can update the weights in the multilayer networks to solve non-linear problems. Over 80% of all current connectionist projects make use of this algorithm (Caudill & Butler, 1992). Although the origins of the algorithm are contested, much of the reason for its success and popularity can be traced to the PDP research group in San Diego (Rumelhart & McClelland, 1986b).

Back propagation modifies the network in a two step process, (a) input is presented to the network and flows throughout the network exactly as it did in Figure 15, and (b) the error result for output cells is calculated between the desired output and current output exactly the same way as was done with the delta rule, but in this case a function of the error is propagated back through all layers of the network for corrective adjustment of the weights. Thus all the weights of the network are updated.

One important factor was that Rumelhart et al. (1986) made the activation function of each cell a continuous non-linear function. The most often used function is the logistic function or S shaped function:

$$activation_j = \frac{1}{1+e^{-x_j}}$$

where x_j is the weighted sum of inputs at cell j. The desirable feature of the logistic function is that it is fairly easy to calculate the rate of change (slope) at any point in the function because it is easy to find the derivative. What one wants to know is how the rate of change of any weight in the network will affect the amount of error in the output. The detailed calculations of how the weights between layers are updated can be found in Rumelhart et al. (1986). It can be shown by mathematical proof that networks of this type will converge to a solution, if such a solution exists, by the method of gradient descent (Rumelhart, et al., 1986, p. 322). The gradient descent approach to problem solving is the reverse of the hill-climbing approach discussed previously in the context of mean-ends analysis. Essentially the idea of learning by gradient descent is one of adjusting the weights in the network to minimize the error between desired output and actual output. This concept is exactly the same as in the delta rule. The difference is that gradient descent, using the back propagation algorithm, can modify the weights between more than one layer, thereby overcoming Minsky and Papert's problem of non-linearity. The gradient descent method via back-propagation is not without problems. For example, the system could fall into a local minima before reaching its goal of a global minima and thus produce a solution that is not optimal. A number of methods have been suggested to overcome this problem (Dawson & Schopflocher, 1992).

2.6.2.2.5 A simplified explanation of back propagation

The idea behind the back propagation algorithm is to calculate an error between network output and the correct answers for a given examplar, each time propagating a function of the error back through the network, and making the required changes to network weights. The ultimate goal is to minimize the calculated error squared.

To understand how the back propagation algorithm works in a connectionist network with hidden units, consider a network that is to be trained to recognize a representation of the concept 'dog.' Suppose the input patterns (representations) are comprised of two numbers, 0 and 1. Therefore, four possible input patterns exist: [0,0]; [0,1]; [1,0]; and [1,1]. If the network is given either pattern [1,0] or [0,1] then it is to respond 'dog' (i.e., TRUE). If the network is given either pattern [0,0] or [1,1] it is to respond 'not dog' (i.e., FALSE).

The input layer of the network is given the input patterns one at a time. Each value in the input pattern is then linked to an input cell in the network. These input cells are then linked to cells in the hidden layer; the hidden layer cells can be linked to other hidden layers or directly to the output cells. Figure 19 shows a simple network consisting of three layers. The network activity spreads from the input layer forward to each higher layer in the network in a sequential fashion. At each layer all cells simultaneously calculate their input activity and their output activation. The final result of the forward spread will be the activation of the output cell. At this point the activation value produced by the output cell will be compared to the desired output for the given input pattern. The resulting difference between the desired output and the network output will be the network error. This begins the back propagation process because a function of the error produced by the network is propagated backwards through the network—this may be thought of as the network trying to adjust itself based on feedback. The aim is to adjust the weights in such a way that when the same input pattern is presented to the network again the resulting output produced by the network will be closer to the desired output. Therefore each weight is adjusted after each input to minimize the error at the network's output.

Examining Figure 19 the input patterns are 'not dog' [0,0], 'dog' [0,1], 'dog' [1,0], and 'not dog' [1,1]. On top we see the desired output 'dog' [1], and 'not dog' [0], corresponding respectively to each input. Suppose the first input pattern for 'not dog' [0,0] enters the network through input cells 1 and 2. Suppose the output layer's only output cell (Cell 5) produces a value of 0.5 after forward spreading activation. The desired output is 0 so the error is 0 - 0.5 which is -0.5. Therefore, a function of -0.5 is propagated back through the network and used to adjust the weights. This process of presenting the input and calculating error will continue until the network has achieved an acceptable level of error, in other words the network is trained to recognize the input patterns as either 'dog' or 'not dog.'

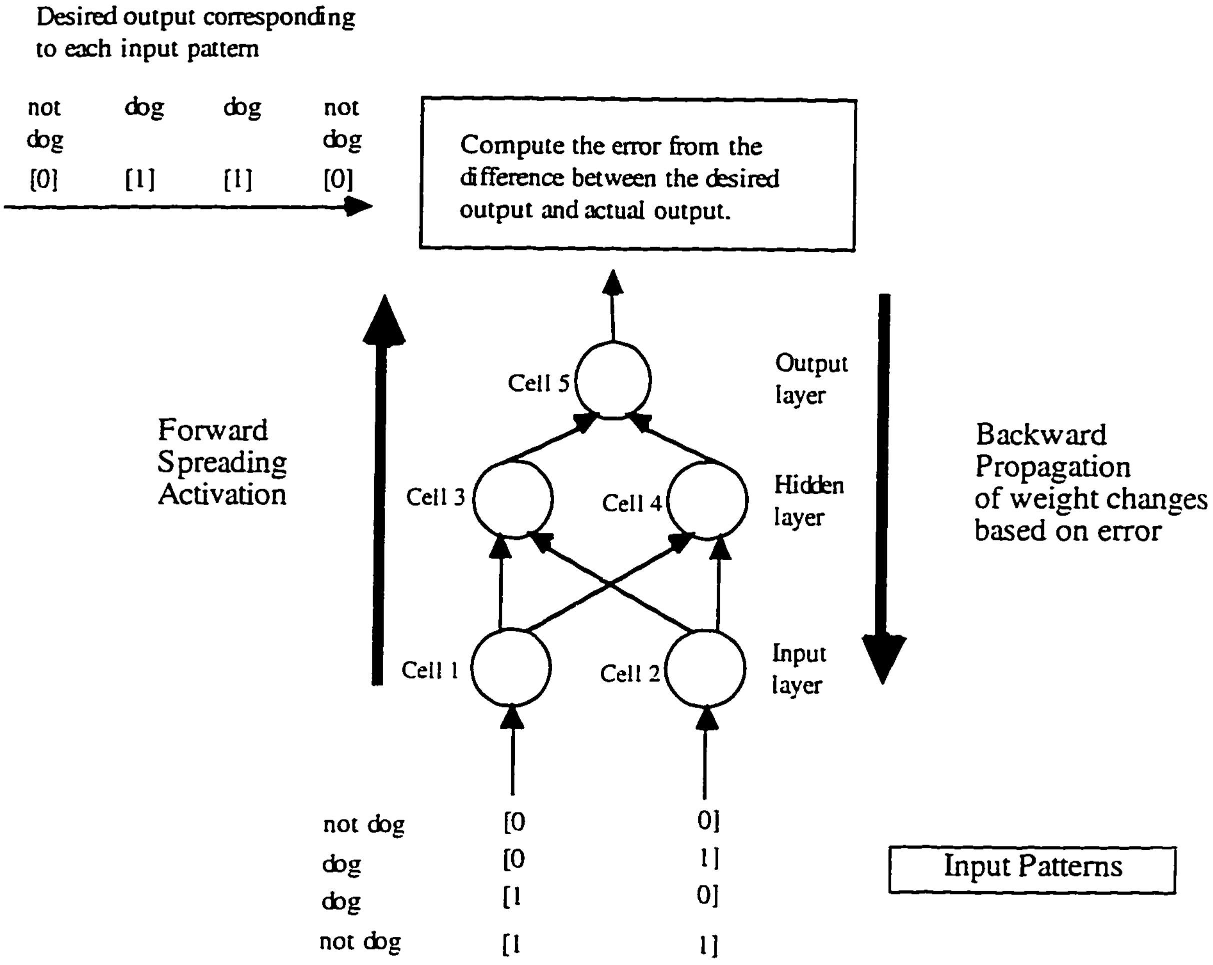

Desired output corresponding to each input pattern
not dog [0]
dog [1]
dog [1]
not dog [0]
Compute the error from the difference between the desired output and actual output.
Forward Spreading Activation
Cell 5
Output layer
Cell 3
Cell 4
Hidden layer
Cell 1
Cell 2
Input layer
Backward Propagation of weight changes based on error
not dog [0 0]
dog [0 1]
dog [1 0]
not dog [1 1]
Input Patterns

Figure 19: Three layer network illustration to learn the concept 'dog' using back propagation.

2.7 Computational modeling examples related to educational research

This section reviews a number of computational models that have strong ties to educational research particularly in the area of development. The importance of developmental theory and its relevance to educational pedagogy is well documented

(Thomas, 1985). What appears to be of most significance is that computational models may offer insight into the transition mechanisms that underlie the developmental process (Elman, et al., 1996; Simon & Halford, 1996b). For educators, this knowledge should allow for a greater understanding of why certain experiences produce cognitive change. Furthermore, precise knowledge of how developmental transitions occur should allow educators to design better instructional interventions to promote such transitions.

The application of computational modeling to study of developmental change is still very much in its infancy (Klahr, Langley, & Neches, 1987; Simon & Halford, 1996b). Although symbolic models of cognitive development have been proposed (Ling & Marinov, 1993; Schmidt & Ling, in press) such models do not attempt to offer the same level of explanation as those that employ a connectionist framework. Some have argued that differences between the two modeling approaches appears to be narrowing (Dinsmore, 1992b; MacWhinney, 1993) or at one level (Turing machine) there is no difference between either form of modeling (Honavar, 1994). Nevertheless, there are clearly implementation and information processing details that separate the two approaches.

2.7.1 Supporting arguments for connectionist modeling of development

Connectionist models offer a profound opportunity to investigate a number of issues fundamental to developmental psychology (Bechtel & Abrahamsen, 1991; Plunkett & Sinha, 1992; Schyns, 1991; Shultz, 1991; Siegler, 1991). Bates and Elman (1993) indicate the significance of such models with respect to developmental issues:

> This is the first precise, formal embodiment of the notion of emergent form—an idea that stood at the heart of Piagét's theory of change in cognitive systems. As such, connectionist systems may have the very property that we need to free ourselves from the Nature-Nurture controversy. New structure can emerge at the interface between "nature" (the initial architecture of the system) and "nurture" (the input to which that system is exposed). These new structures are not the result of black magic, or vital forces. They are the result of laws that govern the integration of information in non-linear systems.[underline added] (p. 15)

Bechtel and Abrahamsen (1991) argue the field of developmental psychology can benefit from four features present in connectionist modeling: (a) a representation of rules which is more robust than that of symbolic modeling, (b) the internal layer of connectionist networks are highly adaptive, (c) the internal units of a network impose a structure of multiple constraints that are satisfied to achieve the best overall solution, and (d) the model is resistant to damage in a similar manner to the actual neural structure (i.e., graceful degradation). Bechtel and Abrahamsen's conclusion is that connectionist models offer multiple constraints that may provide researchers with new interpretations of traditional developmental constructs and data. Similar conclusions have been stated by other connectionist researchers working on developmental problems (Elman, et al., 1996; Plunkett & Marchman, 1990; Shultz, 1991).

The final chapter of Siegler's (1991) book Children's Thinking offers support for a connectionist modeling approach in the study of development. In it Siegler outlines a number of future challenges that confront researchers in the field of child development. The first two challenges, central to understanding children's thinking, are "What develops?" and "How does development occur?" According to Siegler, these are the two most important questions faced by developmental psychologists and are fundamental to understanding what the research is all about. Siegler suggests that what is required to further our understanding of children's thinking is new theories of cognitive development on the same scale as those first postulated by Piaget. Furthermore these theories must be both "broadly applicable and precisely stated" (p. 335). He continues by saying that "perhaps the single greatest obstacle to generating more advanced theories of what develops and how development occurs is our underdeveloped knowledge of the change processes. Many current hypotheses about developmental mechanisms seem to be generally in the right direction, but are too imprecise to generate satisfying predictions or explanations" (p. 336).

For a model to qualify as a theory of developmental change mechanisms the following three components have been suggested as being important (Simon & Halford, 1996a). One is that a model should be capable of simulating the same knowledge and competence of humans on a given task. The second is that specific mechanisms that transform representations into more advanced representations need to be explicit. Finally, the behavior being modeled should account for the same intermediate outputs, both mental and physical, that occurred during observation.

From a general perspective Siegler (1991) identifies four classes of change processes as being highly relevant to understanding developmental change:

1. Automatization: Mental procedures are optimized so that other mental resources are utilized more effectively. Important issues related to automatization are parallel processing, and increases in both processing speed (time) and capacity (space).

2. Encoding: Conceptual entities (e.g. objects, events) are encoded into some type of internal representation with shared features and relations. Of primary concern for the

encoding process are issues such as discrimination, assimilation, and critical feature identification.

3. Generalization: The ability to extrapolate knowledge to novel situations. Basically this involves the process of inductive or analogical reasoning.

4. Strategy construction: The linking of various processes that adjust to the constraints for a given task. Related issues involve such things as rule formation, executive processing, and accommodation.

With respect to these four change processes Siegler states that one major issue future researchers should be concerned with is the construction of precise models of how various aspects of these four change processes may operate. He then states that, "one particularly promising effort to specify how change mechanisms operate involves connectionist models" (p. 338). Siegler continues to describe MacWhinneys, et al.'s (1989) connectionist model of how children acquire a particular aspect of German grammar. He concludes that MacWhinneys et al.'s work "illustrates how very-specific analyses of change mechanisms can yield broadly applicable conclusions about cognitive growth" (p. 342).

A major concern for connectionist modelers is the isomorphic relationship between the connectionist cognitive-model and theoretically defined developmental stages such as those of Piaget (McCloskey, 1991; Raijmakers, Koten, & Molenaar, 1996; Robinson, 1992). For example, Siegler's notion of precise models of change, relative to the four change processes he described, can be said to be a basis for an isomorphic relationship. In other words, the concept of model-precision is directly tied to the level of understanding one has in the temporally based interpretation of the model's process of representation instantiation. The more explicit (formal) one's understanding of the model, the better one can judge the model's reliability and validity relative to its structural and environmental constraints. In this sense, connectionist models offer the possibility of being analyzed at a more detailed level of specificity than flow chart or decision tree models that are often based on traditional statistical studies of cognitive psychology. In summary, McClelland and Jenkins (1991) state,

> The exploration of connectionist models of human cognition and development is still at an early stage. Yet already these models have begun to capture a new way of thinking about processing, about learning and, ..., about development. Several further challenges lie ahead. One of these is to build stronger bridges between what might be called cognitive-level models and our evolving understanding of the details of neuronal computation.

> Another will be to develop more fully the application of cognitive models to higher-level aspects of cognition. The hope is that the attempt to meet these and other challenges will continue to lead to new discoveries about the mechanism of human thought and the principles that govern their operation and adaptation to experience. (p. 71)

Both of the above challenges are directly related to the task of constructing connectionist cognitive models of development within the framework of top-down and bottom-up constraints.

2.7.2 Connectionist models of developmental processes

The application of connectionism to modeling the study of cognitive development has generally been concerned with: (a) connectionist models of language acquisition, and (b) connectionist models of stage-like concept acquisition. These two lines of connectionist research show a great deal of overlap and both emphasize the importance of understanding how a system will change over time. It is interesting to note that these two lines of research address the same two factors (conceptual thought and language) that underlie Vygotsky's (1934/1986) hypothetical model of cognitive growth. Table 4 shows a number of models designed to study development processes.

2.7.2.1 Modeling Siegler's rules for the balance scale problem

According to McClelland and Jenkins (1991) diverse developmental phenomena such as failures of conservation and compensation, progressive differentiation of knowledge, and U-Shaped learning curves in language are the result of a basic learning principle. This learning rule can be stated as follows: "Adjust the parameters of the mind in proportion to the extent to which their adjustment can produce a reduction in the discrepancy between expected and observed events" (p. 45).

It is assumed that this principle operates continuously on different tasks and throughout the cognitive system. In this sense, the learning principle "might well be seen as capturing the residue of Piaget's accommodation process, in that accommodation involves an adjustment of mental structures in response to discrepancies" (p. 45). McClelland and Jenkins state that a connectionist system using the back-propagation of error algorithm represents a system that implements this learning principle.

Table 4. Connectionist models of cognitive development

Authors	Purpose	Network Architecture
Rumelhart and McClelland (1986a)	To develop a simulation which could reproduce the rules children are said to acquire during the learning of English past tense verbs.	Pattern associator network trained with the delta-rule. Used a probabilistic activation threshold function.
Plunkett and Marchman (1990)	Similar to the work of Rumelhart and McClelland (1986a). To simulate the rules children are said to acquire during the learning of past tense.	Three layer network where all units of each layer connected with the units above (feed-forward). The network is trained using the back-propagation learning algorithm.
McClelland and Jenkins (1991)	To model the same rule characteristics that Siegler (1981) described children are said to use when solving the balance scale problem.	Three layer network configured in a nonstandard feed-forward structure. The network is trained using the back propagation learning algorithm.
Plunkett and Sinha (1992)	To model concept formation and vocabulary growth.	Four layer feed-forward network using back-propagation of error.
Schyns (1991)	To model the process of concept acquisition found in young children. The main thrust of the work was on the categorization of prototypes.	Both unsupervised [a variant of Kohonen (1982) competitive network] and supervised learning [pattern associator network] were implemented in two separate modules.[11]
Shultz (1991)	To model cognitive development of Siegler's rule stages that children employ when solving the balance scale problems.	Cascade-Correlation learning network which employs a combination delta-rule and a varient of back-propagation.

As support for their assumptions McClelland and Jenkins constructed a connectionist network that appears to demonstrate the same rule based behavior that children use when solving Piaget's balance scale problem (Siegler, 1976; Siegler, 1981). The balance scale problem involves children deciding which side of a fulcrum (containing equal number of equally spaced pegs on each side) would go down based on the weight and position of metal disks which are placed on the fulcrum pegs. Figure 20 shows a typical setup of the balance scale.

[11] Supervised learning involves comparing the network's output to the desired output and taking the difference (error) to adjust the weights. Unsupervised learning involves no performance evaluation and the network self-organizes based on a clustering of patterns. (Gallant, 1993, p. 6-7)

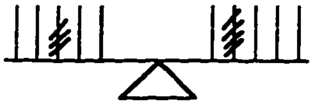


Figure 20: The balance scale with 5 pegs on each side. The left side has 3 metal disks on middle peg. The right side has 5 metal disks on the second peg from the fulcrum.

Siegler (1976) proposed the four rules that children ages five to 17 may use when attempting to solve the balance scale problem. These rules have been rewritten in symbolic form using the following abbreviations: LW (left weight), RW (right weight), LD (left distance), and RD (right distance).

Rule 1:

```
if $LW = RW$ then PREDICT BALANCE
else if $LW \neq RW$ then PREDICT HEAVIEST SIDE DOWN
```

Rule 2:

```
if $LW \neq RW$ then PREDICT HEAVIEST SIDE DOWN
else if $LW = RW$ then PREDICT WEIGHT FARTHEST FROM FULCRUM
        DOWN
```

Rule 3:

```
if $LW = RW$ and $LD = RD$ then PREDICT BALANCE
else if $LW \neq RW$ or $LD \neq RD$ then
        if $LD = RD$ then PREDICT HEAVIEST SIDE DOWN
        else if $LW > RW$ and $LD > RD$ then PREDICT LEFT SIDE DOWN
        else if $LW < RW$ and $LD < RD$ then PREDICT RIGHT SIDE DOWN
        else GUESS
```

Rule 4: (Replace GUESS in rule 3 with the following:)

```
else if $(LW * LD) = (RW * RD)$ then PREDICT BALANCE
else if $(LW * LD) > (RW * RD)$ then PREDICT LEFT SIDE DOWN
else PREDICT RIGHT SIDE DOWN
```

Learning to solve the balance scale task may involve two stage transitions, pre-operational stage to concrete operational stage and from the concrete operational to the formal operational (Raijmakers, et al., 1996). In other words, Siegler's four rules are a formal representation of this stage transition process. Siegler used the rule-assessment methodology with six types of problems to categorize the children's responses as being one of the four rules. The rule-assessment methodology attempts to classify answer patterns numerically based on the problem subjects are given.

Figure 21 shows that the connectionist network designed by McClelland and Jenkins (1991) to handle this problem was defined as a three layer feed forward architecture composed of 20 inputs units, 4 hidden units, and 2 output units (each representing one side of the fulcrum). The network was trained using the back propagation algorithm. The training set was composed of 625 exemplars and each epoch was defined to be a 100 random samples from the training set. Each exemplar (input pattern) was made up of two parts, one for weight, and one for distance. These parts are further subdivided into two more parts based on the fulcrum position (left or right side). For example, the weight side has ten inputs, five are used to describe the weight on the left side the other five are used to describe the weight on the right side. The target patterns were mapped to output units on the basis of which side of the fulcrum had the greatest torque (if both sides were balanced then the activation levels were assumed to be the same). After each epoch the network was tested on a 24 item test taken from one of Siegler's experiments.

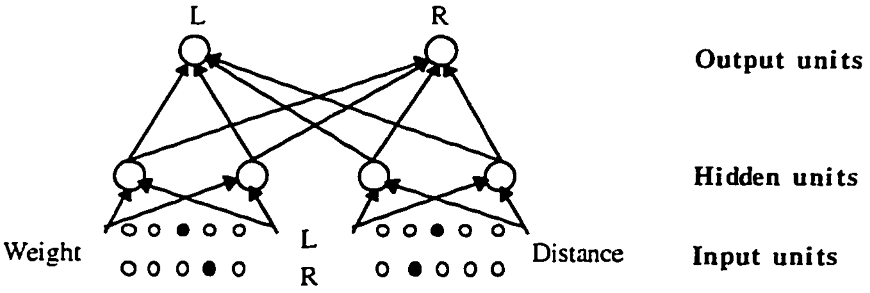


Figure 21: A diagram of the network used by McClelland and Jenkins (1991)

Two important assumptions were made by the modelers. The first is that environmental input is biased so that the weight dimension received more emphasis. This assumption was based on the idea that children have more experience with the weight dimension than the distance dimension. The second assumption is that weight and distance are viewed as separate concepts and analyzed separately. This means the model itself is required to combine both dimensions during the learning phase. Both of these assumptions appear necessary for the model to exhibit stage-like developmental performance during its learning phase. McClelland and Jenkins suggest that future research is required to determine how variations on these assumptions might change the model's performance.

Overall McClelland and Jenkin's model appears to demonstrate analogous transitional behavior of rule learning behavior for children between the ages of 5 - 17. (Siegler, 1976). That is after 100 epochs the simulation exhibited different stages of development that could be closely mapped to the four rules children appear to acquire between the ages of 5 to 17. A distinction between the final two rule stages was difficult to establish because there appeared to be a shifting back and forth between these two stages. According to McClelland and Jenkins, Siegler also noticed this stage shifting behavior.

A variety of interpretive approaches were used to substantiate the isomorphic relationship between Siegler's experimental conclusions and the connectionist models of overt behavior. One method involves evaluating output activation levels, on a test set of exemplars, at various epoch points during training. For example, at 20 epochs the model exhibited an earlier developmental rule type behavior (only considering the dimension of weight), and between epoch 20 and 40 there appears to be gradual learning of the next developmental level of rule type behavior (an understanding of the concept of distance in relation to weight). McClelland and Jenkins also provided detailed examples of how the connectionist weights between input to hidden and hidden to output changed during the rule learning process. In other words they attempted to illustrate a temporally based understanding of how the network's representation changes by a detailed analysis of the network's weight changes. A visual diagram of the weight changes depicts the magnitude of their adjustments over time (see McClelland and Jenkins 1991, p. 59). Also shown are changes in the hidden and output activation values with respect to time. A graphical representation, comparing the network's performance to that of the children, shows a consistent relationship between the network's rule learning behavior and age group use of rules. These approaches to network interpretation are significant in supporting the following claim that their model appears to capture the stagewise qualitative progression, "while at the same time exhibiting an underlying continuity which accounts for gradual change in readiness to move on to the next stage" (McClelland & Jenkins, 1991).

Shultz and Schmidt (1991) also used a connectionist network to model cognitive development on the balance scale problem. In their particular modeling task they employed the Cascade-Correlation algorithm developed by Fahlman and Lebiere (1990). According to Gallant (1993) this algorithm falls into the general class of constructive algorithms that attempt to transform the difficult process of building a network into a simpler problem of single-cell learning. Cascade-Correlation starts with an initial baseline network topology of input and output units and adds additional hidden units until the network learns to solve the problem.[12] Shultz and Schmidt suggest that the Cascade-Correlation method offers a more natural interpretation of cognitive developmental phenomena because qualitative changes are reflected in a dynamic method of network construction while quantitative changes are reflected in the changes of weights. In a series of experiments Shultz and Schmidt were clearly able to distinguish Siegler's four developmental rules. They used Hinton diagrams (Hinton, 1989) to display the weight changes at specific epochs as the network developed (added more units) and learned the balance scale rules.

2.7.2.2 Modeling English past tense acquisition

Rumelhart and McClelland (1986a) designed a connectionist network to account for the three sequential stages of past tense use of English by children. These three stages were identified by Brown (1973) and can be summarized as follows:

1. In stage one the children frequently learn to use a small number of irregular past tense verbs correctly (e.g. came, got, and went). The conclusion is that children simply know the right form of the verb.

2. In stage two children begin to acquire the past tense rule, i.e. add -ed to the stem of the verb. During this stage children often incorrectly add the -ed ending to verbs which have an irregular past tense (e.g. comed) or even to the irregular form itself (e.g. camed). This process is referred to as "overgeneralization" of a rule and the children thus appear to exhibit regression in their linguistic performance.

3. In stage three both the irregular and regular forms of the verb are used correctly. This means the children have regained the correct irregular verb performance and also understand the context of when to apply the past tense rule.

Overall the three stages produced a type of learning behavior often characterized in the form of a U-shaped curve. In other words children move from correct irregular verb use, to incorrect regular verb use, and then back again to correct irregular verb use.

Rumelhart and McClelland (1986a) constructed a simple pattern associator network (i.e., containing only input and output layers) and used a logistic probability function (Boltzman machine) to determine if a unit should be turned on (threshold unit activation). A version of the delta rule was used to train the network. A set of 506 English verb exemplars were selected and divided into three input sets: (a) ten high frequency verbs of which eight are irregular verbs, (b) 410 medium-frequency verbs of which 76 are irregular verbs, (c)

[12] Assuming that such a solution exists.

86 low frequency verbs of which 14 are irregular verbs. At different epoch points during the learning phase the verb sets were introduced to the network. Rumelhart and McClelland used a set of graphs to describe the percentage of correct features for regular and irregular verbs at various epoch training points. These graphs indicate a three stage U-shaped learning curve of the type exhibited by children.

Building on the work of Rumelhart and McClelland (1986), Plunkett and Marchman (1990) were also able to demonstrate U-shaped learning of English past tenses without any discontinuities in the network training regime. In other words, unlike Rumelhart and McClelland, different sets of input patterns were not introduced at different epoch training points. Instead Plunkett and Marchman ran a series of simulations using an incremental epoch expansion learning schedule (i.e., they trained their network on a subset of data and gradually expanded to include the complete data set). This method allows for a continuous increase in the network's knowledge base without undergoing major disruptions in learning. Interestingly Plunkett and Marchman make the following statement with respect to how pre-existing knowledge influences learning:

> Several prominent theories of cognitive development have explored the relationship between the current knowledge state of the child and the nature of the problem domain to which the child is exposed with respect to determining the child's success or failure. For example, Piagét [1953] introduces the notion of *moderate novelty* to summarize the finding that children display the most advancement in those conditions where new problems only exert moderate demand on their current knowledge state. Incorporating slightly different components, Vygotsky [1962] utilizes the notion of the zone of proximal development for similar purposes. In these simulations, the interaction between the current knowledge state of the network, as encoded in the weight matrix, and the nature and size of the problem space to which the network is exposed can be seen to account for many of the observed behaviors. Networks, like children, appear to benefit from a moderate novelty effect, such that if the initial knowledge state of the network is undifferentiated with respect to the problem at hand (e.g., a random weight configuration), overall performance is enhanced if the learning set is initially restricted and then gradually expanded. (p. 15)

This suggests that expansion training procedures can be viewed as setting into place a cognitive schema of domain knowledge that makes network learning easier (Rumelhart &

McClelland, 1986a). Unfortunately, Plunkett and Marchman indicate that the expansion training procedure is not well understood.

The importance of having a dynamic external environment in the form of an expansion training environment should not be overlooked. Bechtel and Abrahamsen (1991) point out that Neisser (1976) argued for the integration of the dynamic environmental stimuli with a dynamic information processing system. From the connectionist prespective the network is dynamic but the inputs have an invariant structure and enter the network either randomly or sequentially. "There is nothing inevitable about this arrangement, however; network processing of dynamic input patterns is an area that is ripe for exploration" (Bechtel & Abrahamsen, 1991, p. 267). Plunkett and Marchman's (1990) use of the expansion training procedure, partially addresses this concern of trying to understand a dynamic input environment. Furthermore, expansion training procedures also offer the possibility of linking the following two educational issues to connectionist modeling: (a) hierarchically structured pedagogical environment models of the type proposed by Gagne (1962; 1979), and (b) the influence of pre-existing knowledge schema on problem solving.

2.8 References

Anderson, J. R. (1990). The place of cognitive architectures in rational analysis. In K. VanLehn (Eds.), Architectures for Intelligence Hillsdale, N.J.: Erlbaum.

Barton, S. (1994). Chaos, self-organization, and psychology. American Psychologist, 49(1), 5-14.

Barwise, J., & Etchemendy, J. (1993). Turing's world 3.0: An introduction to computability theory. Stanford, CA: CSLI Publications.

Bates, E. A., & Elman, J. L. (1993). Connectionism and the study of change. In M. H. Johnson (Eds.), Brain Development and Cognition Cambridge, MA: Blackwell.

Bechtel, W., & Abrahamsen, A. (1991). Connectionism and the mind: An introduction to parallel processing in networks. Cambridge, MA: Blackwell.

Black, I. B. (1991). Information in the brain: A molecular perspective. Cambridge, MA: MIT Press.

Borg, W. R., & Gall, M. D. (1989). Educational research (5th ed.). New York: Longman.

Bower, G. H., & Hilgard, E. R. (1981). Theories of learning (5th ed.). Englewood Cliffs, N. J.: Prentice-Hall.

Broadbent, D. E. (1958). Perception and communication. Oxford: Pergamon.

Brown, G. D. A., Hulme, C., & Dalloz, P. (1996). Modelling human memory: Connectionism and convolution. *British Journal of Mathematical and Statistical Psychology, 49*, 1-24.

Brown, R. (1973). *A first language*. Cambridge, MA: Harvard University Press.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York: John Wiley & Sons.

Bruner, J. S., Olver, R. R., & Greenfield, P. M. (1966). *Studies in cognitive growth*. New York: Wiley.

Buchanan, B. G., & Feigenbaum, E. A. (1978). Dendral and meta-dendral. *Artificial Intelligence, 11*, 5-24.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and Quasi-experimental designs for research. In N. L. Gage (Eds.), *Handbook of research on teaching* (pp. 171-246). Chicago: Rand McNally & Company.

Carr, M., & Jessup, D. (1995). Cognitive and metacognitive predictors of mathematics strategy. *Learning and Individual Differences, 7*(3), 235-247.

Carrier, H. D., & Wallace, W. A. (1989). An epistemological view of decision-aid technology with emphasis on expert systems. *IEEE Transactions on Systems, Man, and Cybernetics, 19*(5), 1021 - 1029.

Caudill, M., & Butler, C. (1990). *Naturally intelligent systems*. Cambridge, MA: MIT Press.

Caudill, M., & Butler, C. (1992). *Understanding neural networks*. Cambridge, MA: MIT Press.

Cendrowska, J., & Bramer, M. A. (1984). A rational reconstruction of the MYCIN consultation system. *International Journal of Man-Machine Studies, 20*, 229-317.

Chase, W. G. (Ed.). (1973). *Visual Information Processing*. New York: Academic Press.

Churchland, P. M. (1986a). Some reductive strategies in cognitive neurobiology. *Mind, 95*, 279-309.

Churchland, P. S. (1986b). *Neurophilosophy: Towards a unified science of the mind-brain*. Cambridge, MA: MIT Press/Bradford Books.

Churchland, P. S., & Sejnowski, T. J. (1989). Neural representation and neural computation. In L. Nadel, L. A. Cooper, P. Culicover, & R. M. Harnish (Eds.), *Neural connections, mental computation* Cambridge, MA: MIT Press.

Churchland, P. S., & Sejnowski, T. J. (1992). *The computational brain*. Cambridge, MA.: MIT Press.

Clancey, W. J. (1983). The epistemology of a rule-based expert system: a framework for explanantion. Artificial Intelligence, 20, 215-251.

Clark, A. (1993). Associative engines: Connectionism, concepts, and representational change. Cambridge, MA: MIT Press.

Collins, A., & Smith, E. E. (Ed.). (1988). Readings in cognitive science: A perspective from psychology and artificial intelligence. San Mateo, CA: Morgan Kaufmann.

Criak, K. J. W. (1943). The nature of explanation. Cambridge: Cambridge University Press.

Crick, F., & Asanuma, C. (1986). Certain Aspects of the Anatomy and Physiology of the Cerebral Cortex. In J. L. McClelland & D. E. Rumlehart (Eds.), Parallel distributed processing: Explorations in the microstructure of cognition Cambridge, MA: MIT Press.

Das, J. P., Naglieri, J. A., & Kirby, J. R. (1994). Assessment of cognitive processes. Toronto: Allyn and Bacon.

Davis, R. (1980). Meta-Rules: reasoning about control. Artificial Intelligence, 15, 179-222.

Dawson, M. R. W. (in Press). The coffee room and cognitive science. Cambridge, MA: MIT Press.

Dawson, M. R. W., & Schopflocher, D. P. (1992). Modifying the generalized delta rule to train networks on nonmonotonic processors for pattern classification. Connection Science, 4, 19-31.

Dinsmore (1992a). Thunder in the gap. In J. Dinsome (Eds.), The symbolic and connectionist paradigms Hilsdale, NJ: Lawrence Erlbaum Associates.

Dinsmore, J. (Ed.). (1992b). The symbolic and connectionist paradigms: Closing the gap. Hillsdale, NJ: Lawrence Erlbaum Associates.

Dreyfus, H. L. (1979). What computers can't do: The limits of artificial intelligence. New York: Harper Colophon.

Dreyfus, H. L., & Dreyfus, S. E. (1985). Mind over machine. New York: Macmillan.

Eimas, P. D. (1994). Categorization in early infancy and the continuity of development. Cognition, 50, 83-93.

Elman, J. L. (1990). Finding structure in time. Cognitive Science, 14, 179-211.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, Parisi, D., & Plunkett, K. (1996). Rethinking innateness: A connectionist perspective on development. Cambridge, MA: MIT Press.

Evans, J. S. B. T. (1989). Bias in human reasoning: Causes and consequences. Hillsdale, NJ: Erlbaum.

Fahlman, S. E., & Lebiere, C. (1990). The cascade-correlation learning architecture (Technical No. CMU-CS-90-100). School of Computer Science, Carnegie Mellon University.

Fetzer, J. (1991). Philosophy and Cognitive Science. New York: Paragon House.

Fischler, M. A., & Firschein, O. (1987). Intelligence: The eye, the brain, and the computer. Don Mills, ON.: Addison-Wesley.

Frankel, J. R., & Wallen, N. E. (1993). How to design and evaluate research in education (second ed.). New York: McGraw-Hill.

Gagne, R. (1962). The acquisition of knowledge. Psychological Review, 69(4), 355-365.

Gagne, R. M., & Briggs, L. J. (1979). Principles of instructional design (2nd ed.). New York: Holt, Rinehart and Winston.

Gallant, S. I. (1993). Neural network learning and expert systems. Cambridge, MA: MIT Press.

Gardner, H. (1985). The mind's new science: A history of the cognitive revolution. New York: Basic Book Inc.

Garey, M. R., & Johnson, D. S. (1979). Computers and intractability. New York: W. H. Freeman and Company.

Gödel, K. (1931). Uber formal unentscheidbare Sätze der Principia Mathematica und vervandter Systeme I. Monatsh. Math. Phys, 38, 173-198.

Greenough, W. T., Black, J. E., & Wallace, C. S. (1987). Experience and brain development. Child Development, 58, 539-559.

Haugeland, J. (1985). Artificial intelligence: The very idea. Cambridge, MA: MIT Press.

Hebb, D. O. (1949). The organization of behavior. New York: Wiley.

Hecht-Nielsen (1990). Neurocomputing. New York: Addison-Wesley.

Hinton, G. E. (1989). Connectionist learning procedures. Artificial Intelligence, 40, 185-234.

Honavar, V. (1994). Symbolic artificial intelligence and numeric artificial neural networks: Towards a resolution of dichotomy (Technical Report No. TR94-14). Iowa State University of Science and Techology.

Hothersall, D. (1990). History of psychology. New York, NY: McGraw-Hill Pub. Company.

Howe, M. L., & Rabinowitz, F. M. (1994). Dynamic modeling, chaos, and cognitive development. Journal of Experimental Child Psychology, 58, 184-199.

Hunt, E. B. (1962). Concept learning: An information processing problem. New York, NY: John Wiley and Sons, Inc.

Inhelder, B., & Piagét, J. (1958). The growth of logical thinking from childhood to adolescence. New York: Basic Books.

Jackson, P. J. (1990). Introduction to expert systems (2nd ed.). Don Mills, ONT.: Addison Wesley.

Jacob, V. S., Gaultney, L. D., & Salvendy, G. (1986). Strategies and biases in human decision-making and their implications for expert systems. Behaviour and Information Technology, 5(2), 119 - 140.

Johnson-Laird, P. N. (1988). The Computer and the mind: An introduction to cognitive science. Cambridge, MA: Harvard University Press.

Klahr, D., Langley, P., & Neches, R. (Ed.). (1987). Production systems of learning and development. Cambridge, MA: MIT Press.

Klahr, D., & Sielger, R. S. (1978). The representation of children's knowledge. In H. W. Reese & L. P. Lipsitt (Eds.), Advances in child development New York: Academic Press.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. Biological Cybernetics, 43, 59-69.

Kolb, B., & Whishaw, I. (1990). Fundamentals of neuropsychology (3rd ed.). New York, NY: Freeman.

Langley, P. (1987). A general theory of discrimination learning. In D. Klahr, P. Langley, & R. Neches (Eds.), Production systems models of learning and development Cambridge, MA: MIT Press.

Languis, M. L., & Miller, D. C. (1992). Luria's theory of brain functioning: A model for research in cognitive psychophysiology. Educational Psychologist, 27(4), 493-511.

Lashley, K. S. (1929). Brain mechanism of intelligence. Chicago: University of Chicago Press.

Lavine, T. E. (1984). From Socrates to Sartre: The philosophic quest. Toronto, ON: Bantam Books.

Ling, C. X., & Marinov, M. (1993). Answering the connectionist challenge: a symbolic model of learning the past tenses of English verbs. Cognition, 49(3), 235-290.

Lippmann, R. P. (1987, April). An introduction to computing with neural nets. IEEE ASSP Magazine, p. 4-22.

Luger, G. L., & Stubblefield, W. A. (1989). Artifical intelligence and the design of expert systems. Don Mills: Benjamin/Cummings.

Luria, A. R. (1980). Higher cortical functions in man (2nd ed.). New York: Basic Books.

MacWhinney, B. (1993). Connections and symbols: closing the gap. Cognition, 49, 291-296.

MacWhinney, B., Leinbach, J., Taraban, R., & McDonald, J. (1989). Language learning: Cues or rules? Journal of Memory and Language, 28, 255-277.

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information. San Francisco: W. H. Freeman.

Massaro, D. W., & Cowan, N. (1993). Information processing models: Microscopes of the mind. Annual Review of Psychology, 44, 383-425.

McClelland, J. L., & Jenkins, E. (1991). Nature, nurture and connections: Implications of connectionist models for cognitive development. In K. VanLehn (Eds.), Architectures for intelligence: Twenty-second Carnegie symposium on cognition (pp. 41-73). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

McClelland, J. L., & Rumelhart, D. E. (1986). Parallel distributed processing explorations in the microstructure of cognition: Psychological and biological models. Cambridge, MA: MIT Press.

McClelland, J. L., & Rumelhart, D. E. (1989). Explorations in parallel distributed processing: A handbook of models, programs and exercises. Cambridge, MA: MIT Press.

McClelland, J. L., Rumelhart, D. E., & Hinton, G. E. (1986). The appeal of parallel distributed processing. In D. E. Rumlehart & J. L. McClelland (Eds.), Parallel distributed processing: Explorations in the microstructure of cognition: foundations (pp. 3-44). Cambridge, MA: MIT Press.

McCloskey, M. (1991). Networks and theories: the place of connectionism in cognitive science. Psychological Science, 2, 387-395.

McCulloch, W. S., & Pitts, W. H. (1943). A logical calculus of the ideas immanent in nervous activity. In W. S. McCulloch (Eds.), Embodiments of the mind (pp. 19-39) (pp. (Reprinted from the Bulletin of Mathematical Biophysics, 115-133)). Cambridge, MA: MIT Press.

McDermott, J. (1982). R1: A rule-based configurer of computer systems. Artificial Intelligence, 19, 39-88.

Memmi, D. (1990). Connectionism and artificial intelligence as cognitive models. Artificial Intelligence and Society, 4, 115-1136.

Miller, G. A., Galanter, E., & Pribram, K. (1960). Plans and the Structure of Behavior. New York: Holt, Rinehart & Winston.

Minsky, M. (1968). Semantic information processing. Cambridge, MA: MIT Press.

Minsky, M. L., & Papert, S. A. (1969). Perceptrons: An introduction to computational geometry. Cambridge, MA: MIT Press.

Moody, T. (1993). Philosophy and artificial intelligence. Englewood Cliffs, NJ: Prentice Hall.

Moore, O. K., & Anderson, S. B. (1954). Modern logic and tasks for experiments on problem solving behavior. Journal of Psychology, 38, 151-160.

Neisser, U. (1976). Cognition and reality: Principles and implications of cognitive psychology. San Francisco,CA: W. H. Freeman.

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. G. Chase (Eds.), Visual Information Processing New York: Academic Press.

Newell, A. (1982). The knowledge level. Artificial Intelligence, 18, 87-127.

Newell, A. (1990). Unified theories of cognition. Cambridge MA.: Harvard University Press.

Newell, A., Shaw, J. C., & Simon, H. A. (1958). Elements of the theory of human problem solving. Psychological Review, 65, 151-166.

Newell, A., Shaw, J. C., & Simon, H. A. (1959). Report on a general problem-solving program (Technical No. P-1584). The RAND Corporation Carnegie Institute of Technology.

Newell, A., & Simon, H. (1976). Computer Science as empirical inquiry: Symbols and search. Communications of the ACM, 19, 113-126.

Newell, A., & Simon, H. A. (1956). The logic theory machine. Transactions on Information Theory, IT-2(3).

Newell, A., & Simon, H. A. (1972). Human problem solving. Englewood Cliffs, N.J.: Prentice-Hall.

Nilsson, N. J. (1971). Problem-solving methods in artificial intelligence. Toronto, ON: McGraw-Hill.

Ohlsson, S. (1988). Computer simulation and its impact on educational research and practice. International Journal of Educational Research, 12, 5-34.

Piaget, J. (1954). The construction of reality in the child. New York: Basic Books.

Plunkett, K., & Marchman, V. (1990). From rote learning to system building (Technical No. 9020). University of California, San Diego.

Plunkett, K., & Sinha, C. (1992). Connectionism and development psychology. British Journal of Developmental Psychology, 10, 209-254.

Posner, M. I. (Ed.). (1989). Foundations of cognitive science. Cambridge, MA: MIT Press.

Posner, M. I., Peterson, S. E., Fox, P. T., & Raichle, M. E. (1988). Localization of cognitive operations in the human brain. Science, 240, 1627-1631.

Putnam, H. (1960). Minds and machines. In S. Hook (Eds.), Dimensions of mind New York, NY: New York University Press.

Pylyshyn, Z. W. (1989). Computing in cognitive science. In M. I. Posner (Eds.), Foundations of cognitive science (pp. 49-92). Cambridge, MA: MIT Press.

Quinlan, P. (1991). Connectionism and psychology: A psychological perspective on new connectionist research. Chicago, IL.: The University of Chicago Press.

Raijmakers, M. E. J., Koten, S. v., & Molenaar, P. C. M. (1996). On the validity of simulating stagewise development by means of PDP networks: Application of catastrophe analysis and an experimental test of rule-like network performance. Cognitive Science, 20, 101-136.

Reber, A. S. (1985). Dictionary of psychology. Toronto, ON.: Penguin Books.

Reitman, W. R. (1964). Information-processing models in psychology. Science, 14, 1192-1198.

Robinson, D. A. (1992). Implications of neural networks for how we think about brain function. Behavioral and Brain Sciences, 15, 644-655.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65, 386-408.

Rumelhart, D. E. (1989). The architecture of mind: A connectionist approach. In M. I. Posner (Eds.), Foundations of cognitive science Cambridge, MA: MIT Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumlehart & J. L. McClelland (Eds.), Parallel distributed processing: Explorations in the microstructure of cognition Cambridge, MA: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (1986a). On Learning the past tenses of English verbs. In J. L. McClelland & D. E. Rumlehart (Eds.), Parallel distributed processing: Psychological and biological models (pp. 216-268). Cambridge, MA: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (Ed.). (1986b). Parallel distributed processing: Explorations in the microstructure of cognition. Cambridge, MA: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (1986c). PDP models and general issues in cognitive science. In D. E. Rumelhart & J. L. McClelland (Eds.), Parallel distributed processing: Explorations in the microstructure of cognition Cambridge, MA: MIT Press.

Schank, R. C. (1973). Computers models of thought and language. San Francisco, CA.:

Schmidt, W. C., & Ling, C. X. (in press). A decision-tree model of balance scale development. Machine Learning, 1-30.

Schneider, W. (1987). Connectionism: Is it a paradigm shift for psychology. Behavior Research Methods, Instruments, & Computers, 19(2), 73-83.

Schneider, W., & Graham, D. J. (1992). Introduction to connectionist modeling in education. Educational Psychologist, 27(4), 513-530.

Schunk, D. H. (1996). Learning theories: An educational perspective (2nd ed.). Toronto: Prentice Hall.

Schyns, P. G. (1991). A modular neural network model of concept acquisition. Cognitive Science, 15, 461-508.

Sejnowski, T. J., & Churchland, P. M. (1989). Brain and cognition. In M. I. Posner (Eds.), Foundations of cognitive science Cambridge, MA: MIT Press.

Selverston, A. I. (1988). A consideration of invertabrate central pattern generators as computational data bases. Neural Networks, 1, 109-117.

Selverston, A. I., & Moulins, M. (1987). The crustacean stomato-gastric system: A model for the study of the central nervous system. Berlin: Springer-Verlag.

Shallice, T. (1988). From neuropsychology to mental structure. Cambridge: Cambridge University Press.

Shultz, T. R. (1991). Simulating stages of human cognitive development with connectionist models. In L. Brinbaum & G. Collins (Eds.), Machine learning: Proceedings of the eight international workshop (pp. 105-109). San Mateo, CA: Morgan Kaufmann.

Shultz, T. R., & Schmidt, W. C. (1991). A cascade-correlation model of the balance scale phenomena. In Proceedings of the thirteenth annual conference of the cognitive science society, (pp. 635-640). Hillsdale, NJ. : Lawrence Erlbaum.

Siegler, R. S. (1976). Three aspects of cognitive development. Cognitive Psychology, 8, 481-520.

Siegler, R. S. (1981). Developmental sequences within and between concepts. Monographs of the Society for Research in Child Development, 46(189), 1-74.

Siegler, R. S. (1991). Children's thinking (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Siegler, R. S. (1995). How does change occur: A microgenetic study of number conservation. Cognitive Psychology, 28, 225-273.

Silverman, B. G. (1992). Evaluating and refining expert critiquing systems: A methodology. Decision Sciences, 23, 86 - 110.

Simon, H. A. (1992). What is an "explanation" of behavior? Psychological Science, 3(3), 150-161.

Simon, H. A., & Kaplan, C. A. (1989). Foundations of cognitive science. In M. I. Posner (Eds.), Foundations of cognitive science (pp. 1-48). Cambridge MA: MIT Press.

Simon, T. J., & Halford, G. S. (1996a). Computational models and cognitive change. In T. J. Simon & G. S. Halford (Eds.), Developing cognitive competence: New approaches to process modeling (pp. 1-29). Hillsdale, NJ: Lawrence Erlbaum Associates.

Simon, T. J., & Halford, G. S. (Ed.). (1996b). Developing cognitive competence: New approaches to process modeling. Hillsdale, NJ: Lawrence Erlbaum Associates.

Smolensky, P. (1988). On the proper treatment of connectionism. Behavioral and Brain Sciences, 11, 1-74.

Thelen, E., & Smith, L. B. (1994). A dynamic systems approach to the development of cognition and action. Cambridge, MA: MIT Press.

Thomas (1985). Comparing theories of child development. Belmont, California: Wadsworth Publishing Co.

Treisman, A. M. (1964). Verbal cues, language, and meaning in selective attention. American Journal of Psychology, 77, 206-219.

Turing, A. M. (1936). On computable numbers, with an application to the entscheidungs problem. Proceedings of the London Mathematical Society, Series 2(42), 230-265.

von Neumann, J. (Ed.). (1963). Collected works: Design of computers, theory of automata and numerical analysis. New York: Pergamon.

Vygotsky, L. (1934/1986). Thought and language (newly revised and edited by Alex Kozulin, Trans.). Cambridge, MA: MIT Press.

Wason, P. C. (1966). Reasoning. In B. M. Foss (Eds.), New horizons in psychology Harmandsworth: Penguin.

Watson, J. B. (1913). Psychology as the behaviorist views it. Psychological Review, 20, 158-177.

Wenger, E. (1987). Artificial intelligence and tutoring systems. Los Altos, CA: Kaufmann.

Whitehead, A. N., & Russell, B. (1950). Principia mathematica (2nd ed.). London: Cambridge University Press.

Widrow, G., & Hoff, M. E. (1960). Adaptive switching circuits. Institute of Radio Engineers, Western Electronic Show and Convention Record, Part 4, 96-104.

Wiersma, W. (1995). Research methods in education (6th ed.). Toronto: Allyn and Bacon.

Williams, M. R. (1985). A history of computing technology. Englewood Cliffs, NJ: Prentice-Hall.

Winograd, T., & Flores, F. (1987). Understanding computers and cognition: A new foundation for design. Don Mills, ON: Addison-Wesley.

Chapter 3

3. Recasting Bruner in a connectionist framework

3.1 Introduction

An examination of any educational learning theory textbooks usually reveals sections covering Vygotsky's (1934/1986) and Bruner's theories (Bruner, 1966; Bruner, Goodnow, & Austin, 1956; Bruner, Olver, & Greenfield, 1966) concerning learning and development. Often cited is Vygotsky's postulated Zone of Proximal Development that represents the degree of possible learning students can achieve in a suitable instructional environment. Bruner (1966) suggested the idea of discovery learning where students inductively reason from specific instances to form rules and concepts.

Both Vygotsky and Bruner argued that children pass through specific stages of conceptual development. From Vygotsky's (1934,1986) perspective, the early stages of concept development involves the creation of unorganized-heaps where objects are randomly associated together. Eventually, according to Vygotsky, children begin to form more complex factual relationships between objects. These relationships develop into higher level conceptual representations directly linked to the acquisition of language skills. Bruner et al. (1966) viewed conceptual development as occurring in three representational stages: enactive (representation in the muscles), iconic (mental images), and symbolic (letters, words). Bruner believed that these conceptual stages develop sequentially and remain throughout our lives. Although currently the stage theory of conceptual development is not believed to occur the way Vygotsky, Bruner or even Piaget suggested (Eimas, 1994), at a gross level of analysis it still remains a formative measure of the developmental progress (Siegler, 1995).

Many of the conclusions arrived at by Vygotsky and Bruner were grounded in the empirical experimental results from using conceptual learning tasks. Much of their work concerned such fundamental questions as: (a) What is a concept? (b) Why do we need concepts? (c) How do we form concepts? and (d) What is the relationship between concepts and reasoning? Answers to such questions have been shown to be of importance with respect to understanding conceptual development in children (Siegler, 1991) and adults (Smith & Medin, 1981). Furthermore, concept representation "remains as a cornerstone issue in all aspects of cognitive science" (Medin, 1989, p. 1469).

Of particular importance is the experimental process utilized by both Bruner and Vygotsky. For Bruner et al. (1956), Vygotsky's (1934/1986) early research on conceptual development provided fundamental insight into how a researcher could approach the study of such a topic. More specifically, Bruner et al. (1956, p. 41) mention Vygotsky's "block test" as being one of the ways an experimenter can study concept attainment tasks. Vygotsky's test involved a sorting task whereby small wooden blocks that varied in size, shape and color were organized into groups based on nonsense labels. Essentially, the test was designed to study one's ability to induce conjunctive concepts (AND relationship between attributes). Bruner et al. (1956) expanded on Vygotsky's work and developed an experimental procedure to explore two additional types of conceptual induction tasks: (a) disjunctive concepts (OR relationship between attributes), and (b) relational concepts (a specifiable relationship between attributes). Thus many of Vygotsky's ideas on concept learning are subsumed within Bruner et al.'s (1956) later research.

The research described here will re-examine Bruner et al.'s (1956) experimental techniques on concept attainment by using techniques of connectionist modeling. Connectionist models are based on an analogy of the network of interconnections, between neurons, within the brain. Like the brain, connectionist models involve parallel processing of interconnected units, whereby learning occurs through the strengthening or weakening of these connections (Quinlan, 1991). In essence, the connectionist modeling approach offers cognitive researchers microlevel models of postulated cognitive processes (Simon & Halford, 1996a).

The ideal underlying premise of connectionist modeling assumes that there exists an isomorphic relationship between Bruner et al.'s (1956) theory of concept learning and its connectionist model counterpart. This assumption may not be the case. It may be that such models produce phenomena that are not directly explainable with respect to the theory being modeled. In general there are two possibilities for such results: (a) the model may be deemed a less than adequate representation of the theory; or (b) the theory, on which the model is based, is inadequate to explain the phenomenon in question. In either case, the connectionist model can assist in the interpretation of performance at a level of analysis below the flow-chart and decision tree descriptions of a process (Klahr, 1992).

Essentially, the role of a connectionist model is to make processes (e.g., concept attainment) more explicit for investigation. At a coarse level of analysis this involves comparing the model's performance with empirical data, whereas at a fine level of analysis this involves an interpretation of how the model comes to represent the process being studied. An important aspect of the fine level of analysis is the underlying structural assumptions used by the model to instantiate its representation. These assumptions can be viewed as methods to preserve the model's ecological validity. As a result, in addition to recasting Bruner et al.'s (1956) classic study of concept attainment in a connectionist model, the implications of educational learning theory for connectionist model building will be examined with respect to improving the model's ecological validity. Specifically, it will demonstrate how Gagne's (1962; 1984; 1985) work can be used to build certain structural assumptions of the model.

The chapter is organized as follows: The second section reviews the nature of concepts, how concepts might be learned, and an example of a connectionist model of concept learning. Section three examines Bruner et al.'s (1956) classic study in concept attainment and its relationship to inductive hypothesis generation. Section four describes a connectionist model of Bruner et al.'s (1956) concept attainment task and describes initial empirical results from the model. Section five presents an argument for restructuring the network based on these initial results and Gagne's theory of concept learning. Section six describes the experimental results obtained from the restructured network model. The final section suggests the implications this work has for educational research.

3.2 Nature of concepts

Smith (1988) points outs that "the notion of a concept is essential for understanding thought and behavior" (p. 19). Basically concepts can be viewed as mental representations of classes and one of their most important purposes is that of cognitive economy (Bruner, et al., 1956; Rosch, 1988). By mentally structuring the world into classes or categories we reduce the amount of information that must be processed and thereby conserve our cognitive resources. If we were not capable of forming categories the problem of remembering each individual instance of every object or experience would place our cognitive processing system into a state of computational overload. As a result, conceptual development is directly tied to our ability to categorize, where a concept functions as a categorization device (Smith & Medin, 1981).

The simplest way to view the research on major theories of concepts is chronologically. According to Medin (1989), there have been two important shifts with respect to theories of concepts. The first shift was from the so-called "classical view" to that of "probabilistic theories." The second shift is from probabilistic theories to "theory-based" theories.

The classical view asserts that examples of a category share necessary and sufficent properties and that these properties can be used to determine if a concept belongs in a given category. In their classical form, concepts exist as dictionary definitions. For example, an island is a body of land surrounded by water. An island has two properties: (a) land, and (b) surrounded by water. Each of these properties is necessary because something can't be an island without being both a piece of land and being surrounded by water. The two properties are jointly sufficient, that is, if something is a piece of land and surrounded by water it must be an island. Although both Vygotsky and Bruner tend to present the classical view of conceptual theory (Siegler, 1991), an entire chapter of Bruner et al.'s (1956) book is devoted to the idea that some concepts are represented probabilistically.

Smith and Medin (1981) identified three problems with the classical view. First, it is often difficult to identify a set of defining features. For example, a person may say "made of leather" is a defining property of ice skates but not all ice skates are made of leather. Second, people don't often refer to the dictionary definition to define a concept; instead, they base the notion of a concept on the extent to which the instances are "typical" of a concept. For example, people rate an apple as being more typical of the concept "fruit" than they would an olive. Finally, there are cases where category membership is unclear and a dictionary definition is not suffice, (e.g., is a clock a piece of furniture?).

The weakness in the classical view of concepts gave way to the probabilistic view that states categories are fuzzy entities that are composed from correlated values of attributes (Rosch, 1975; Rosch, 1988). That is, categories represent concepts with respect to the degree that features of a concept share common characteristics. One perspective of the probabilistic view is that mental representations exist in the form of prototypes. A prototype is the most representative instance of a concept, meaning it has the most typical attributes. The basic notion underlying the probabilistic view is that people classify things according to the central tendency of features (e.g., a robin is more birdlike than a penguin). This helps to overcome the classification problem of what might constitute a "typical" concept for a given category like fruit. (Bruner (1956), on page 64-65, also provides an early account of the notion of a typical representation.) Problems with the probabilistic view concern determining what constitutes a feature and the similarity between two features. For example, for the concept "beautiful face": (a) What are the features? (b) How similar (highly correlated) will these agreed upon features be for a group of beautiful faces?

In the theory-based view, categories represent causal relationships between states of an organized system. The theory-based view advocates analyzing the relational properties, rather than the features, when forming categories. For example, a concept that has the following properties: children, pets, photographs, and rare books, might make sense only when one creates a category called "things to get when the house is on fire" (Barsalou, 1983). Interestingly, Kiel (1987) points out that both Vygotsky and Bruner also noticed the shifts in children from the probabilistic view to the theory-based view.

In summary, the three major theories regarding the nature of concepts offer insight into how concepts might be understood. It appears Bruner et al.'s (1956) early work on the nature of concepts addressed all three of these theorical issues. This indicates the significant insight these early researchers had and the importance of their experimental techniques and analysis.

3.2.1 Concept learning from an instructional perspective

Concept learning refers to forming representations to identify attributes, generalize them to new examples, and discriminate examples from nonexamples (Schunk, 1996, p. 218). Gagne (1985; 1979) developed a view of learning and instruction that makes concept learning one of the central issues. According to Dick and Carey (1985), Gagne asks the question "What does the student already have to know how to do so that, with a minimal amount of instruction, this task can be learned?" (p. 48). In answering this question Gagne (1962; 1985) developed the notion of "instructional learning hierarchies." A learning hierarchy is composed of instructional objectives and a set of prerequisite relationships connecting the objectives. Support for learning hierarchies can be found in a meta-analysis study by Horon and Lynn (1980) who reviewed 15 studies that compared the effectiveness of learning hierarchies in the sequencing of objectives. They concluded such hierarchies increase achievement and reduce learning time.

With respect to concept learning, the question becomes what prerequisite capabilities must learners have to discriminate among stimulus attributes to determine their relevancy in the context of learning new concepts. Gagne (1985) identifies the domain of intellectual skills as having an important effect on the learning outcomes related to concepts, rules, and problem solving. At the most complex level intellectual skills are synonymous with procedural knowledge as described by Anderson (1993). The following set of seven prerequisite categories are said to comprise the domain of intellectual skills, ranging from least complex to most complex.

1. stimulus-response connection: A single connection is formed between a stimulus and a response.

2. chaining: This involves linking individual stimulus-response elements into a sequence.

3. making verbal associations: Connecting verbal stimulus-responses in a sequence.

4. making discriminations: "A capability of making different responses to stimuli that differ from each other along one or more physical dimension" (Gagne & Briggs, 1979, p. 63).

5. learning concepts: "A capability that makes it possible for an individual to identify a stimulus as a member of a class having some characteristics in common, even

though such stimuli may otherwise differ from each other markedly" (Gagne & Briggs, 1979, p. 64).

6. learning rules: The learner is able to respond with regularity to different situations with a class of relationships among classes of objects and events.

7. solving problems: The ability to combine simple rules to form complex ones and then use these rules to solve problems.

Note that for Gagne (1985) a key step prior to learning concepts is learning to make discriminations between different members of a particular collection. Thus two different kinds of learning may occur when the learner is confronted with a set of stimulus objects. As Gagne states:

> In some circumstances of discrimination learning, the learner may have to acquire a response that differentiates (by name or otherwise) the stimulus features of a single member of a set from those of other members of a set, making a different response to each. Having previously learned to make these distinctions, the learner may be required to acquire the capability of responding to the set of stimuli as a class and distinguishing members of the class from non-members; this is what is meant by "learning a concept." Of course, the learner may be required to respond to a stimulus set in both these ways on different occasions, and such a response is perfectly possible (p. 90).

Discrimination learning often involves having the learner identify distinctive features of stimulus objects. Individuals learn differential responses to distinguishing characteristics of objects like shapes, sizes, colors, and textures (Gagne, 1985, p. 91). The following three stage model, presented by Gagne (1985, p. 97), describes the nature of discrimination learning and its relationship to concept learning that follows from it.

1. In the first stage a stimulus feature is presented as an instance of the concept along with a noninstance. The learner demonstrates the ability to make the discrimination between two stimuli such as curved line vs. a straight line.

2. During the second stage, referred to as the generalization stage, the learner identifies instances from noninstances. For example, a straight line paired with different curved lines.

3. Finally, in stage three, the stimulus features such as a curved and straight line are varied across dimensions and presented as instances and non-instances of a class. For

example, if straightness is the concept to be attained non-instances would be nonstraight lines.

Throughout the learning process, reinforcement is used for responding correctly. To test that the concept has been attained the learner is asked to identify whether novel instances are part of the concept category.

Gagne's (1985) ideas on concept learning are analogous to those of Klausmeier (1992) who postulates a four stage approach for concept learning: (a) concrete, (b) identity, (c) classificatory, and (d) formal. Each stage in the model is a prerequisite for the later stage and is based on the interaction between development, experience, and instruction. At the concrete level the context and spatial orientation remains the same such that two items shown at different times can be recognized as the same. This requires the learner to discriminate the item based on defining attributes. For example, two rectangles shown at different times would still be recognized as rectangles.

When two items can be recognized as being the same, even if they are observed from different perspectives, the learner is said to be at the identity level. This stage involves the process of generalization. For example, a rectangle presented at different orientations is still recognized as a rectangle.

Formation at the classificatory level means that two items can be recognized as being equivalent. For example, a large rectangle is conceptually the same as a small rectangle.

Finally, the formal level requires the learner to recognize concept instances from noninstances, name the concept and defining attributes, and indicate why certain attributes separate the concept from closely related concepts. At this stage, cognitive processes such as evaluating, hypothesizing, and inductive reasoning are important.

In either case, for both Gagne (1985) and Klausmeier (1992), concept learning is an outgrowth of a sequential process whereby subordinate objectives (feature discrimination) must be mastered prior to superordinate objectives (attainment of concept). More specifically recognizing a concept instance from non-instance involves discrimination (identifcation) of attributes. As will be shown, such an approach can have important implications for designers of connectionist concept attainment networks. A fundamental question for network designers is what prerequisite skills (critical subordinate skills) must the network either have or learn in order to solve a concept attainment problem?

3.2.2 Connectionist approach to modeling concepts

Traditionally, concept learning has been an important area of research in cognitive science (Bower & Clapper, 1989). If concepts are "essentially pattern-recognition devices"

(Smith & Medin, 1981, p. 8), then connectionist modeling is ideally suited for the study of concepts (Clark, 1993). Connectionist modeling of concept learning has been done with both supervised (Gluck & Bower, 1988; Plunkett & Sinha, 1992) and unsupervised (Carpenter & Grossberg, 1988; Schyns, 1991) networks. What follows is a more detailed example of one connectionist model used to study concept learning.[13]

3.2.2.1 Plunkett and Sinha

Plunkett and Sinha (1992) built a connectionist model of concept formation and vocabulary growth that attempts to provide insight into further understanding semantic development in early childhood. The model is build on the premise that vocabulary growth depends on the formation of non-linguistic categories but the "structure of the vocabulary of the language being acquired plays the role of highlighting which categories should be semantically represented" (Plunkett & Sinha, 1992, p. 229). Such ideas have long been associated with the work of Vyogtsky (1934/1986).

The model built by Plunkett and Sinha (1992) was derived from earlier network modeling work done by Chauvin (1988) and by experimental procedures and results described by Posner and Keele (1968; 1970). The goal of the network is to associate labels (nonsense labels) and images (dot patterns). Thirty-two randomly generated dot pattern images were formed and designated as prototypes. Each prototype contained nine dots. No prototype pattern had more than two dots in common with any other prototype. From each prototype six distorted images were created by randomly moving each of the nine dots by a specified distance. Some dot patterns were more distorted than others. A total of 192 image patterns were created from the original 32 prototypes, providing 32 categories with six entries in each. The original dot images were preprocessed by a mathematical function that mapped a 2,100 pixel image into 171 retinal units (Plunkett & Sinha, 1992, p. 230). In doing this transformational mapping the 171 retinal units were encoded into a dot neighboring relationship not present in the original image. The underlying goal was to preserve the two dimensional bit image in a one dimensional input vector. In other words, the input image data is represented in a compressed format to reduce network input.

The labels for the prototypes were constructed as 32 element bit vectors (one for each category) with only a single bit set on. The relationship between the prototype

[13] The topic of connectionist networks will not be reviewed here. For an introduction to connectionist modeling in education see Schneider and Graham (1992); for an introduction to connectionism or neural networks in general, see Rumelhart and McClelland (1986), Hecht-Nielsen (1990), or Gallant (1993); for

categories and label was completely arbitrary. All label vectors were orthogonal to ensure a linear mathematical relationship of zero.

The four layer network architecture used by Plunkett and Sinha (1992) is shown in Figure 22. The network consists of two separate sets of input units, one called retinal units (for dot images), and the second called label units (for label vectors). The network task is simple yet non-trivial. The network must reproduce the input vectors at the output units. The non-trivial task involves the intermediary representations that must be instantiated inside the network for an association to be formed between input and output. The first hidden layer receives input from two separate channels. On one channel, 32 hidden units receive input from the 171 retinal units. On another channel, 32 hidden units receive input from the label input units. Fifty hidden units at the second hidden layer combine the input from both channels coming from the first hidden layer. Finally, the output layer is exactly the same as the input layer. It is composed of two separate output channels, one of retinal units (171) and the other of label units (32).

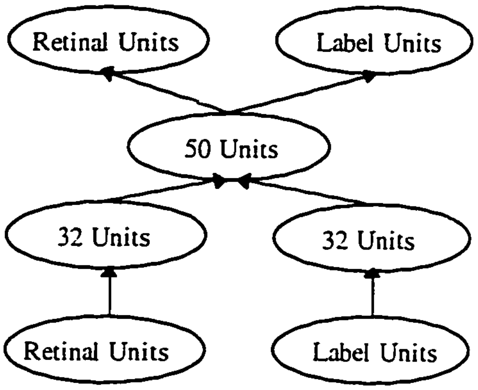


Figure 22: The four layer network architecture used by Plunkett and Sinha (1992).

Training the network involves three phases. In the first phase the retinal input vector is presented to the network. The output from the network's retinal output units is compared to the input representation and an error measure is calculated (only for the retinal output units). Using a back propagation algorithm, the internal network weights are updated (only for the retinal channel). In the second phase the exact same procedure is repeated on the label channel (i.e., input the label vector, calculate error, update weights on

information on the relationship between connectionism to psychology see Quinlan (1991), Bechtel and label channel). The final phase involves presenting both input vectors (label and image) simultaneously and updating all internal network weights. This sequence is repeated for all image/label pairs. Each image/label pair is randomly chosen. The network is never trained on the prototype image, only the distorted images. The network learning process can be viewed as a cyclic process, going from image learning then to label learning, and finally to the association (learning) between image and label.

Results from the network's performance appear to demonstrate three properties of human concept formation and vocabulary growth: (a) prototype extraction, (b) vocabulary growth spurt , and (c) comprehension/production asymmetry. In the case of prototype extraction, the network performance compared favorably with results from earlier human studies. In these studies subjects were able to categorize prototypes that were never seen during category formation (Posner & Keele, 1968; Posner & Keele, 1970). Similarly, the network was able to correctly categorize the prototype images (activate the appropriate label output units) when these images were presented to the network after training. What the network is actually doing is extracting the "central tendency" for each of the category clusters of distorted images.

A second interesting phenomenon was the network's tendency to undergo vocabulary growth spurts. Results from network output indicate that at a fairly early time during training (between 20-30 epochs)[14] the network's ability to activate the correct label (for a given image input) unexpectedly increases. This parallels research findings that show children typically undergo a dramatic increase in vocabulary growth at a point during their second year (Bates, Bretherton, & Snyder, 1988).

A third phenomenon produced by the network is comprehension/production asymmetry. Comprehension is defined as the network's ability to activate the appropriate retinal image when presented with label input. Production is defined as the network's ability to activate the appropriate label when presented with retinal image at input. According to Plunkett and Sinha (1992) this type of asymmetry between comprehension and production has been documented in studies of children. The simulations produced a higher comprehension rate than production rate during network learning. In other words, mapping the label to the image is much easier for the network to learn than mapping image to label. In network terms, this has do with the fact that label vectors are orthogonal whereas the retinal vectors for a given prototype are correlated (thus harder to distinguish for category membership).

Abrahamsen (1991), or Simon and Halford (1996b).

An interesting result reported from the simulation is that the network took a longer period of time to discover categories when label input is not present during training. This indicates that the categories can developed without labels but that the use of labels speeds up category learning. As Plunket and Shina (1992, p. 237) point out, this supports Vygotsky's (van der Veer & Valsiner, 1991) interactional account of semantic development between concept formation and language growth. The authors indicate that other simulation results suggest the creation of idiosyncratic categorizations as postulated by Vygotsky but that little is presently known about such processes and more experimental investigations with children in naturalistic settings is required.

3.3 Bruner's classic study of concept attainment

In 1956, Bruner, Goodnow, and Austin published a book entitled A Study of Thinking, the goal of which was to postulate ideas regarding the human processes of categorization and conceptualizing. According to Bruner et al., learning to categorize objects is directly tied to the act of concept attainment. "Attainment refers to the process of finding predictive defining attributes that distinguish exemplars from nonexemplars of the class one seeks to discriminate" (Bruner et al., 1956 , p. 22). Thus concept attainment involves the act of forming a hypothesis with respect to how an object should be categorized. "To categorize is to render discriminably different things equivalent, to group the objects and events and people around us into classes, and to respond to them in terms of their class membership rather than their uniqueness" (Bruner et al., 1956 , p. 1). The identity of a concept and its classification into a distinct category depends on the way different attributes of a concept are defined and combined. Bruner et al. define an attribute as "any discriminable feature of an event that is susceptible of some discriminable variation from event to event" (p.26). Bruner et al.'s definition of a concept "is the network of inferences that are or may be set into play by an act of categorization" (p. 244) and closely resembles a number of the recent theory-based aspects of conceptual theory (see Medin 1989, p. 1475). In this sense, a concept is viewed as a way of reasoning about an object or event with respect to its category membership. In other words, a concept can be viewed as a reasoning chain of hypothesis formation and hypothesis evaluation, very similar to Newell and Simon's (1972) state space search, Gagne's (1984) learning outcomes category of intellectual skills, and Anderson's (1993) view of problem solving.

[14] One epoch is the presentation of all the exemplars in the training set to the network.

As noted earlier, the basis for Bruner et al.'s (1956) view of concepts was grounded in their experimental studies that built upon the earlier work of researchers like Vygotsky. Bruner et al. (1956, p. 42) used a set of 81 cards (much like playing cards) to study concept attainment. Their cards consisted of four attributes, each of which varied on three values. Figure 23 illustrates the card set that was used by Bruner et al. as the generic stimuli. The four attributes and their associated values are (a) object shape (cross, circle, square); (b) color (green, black, red); (c) number of objects (one, two, or three); and (d) number of borders (one, two, or three). Thus each card instance combined four attributes and varies in accordance with the values. A category or a concept is defined with respect to a subset of cards, where each card shares a specific set of attribute values. Participants were told to discover the concept (induce a hypotheses) that would best describe a subset of instances. A simple task might involve giving individuals both positive and negative card instances to see if they were able to infer a hypotheses that defined a conceptual relationship between the instances. In other words, the individual would postulate a possible concept (hypothesis) based on the finite amount of information that was extracted from a subset of cards taken from the domain of 81 cards. Because the stimuli used were generic (cards) the concepts were artificial, meaning they did not exist beyond the population of 81 cards.

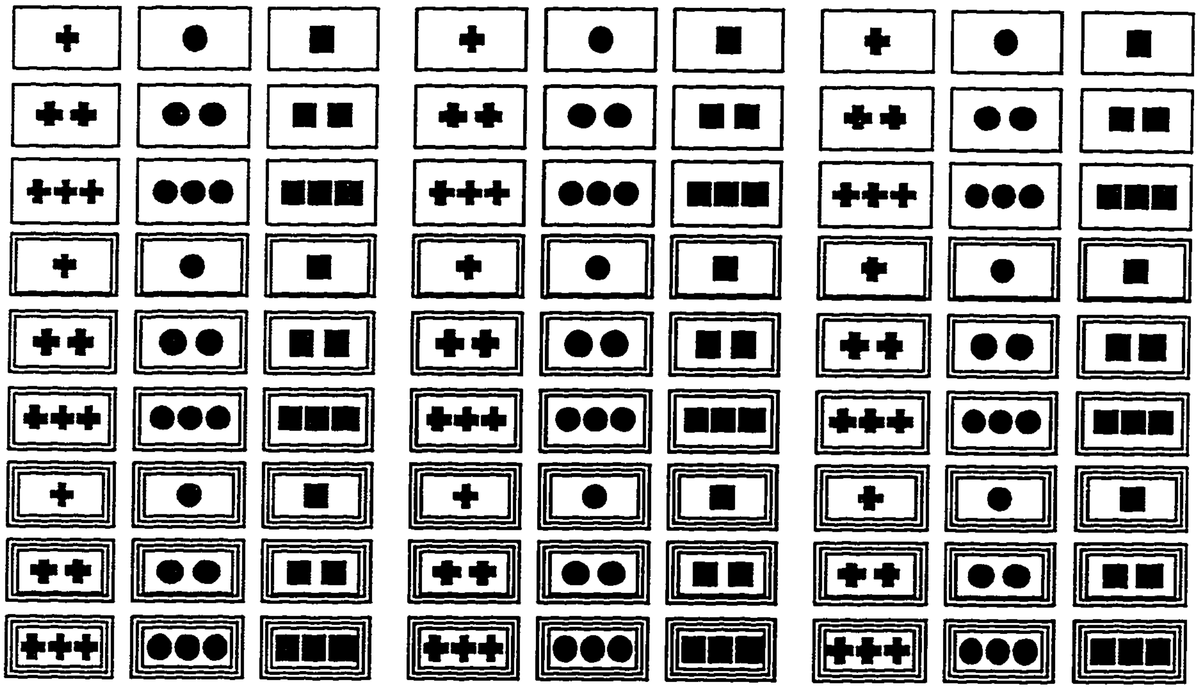


Figure 23: The set of 81 cards used by Bruner et al. (1956) to study concept attainment.

Each instance in the array consists of a combination of attributes and their respective values. Columns 1-3 are green, columns 4-6 are black, and columns 7-9 are red.

How attributes are combined to generate inferential hypotheses is directly linked to the conceptual category types available. Bruner et al. (1956) describe three basic category types:

1. conjunctive concepts: The joint presence of attributes values (e.g, cards that contain one object and that object is square).

2. disjunctive concepts: Either the joint presence of attribute values or the presence of either one of the attribute values independently (e.g., three red circles or three borders of any color).

3. relational concepts: This is defined by the specific relationship between attributes (e.g., cards with fewer borders than objects).

3.3.1 Anderson's example

Figure 24 is adopted from Anderson (1990, p. 307) and contains a hypothetical sequence of instances that could be presented to individuals. The three columns represent the three category types, conjunctive in column one, disjunctive in column two, and relational in column three. The positive and negative instances of a concept would be presented to participants one at a time. For example, reading columns one from top to bottom, the first card (concept instance) presented to an individual is a positive instance of the concept. That card contains crosses (X), is colored green (G), has two objects (2) and one border (1) producing the acronym XG21.[15] The second card (SR21) represents a negative instance, the third card (XG12) is also a negative instance and so on. With only two positive instances and three negative instances the individual should have enough information to formulate the hypothesis for the conjunctive concept, two objects and crosses. The second column shows the disjunctive concept, two borders or circles. Finally the third column shows the relational concept, number of objects must equal number of borders.

Bruner et al.'s (1956) experimental procedure involved having participants discover primarily conjunctive and disjunctive concepts. Very little work was done on relational

[15] The following acronym coding scheme for the cards is used: For attribute shape: Circle (C), Square (S), Cross (X); for attribute color: Green (G), Black (B), Red (R); for attribute number of objects: One (1), Two (2), Three (3); for attribute number of borders: One (1), Two (2), Three (3). The order of the acronym is shape, color, number of objects, and number of borders.

concepts. To make the task easier the experimenter would often tell the individual the type of concept and the number of attribute values.

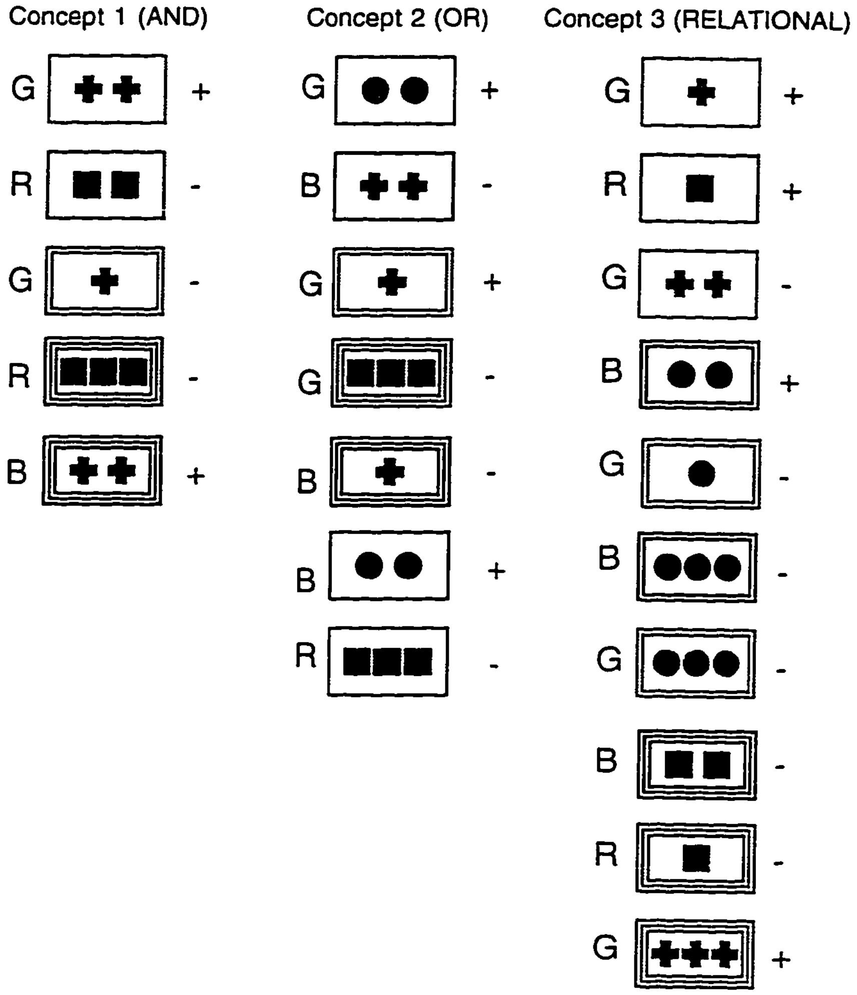


Figure 24: Three examples of concepts categories suggested by Anderson (1990).
A plus (+) indicated the card is a positive instance of the concept, a minus (-) indicates a negative instance.

3.4 Modeling Bruner's concept attainment task

Modeling Bruner et. al.'s (1956) concept attainment task was approached in the following way:

1. The three subsets of cards in Figure 24 were identified as the input exemplars for the network. These examples were selected for initial investigation because: (a) they are

general enough to convey the negative/positive instance relationship of the intended concepts; and (b) there is a specific example for each of the category types (AND, OR, and RELATIONAL).

2. The "real world" visual representation of the cards was then mapped into a mathematical representation that was suitable for network input. Considering the type of task and the procedure used by Bruner et al. (1956), it would important that such a mapping not interfere with the way the cards were to be presented to the network model.

3. A network architecture was then proposed that would accept the exemplar card sets as input and learn the concept attainment task.

4. The network was then trained on an exemplar subset to an acceptable level of conceptual representation.

5. The trained network's behavior was then evaluated on the degree of concept attainment achieved.

3.4.1 Network Input

The approach undertaken was to map an input representation scheme for the 81 cards (visual stimuli) that attempts to preserve as much of the cards' real world representation as possible. The intention was to have an input representation which closely matched that of the cards observed by an individual (as described by Bruner et al. 1956) and thus to increase the level of ecological validity during the experimental process (Bracht & Glass, 1968). In other words, the independent variables (the properties of the cards) were explicitly described for the network input to minimize the differences between their real world form and their network input form.

To accurately reproduce the cards illustrated on page 42 of Bruner et al.'s (1956) book, each pixel was mapped one-to-one to cells in a 2-dimensional numeric array. The size of the array was 26 vertical pixels by 44 horizontal pixels for a total of 1144 pixels. The corresponding numeric array thus had 1144 cells, where colored pixels green, red and black were assigned the values of 0.25, 0.5, and 0.75 respectively. All other cells contained values of 0.0. As an example, the card XR22 of Figure 25 shows a one-to-one mapping between the picture representation and its corresponding numeric array. The numeric array that designates this card is made up of 0.5 in all the cells that contain red pixels. The outside border, row 1 columns 1 - 44, row 26 columns 1-44, column 1 rows 1-26, and column 44 row 1-26, all had 0.5 values. The inside border, row 3 columns 3-41, row 23 columns 3-41, column 3 rows 3-23, column 41 rows 3-23 also contain values of 0.5. The cross shapes were mapped to their respective X,Y positions inside the borders and also contain the 0.5 values. This same mapping procedure was used on all 81 cards.

5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5
5	0	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	0	5
5	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	5
5	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	5
5	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	5
5	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	5
5	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	5
5	0	5	0	0	0	0	0	0	0	0	0	0	5	5	5	5	0	0	0	0	0	0	0	0	0	0	5	5	5	5	0	0	0	0	0	0	0	0	0	0	5	0	5
5	0	5	0	0	0	0	0	0	0	0	0	0	5	5	5	5	0	0	0	0	0	0	0	0	0	0	5	5	5	5	0	0	0	0	0	0	0	0	0	0	5	0	5
5	0	5	0	0	0	0	0	0	0	0	0	0	5	5	5	5	0	0	0	0	0	0	0	0	0	0	5	5	5	5	0	0	0	0	0	0	0	0	0	0	5	0	5
5	0	5	0	0	0	0	0	0	0	5	5	5	5	5	5	5	5	5	5	0	0	0	0	5	5	5	5	5	5	5	5	5	5	0	0	0	0	0	0	0	5	0	5
5	0	5	0	0	0	0	0	0	0	5	5	5	5	5	5	5	5	5	5	0	0	0	0	5	5	5	5	5	5	5	5	5	5	0	0	0	0	0	0	0	5	0	5
5	0	5	0	0	0	0	0	0	0	5	5	5	5	5	5	5	5	5	5	0	0	0	0	5	5	5	5	5	5	5	5	5	5	0	0	0	0	0	0	0	5	0	5
5	0	5	0	0	0	0	0	0	0	5	5	5	5	5	5	5	5	5	5	0	0	0	0	5	5	5	5	5	5	5	5	5	5	0	0	0	0	0	0	0	5	0	5
5	0	5	0	0	0	0	0	0	0	0	0	0	5	5	5	5	0	0	0	0	0	0	0	0	0	0	5	5	5	5	0	0	0	0	0	0	0	0	0	0	5	0	5
5	0	5	0	0	0	0	0	0	0	0	0	0	5	5	5	5	0	0	0	0	0	0	0	0	0	0	5	5	5	5	0	0	0	0	0	0	0	0	0	0	5	0	5
5	0	5	0	0	0	0	0	0	0	0	0	0	5	5	5	5	0	0	0	0	0	0	0	0	0	0	5	5	5	5	0	0	0	0	0	0	0	0	0	0	5	0	5
5	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	5
5	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	5
5	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	5
5	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	5
5	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	5
5	0	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	0	5
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5
5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5

Figure 25. A numeric array representing the card XR22. The value of 5 (actually 0.5) represents a red pixel of the card and a value of 0 represents white pixels.

3.4.2 Basic network architecture

A three layer network architecture was constructed where each unit in a given layer was connected to all units in the layer above (multilayer feed forward Parallel Distributed Processing design). Figure 26 shows the initial physical structure of this network designed to learn the conjunctive concept task described in Figure 24. Note the network contains one output unit, five hidden units, and 1144 input units. The input values to each unit in the network, with the exception of the input units, and the activation values for the hidden and output units were computed using the following functions respectively:

$$unit_i = \sum_j w_{ij} a_j + bias_i \qquad \text{and} \qquad a_i = \frac{1}{1+e^{-unit_i}}$$

In these equations, *j* ranges over the number of units sending input to *unit*$_i$. The value of w_{ij} is the connection weight value to unit *i* from the unit *j* and a_j is the activation level at unit *j*. The value of the bias is represented by *bias*$_i$. A standard back propagation of error correction algorithm was used to train the network connection weights (Rumelhart, Hinton, & Williams, 1986). Connection weights were assigned randomly between +1.0 and -1.0. Momentum was set to 0.0 and the bias value was turned off. The learning rate was set to

1.0. The network error output for each exemplar was monitored after a set number of epochs (e.g., 50) until an acceptable convergence value was rea:hed. Three networks were constructed, one for each exemplar set in Figure 24 (AND, OR, and RELATIONAL). The difference between the networks were the number of hidden units and the number of training epochs required to reach convergence. The input exemplars were selected sequentially in accordance with their position in Figure 24. Each input exemplar consisted of 1144 inputs (see section 3.4.1). The complete simulation including data plots was coded using the Mathematica® programming language (Wolfram, 1991).[16]

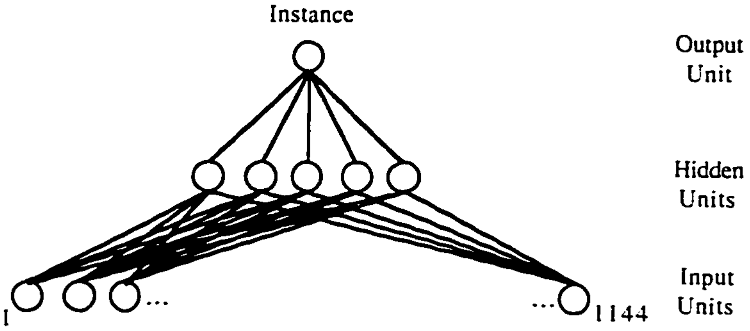


Figure 26. A three layer network structure used to study Bruner et al.'s (1956) conjunctive concept attainment task. Note the output layer contains only one output unit.

3.4.3 Initial results from network with one output unit

Empirical results demonstrated that early versions of the network, using one output unit (see Figure 26) to classify cards as either a positive or negative instances, provided a poor model of concept attainment. Figure 27 shows a graph of the root mean square (RMS) error for the network learning the conjunctive concept (two and cross) in Figure 24.

[16] Mathematica® is an excellent tool to both construct connectionist models and to study their functional properties through the use of data visualization (Freeman, 1994).

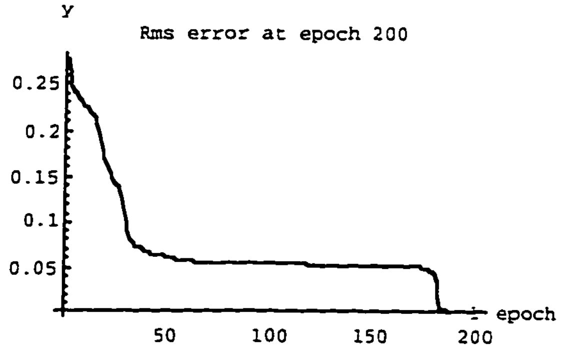


Figure 27. RMS error for the one output unit.

The x-axis indicates the number of training epochs. The y-axis indicates the RMS error at each epoch. Note the error rate falls quickly and levels off at point 0.05 until 180 epochs where it falls quickly toward convergence. The minimal acceptable RMS error is often selected 0.01, for the total responses given at the output layer, after one epoch (McClelland & Rumelhart, 1989). When this network was presented with a novel card representing the concept (a card not in the training set), it would sometimes fail to classify it as a positive instance. For example, the histograms in Figure 28 show the output unit response for this network when presented the nine cards, from the 81 card population, that represent the conjunctive concept two and cross. The x-axis on each histogram indicates the output unit number (in this case there is only one). The y-axis indicates the output unit response. The target response output of such a network is 1.0 for a TRUE instance or 0.0 for a FALSE non-instance. Note that cards XG21 and XB23 are part of the training set and for this reason produce a value close to 1.0. On the other hand, of the seven remaining novel cards, three (XG22, XR22, XB22) were classified as being a negative instance of the concept. These results indicated the network was not being trained on a target output that reflected the implicit properties of the concept attainment task being modeled. Therefore, it appears that association among the input stimuli (the numerical array that uniquely defines each card) and instance values, were not being established with generality. If their association had been adequately established, exemplars XG22, XR22, and XB22 would have had response outputs approaching 1.0. In this sense, the hidden layer unit weights must converge so that the functional relationship between the input and output units is effective over cards beyond the training set.

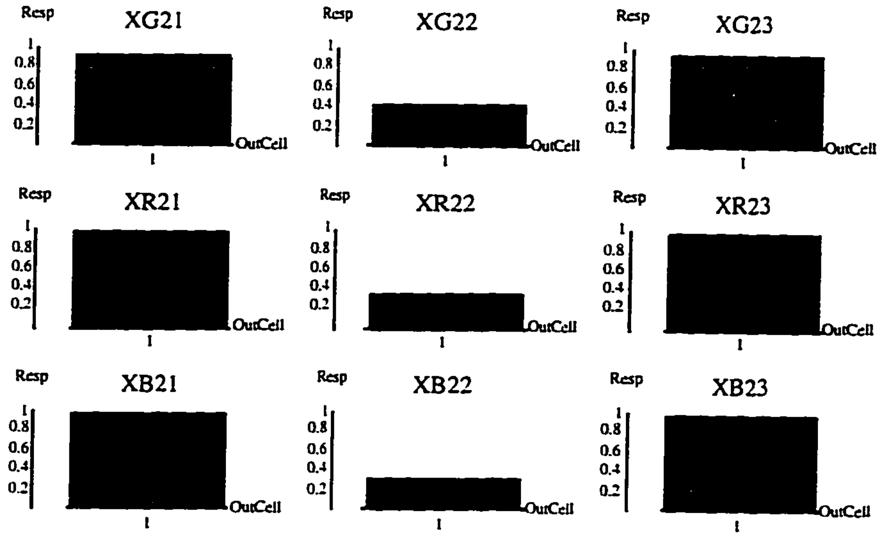


Figure 28. The output unit response for each of the nine cards that contain the concept two and cross.

3.5 Expanding network output for attribute context constraints

Appealing to Gagne's (1962; 1984; 1985) theory on concept learning discussed earlier, subordinate objectives should be mastered prior to higher level objectives. Specifically, the learning of concepts requires that a learner be capable of making discriminations. Within the Bruner et al.'s (1956) card set these discriminations are viewed as the attributes shape, color, number of borders, and number of objects. To successfully model the concept attainment task either: (a) apriori attribute discrimination knowledge must be present in the model, or (b) the capability of acquiring discrimination knowledge simultaneously during concept learning must be present in the model. The goal of the network is to instantiate a model that can attain the inherent concept category (conjunctive, disjunctive, or relational) defined by a set of input instances like those shown in Figure 24. When such a network model is presented with a novel instance of the concept it can "inductively reason" (in the generalization sense) the instance's truth value. In sum, Gagne's ideas suggest that concept attainment could not be achieved until a sufficient level of attribute discrimination knowledge is present in the model.

The intention is to have additional output units satisfy Gagne's (1985) criteria by providing evidence of attribute discrimination during network learning. The idea of employing additional output units to improve network learning has previously been used in engineering to increase the probability of the network finding a good representation and to reduce network learning time; the process is sometimes referred to as "injection of hints" or "extra output learning" (Suddarth, Sutton, & Holden, 1988; Yu & Simmons, 1990).[17] Essentially, the idea is to provide additional output units that express some knowledge of the problem. No modification of the network algorithm or error correction is required. For example, Gallmo and Carlstrom (1995) found training is faster and generalization is better (lower probability of getting trapped in local minima) for a multilayer network solving the XOR problem using AND as an extra target. Thus, while the network is learning to classify a card as either a positive or negative concept instance, it must simultaneously learn to make the necessary discriminations between the attribute values associated with the card instances. In a psychological context, extra output units can be viewed as attribute context constraints[18] that must also be learned in association with either a positive or negative concept instance.

Given that a network is a function learning device (Hecht-Nielsen, 1990) and that the input data, which describes the function, is seldom complete, the network's output is most often an approximation of the true mathematical function. Following the rationale used by Gallmo and Carlstrom (1995) to support the inclusion of extra output units, let **F** denote the function that describes the concept being attained by the network. The set of 81 cards describes the population from which all possible concepts are derived. In Bruner et al.'s (1956) task a single concept is defined as a subset of positive and negative card instances. The complete function definition of the concept **F** would require all 81 cards to be identified as positive and negative instances with respect to the concept.

Given that the network is only presented with a subset of instances the function produced by the network is an approximation of **F**, let {f} be the set of possible functions that are approximations for **F**. In other words, for a given subset of card instances defined as input training set **T**, the network will converge on one of the possible functions defined in {f}. Convergence is achieved when the network error value reaches an acceptable level. This indicates the concept, as defined in **T**, has been learned. The possible size of {f} may

[17] It should be noted that increasing the number of output units does not increase the representational power of the network. The representational power of the network resides in the number of hidden units. Any multilayer network trained using the back propagation algorithm can behave as a universal function approximator, meaning it can represent any function to any degree of accuracy if given a sufficient number of hidden units (Hornik, Stinchcombe, & White, 1989). Therefore, given a finite training set, the extra outputs may assist in finding a good representation from the set of possible representations.

[18] Analogously one can view these constraints as being similar to the constraint that eigenvectors of a matrix be found conditional upon being normalized, the difference being that these attribute constraints have a psychological basis and are not strictly formalized to an absolute value, such as 1.0 of a normalized vector, nor is the solution non-unique.

be large and some elements of {f} appear to be good function approximations (i.e., representing global minima in the error space). This may not be true because **T** is defined as a subset of the card population and thus the chosen element from {f} may only reflect a global minima with respect to this limited subset. The ability to generalize across the entire card population may not be resident in the chosen element from {f}. That is the network's conceptual representation may be false (i.e., the network has a misconception).

For example, assume the five cards for the conjunctive concept in Figure 24 is used to defined **T**. The concept to be attained is conjunctive—number of objects is two and these objects must be crosses. A network trained on **T** will produce TRUE for a positive card instance in **T** and FALSE for a negative card instance in **T**. To test network concept attainment a novel card instance is presented to the network (a card from the population not in **T**). Failure to correctly classify this card as either TRUE (instance) or FALSE (non-instance) would indicate the network's lack of concept attainment (generalization). Such was the case for early versions of the network that compared the output from one unit against positive and negative targets.

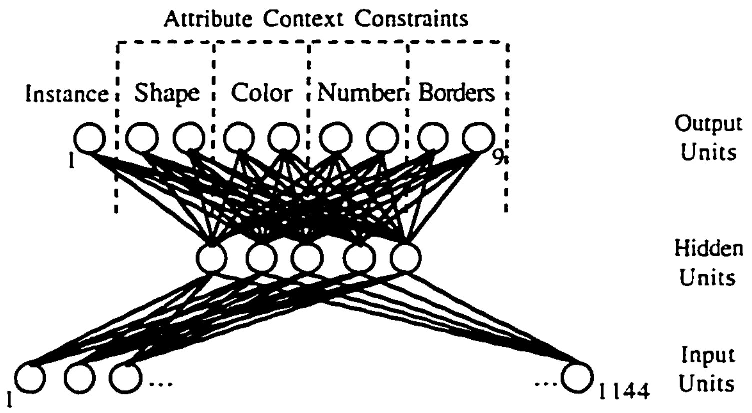


Figure 29. A three layer constrained network structure used to study Bruner et al.'s (1956) conjunctive concept attainment task. Note the addition of the extra output units to classify the card according to the attributes associated with the output instance unit.

To achieve a better network representation more knowledge of the input must be reflected in the output. Let **T′**, **T″**, **T‴**, and **T⁗** be duplicate sets of **T**, where **T′** is defined as the training set for the color function, **C** (red, green, black). **T″** is defined as the training set for the shape function, **S** (cross, square, circle). **T‴** is defined as the training set for the number of boarders function, **B** (three). **T''''** is defined as the training set for number of objects function, **O** (three). Analogous to the relationship between **F** and the set {f}, let {c}, {s}, {b}, and {o} denote the sets of function approximations the network can implement to satisfy **C**, **S**, **B**, and **O** respectively.

Given that all these functions **F**, **C**, **S**, **B**, and **O** are defined by the same training set **T** the network structure can be modified to accommodate the simultaneous learning of these functions by extending the output layer. For the output layer in Figure 29 the first unit provides evidence for the instance value, **F**, and the remaining eight provide evidence for attributes (**C**, **S**, **B**, **O**). In this sense, the attribute output units have a natural association with the instance output unit. Note the internal hidden unit structure is not changed but the network is forced to learn all these functions simultaneously (see Figure 30) such that the intersection (i.e., $\{f\} \cap \{s\} \cap \{c\} \cap \{b\} \cap \{o\}$) of all functions define the set of approximations that best describe the concept being attained. These extra output functions can be viewed as additional attribute constraints that must be learned in the context of positive and negative card instances in order to achieve concept attainment. The intersection reflects a subfunction that satisfies these attribute constraints. The network is directed to converge on the functional intersection and to instantiate a better model of concept attainment. This means it should be more capable of generalizing to novel instances. As a result, context dictates the type of attributes that can be used to constrain the space of representations so that the network can find a good representation. Note the number of hidden units in Figure 29 remains the same as in Figure 26, only the output layer has been restructured.

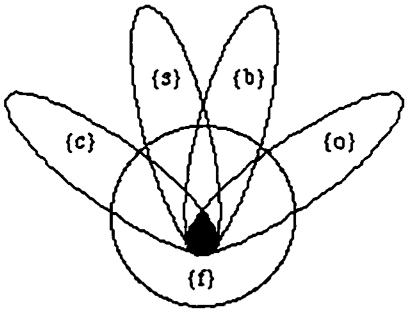


Figure 30. The relationship between attribute subfunctions and the instance function. The intersection defines a set of more robust concept attainment models.

The specific target output representation, a binary string, consists of nine output units that uniquely symbolize the positive or negative card instance and its attributes. A target output representation for a given card is constructed in a two-step process. First, one binary value is used to designate the truth or falsity of a concept instance. In other words, a TRUE concept instance is the binary number 1, and a FALSE concept instance is the binary number 0. Second, a string of binary vectors, one for each of the card's attribute values, were concatenated together to comprise the remaining eight output units. Table 5 shows the four attributes and their associated values that can be combined to form the attribute context constraints. Below each value in the table is the binary code which symbolizes its respective value. Using this coding scheme, the target output representation for card XR22 in Figure 25, (assuming it represents a positive instance) is as follows: The first target element (reading left to right) for output unit one is the binary value 1. The next eight values, required to complete the target, are constructed by concatenating the 2 bits for each of shape (01), color (10), number (10), and borders (10), to produce the following unique vector 101101010.

Table 5. Coding scheme for attribute context constraints

Attributes	Values		
Shape	Cross 01	Circle 10	Square 11
Color	Green 01	Red 10	Black 11
Number	One 01	Two 10	Three 11
Borders	One 01	Two 10	Three 11

3.5.1 Network training with attribute context constraints

The network was trained in the same manner as descibed in section 3.4.2. That is, the five cards in Figure 24 defining the conjunctive concept were used to train the attribute context constrained network to an acceptable error convergence rate (less than 0.01% RMS). In this case the target output had nine output values to match at the output layer. For example, the network was presented with a card instance XG21 (the first card) and the output produced by the network was compared against its associated target bit string (101011001). Table 6 shows the target output for the five cards that were presented to the network. As mentioned previously, the back propagation algorithm was used to adjust the weights of the network during training. The same random number seed was used to initialize the network weights.

Once the network was trained it was tested using card instances outside the training set. In other words, the network was tested with cards not in the training subset but from within the Bruner et al. (1956) card universe. These selected cards included both positive instances and negative instances from the training set.

Table 6. Description of the cards and their associated target output for training set.

	Target Output Units 1-9								
	1	2	3	4	5	6	7	8	9
Card description	Instance	Shape		Color		Number		Borders	
Green with two crosses and one border. (XG21) Positive instance +	1	0	1	0	1	1	0	0	1
Red card with two squares and one border. (SR21) Negative instance -	0	1	1	1	0	1	0	0	1
Green card with one cross and two borders. (XG12) Negative instance -	0	0	1	0	1	0	1	1	0
Red card with three squares and three borders. (SR33) Negative instance -	0	1	1	1	0	1	1	1	1
Black card with two crosses and three borders. (XB23) Positive instance +	1	0	1	1	1	1	0	1	1

3.5.1.1 Speed of learning and categorization of concept instances

This section examines the speed at which the attribute constrained network learned, and the degree to which the instances and attributes in the training subset were also learned. Figure 31 shows a set of RMS error graphs for the nine output units at various epochs (10, 30, and 140 respectively). Each RMS error line in the graphs are labeled in accordance with their respective position in the output target, i.e., output unit one (instance unit) is labeled 1, output units two and three (shape units) are labeled 2 and 3, and so forth. When the results from Figure 31 to those in Figure 27 are compared, it is clear that the attribute context constraints helped the network to learn faster (i.e., the constrained network reached the same RMS error in approximately 140 epochs that was reached in 180 epochs in the unconstrained network). The output for cell one (instance unit) has a higher error rate than the output units for the attributes. As network learning proceeds the attributes are learned at a faster pace than are the instance/non-instance relationships. Thus learning the attributes appears to have the effect of reducing the instances' unit error rate. Note that the card attributes were learned sooner than instance as would be expected of a hierarchy in which knowledge of attributes is a prerequsite.

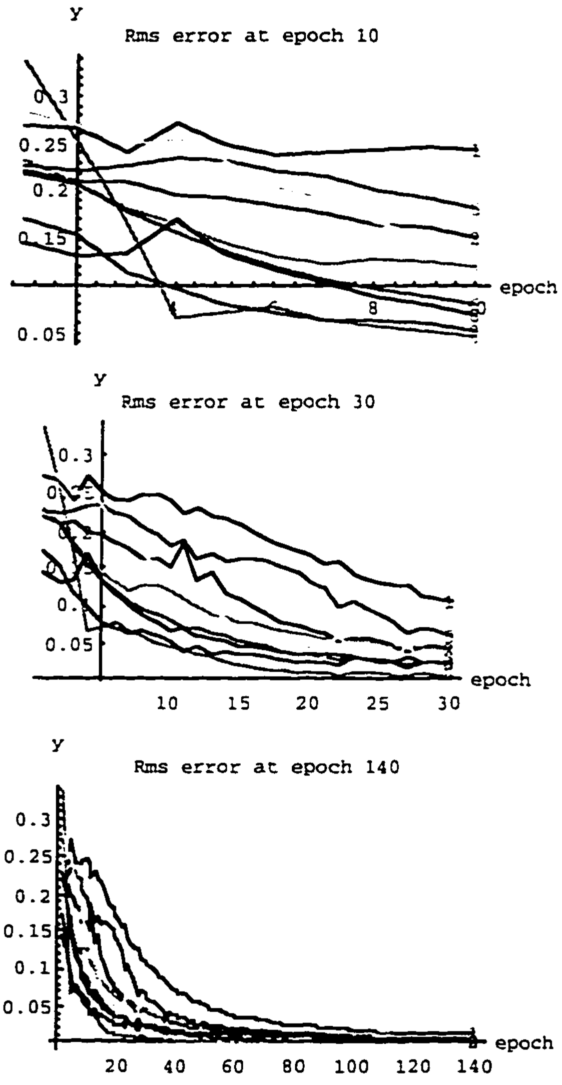


Figure 31. RMS error for the nine output units at three different epoch points.

Figure 32 shows the set of five histograms (one for each card in the examplar training set) that indicate the nine output units responses after training. As described earlier, the y-axis shows the output reponse level (from 0.0 to 1.0) for the output units and the x-axis shows the individual histogram bars for each of the nine output units. A comparison can be made between the histogram bars produced by each card and their respective target as described in Table 6.

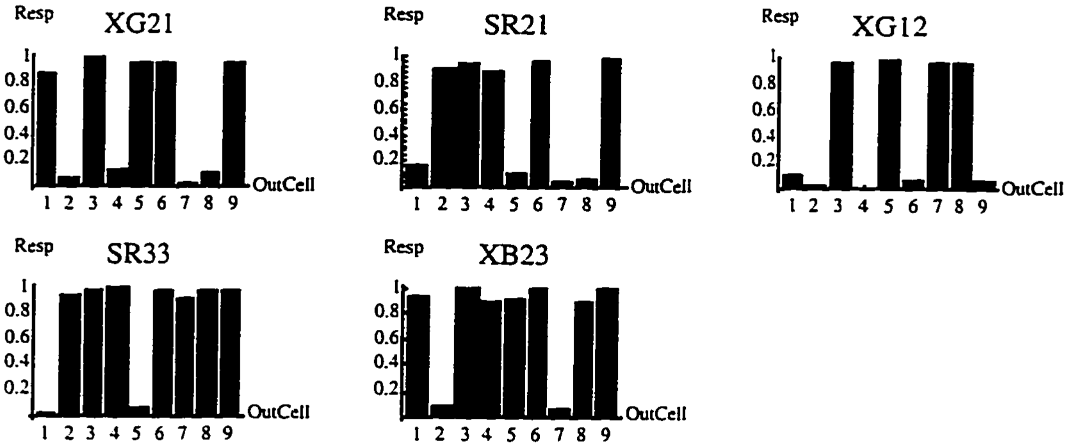


Figure 32. The output responses for the five exemplar training set cards.

In Figure 32 for the card XG21, the first bar represents output cell one and is very close to 1.0. This indicates that the network identifies XG21 as a positive instance of the concept. This is also the case for the card XB23. In other words, in both cases (XG21 and XB23) the first output unit responded TRUE for the concept two and cross. The remaining eight output units describe the card attributes: shape (units two and three), color (units four and five), number (units six and seven), and borders (units eight and nine). The network learned that SR21, XG12, and SR33 are negative instances, as shown by the value of the first output unit of their respective histograms (these values are close 0.0). Figure 32 also shows that attributes shape, color, number and borders were correctly identified. These results are summarized in Table 7.

Table 7. Training set cards and learned instances and attributes

Card	Instance	Shape	Color	Number	Borders
XG21	√ (+)	√	√	√	√
SR21	√ (–)	√	√	√	√
XG12	√ (–)	√	√	√	√
SR33	√ (–)	√	√	√	√
XB23	√ (+)	√	√	√	√

3.5.1.2 Categorization of constraint attributes

Using the final weights obtained, the network was presented with all the cards from the Bruner et al. (1956) card set that contained the concept two and cross. In each case the network categorized them to be true instances of the concept. Figure 33 shows the nine plots for each of the cards that contain the concept two and cross. For the purposes of intrepreting the heights of the bars in each plot, the position is taken that the response values represent the degree of certainty that the network developed that the output is 1.0 for a given output cell. For example, XR22 has a "degree of certainty" of 0.75 that the output is truly 1.0, i.e., a positive instance.[19] These results are summarized in Table 8. Thus the network was able to generalize (inductively reason in the Bruner et al. 1956 sense) to novel card instances never before seen with a higher rate of accuracy than without the attribute context constraints. Note that the three cards not learned previously are now learned when the attribute constraints are added (compare Figure 28 with Figure 33). Similar results were obtained for the disjunctive concept set in Figure 24. The level of ecological validity has increased for two reasons: (a) generalization to the card population is better, and (b) the response at the output layer provides a more valid description of what has been learned with respect to the input representation.

[19] The term degree of certainty cannot be interpreted as a true statistical probability.

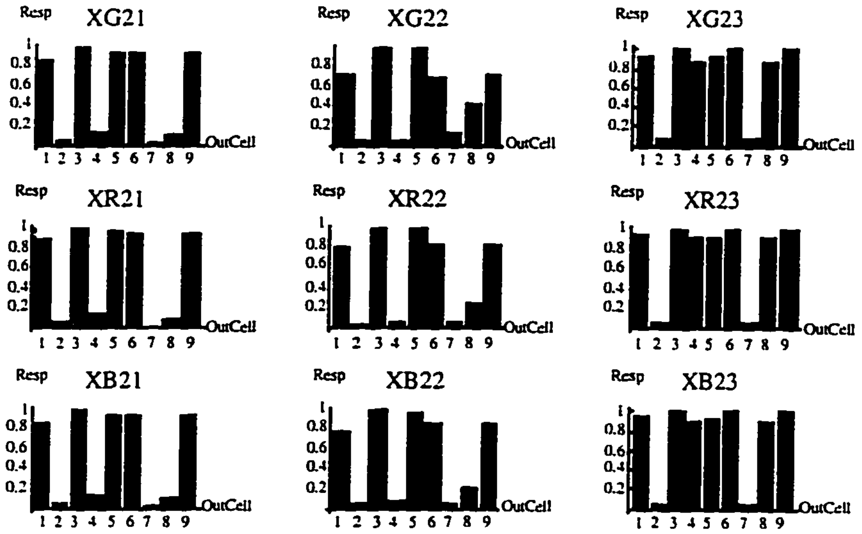


Figure 33. The output responses for all nine cards Bruner et al.'s card set that contain the concept two and cross.

Table 8. Network response to instance and attributes for cards in Figures 32, 33, 34.

Card	Instance	Shape	Color	Number	Borders
*XG21**	√ (+)	√	√	√	√
XG22	√ (+)	√	√	√	?
XG23	√ (+)	√	×	√	√
XR21	√ (+)	√	×	√	√
XR22	√ (+)	√	×	√	?
XR23	√ (+)	√	×	√	√
XB21	√ (+)	√	×	√	√
XB22	√ (+)	√	×	√	×
*XB23**	√ (+)	√	√	√	√
SR22	√ (−)	√	√	√	×
CR23	× (−)	×	?	√	√

* Training set card.
? Low certainty on one of two outputs

The network was then presented with cards that contained squares and it classified them as negative concept instances. For example, Figure 34 shows the output for card

SR22 which was typical for cards that represent negative instances. The network solution was then tested with cards that contain two circles. Figure 35 shows the typical output units response for cards with two circles (in this case the CR23 card). The results for the cards SR22 and CR23 are summarized in Table 8. The network generalized card CR23 to be a true instance of the concept but with less certainty than the positive instance of the concept (see Figure 32). In other words, the concept two and circle reflects a positive instance of the hypothesized concept. This result appears counterintuitive given that the concept of the network was assumed to be two and cross. The network's behavior would indicate a type of overgeneralizing to unseen instances that are closely related to the concept class. Note that the visual stimuli (CR23) contained a shape (a circle) never before seen by the network during training. Also note that these circles are in the context of similar surrounding attributes for training set cards SR21 and SR33 that are negative instances, and for training set card XB23 that is a positive instance. In this situation the network appears to recognize a novel unseen instance as a previously learned one and classifies it as part of the concept class. Over generalization of a visual prototype is not uncommon for human infants. For example, Bomba and Siqueland (1983) found that 3-4 month olds exhibited overgeneralized behavior in the context of prototype dot image patterns. Table 8 also indicates that the attribute of color was in error for most cards representing a positive instance and which were not in the training set. Black and red were incorrectly categorized twice, and green once. Borders also showed incorrect categorization, but not to the same extent as color. The attributes of shape and number had almost no errors of categorization.

Further interpretation of network behavior and a more detailed analysis of the internal weight space representation is presented in the next chapter.

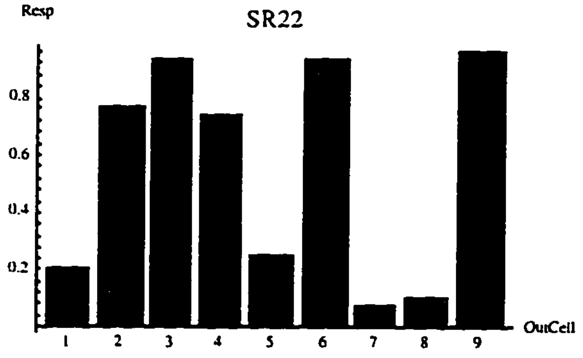


Figure 34. The output responses for the red card with two squares and two borders.

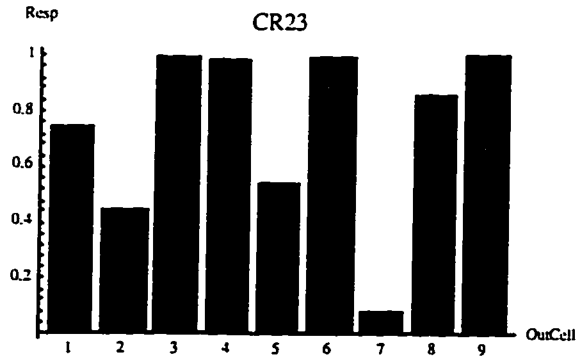


Figure 35. The output responses for the red card with two circles and three borders.

3.5.2 Summary of results for constrained and unconstrained networks

Although it might be expected that the constrained network takes more presentations to reach a given level of categorization than an unconstrained network because more calculations are involved in learning a larger number of attributes, the results indicate the opposite. The constrained network took less time to learn the appropriate categorizations of instances and attributes than the unconstrained network took to learn the categorization of instance only. The results also indicate that the constrained network had better capacity to generalize to card instances not seen previously. The latter is indicated by comparing the instance output for cards XG22, XR22, and XB22 having positive instances in Figures 28 and 32. Even though the attributes exhibited in the training card set were correctly learned by the constrained network, when the network was presented with previously unseen positive instances having different combinations of attributes, the network showed the greatest number of errors for color, and a lower degree of certainty for borders for some exemplars. No errors of categorization appeared for the attributes of shape and number. The network is thus able to discriminate on the critical concept features of "number of objects" and "shape" that define the concept instance. Furthermore the constrained network over generalized the concept by categorizing circles as crosses.

3.6 Conclusion

Traditionally experimental methods in psychology have attempted to limit the effects of extraneous variables in an effort to increase validity. Network models of cognitive processes often make various assumptions regarding the structure and function of the input. These assumptions can directly impact the ecological validity of the network and influence the network performance.

Two ways of preserving network ecological validity were investigated. The first involves structuring the input to reflect the "real world" objects used during the experimental procedures with subjects. In this case a mapping strategy was developed to retain, as much as possible, the visual representation of the actual Bruner et al. (1956) card set.

The second method is influenced by Gagne's (1985) work that suggests discrimination learning plays a significant role in the conceptual development of humans. To address this issue of discrimination learning the number of units in the output layer was expanded to include constraints for a given set of attributes in a given context (thus the phrase "attribute context constraints"). These attribute context constraints direct the network towards a better overall representation by allowing for a higher level of generalization to the overall card population of exemplars. A suggested reason for the success of these attribute context constraints is the functional relationship formed between the input and output layers. Another added benefit of using these constraints appears to be an increase in the speed of network learning.

The role of extra output units to direct network learning is largely unexplored in the context of connectionist modeling of concepts. For example: Can attribute context constraints be used in other connectionist studies of concepts? How does the internal structure (hidden weight space) of the network change as extra output units are added? (see Chapter 4). Furthermore, attribute context constraints offer the possibility of experimenting with Gagne's (1962) learning hierarchies, which are an important component of many pedagogical strategies. Thus knowledge from instructional theory, in the form of attribute context constraints, can play an important role in defining the physical structure of the network.

Finally, the results indicate a connectionist network can be constructed to solve Bruner et al.'s (1956) concept attainment task. Further research would be a more detailed connectionist investigation of Bruner et al.'s work, namely modeling the strategies people employ to determine the relevant features of a concept. That is, given that networks can be structured to achieve the goal of concept attainment, do they implement similar types of strategies that people do?

3.7 References

Anderson, J. R. (1990). Cognitive psychology and its implications (3nd ed.). W. H. Freeman and Company: New York.

Anderson, J. R. (1993). Problem solving and learning. American Psychologist, 48(1), 35-44.

Barsalou, L. W. (1983). Ad hoc categories. Memory and Cognition, 11, 211-227.

Bates, E., Bretherton, I., & Snyder, L. (1988). From first words to grammar: Individual differences and dissociable mechanisms. Cambridge, MA: Cambridge University Press.

Bechtel, W., & Abrahamsen, A. (1991). Connectionism and the mind: An introduction to parallel processing in networks. Cambridge, MA: Blackwell.

Bomba, P. C., & Siqueland, E. R. (1983). The nature and structure of infant form categories. Journal of Experimental Child Psychology, 35, 294-328.

Bower, G. H., & Clapper, J. P. (1989). Experimental methods in cognitive science. In M. I. Posner (Eds.), Foundations of cogntive science Cambridge, MA: MIT Press.

Bracht, G. H., & Glass, G. V. (1968). The external validity of experiments. American Educational Research Journal, 5(4), 437-474.

Bruner, J. S. (1966). Toward a theory of instruction. Cambridge, MA: Harvard University Press.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). A study of thinking. New York: John Wiley & Sons.

Bruner, J. S., Olver, R. R., & Greenfield, P. M. (1966). Studies in cognitive growth. New York: Wiley.

Carpenter, G. A., & Grossberg, B. (1988). The art of adaptive pattern recognition by a self-organizing neural network. IEEE Computer, 21(3), 77-88.

Chauvin, Y. (1988) Symbolic acquisition in humans and neural (PDP) networks. unpublished doctoral dissertation, University of California.

Clark, A. (1993). Associative engines: Connectionism, concepts, and representational change. Cambridge, MA: MIT Press.

Dick, W., & Carey, L. (1985). The systematic design of instruction (2nd ed.). Glenview, IL: Scott, Foresman and Company.

Eimas, P. D. (1994). Categorization in early infancy and the continuity of development. Cognition, 50, 83-93.

Freeman, J. A. (1994). Simulating Neural Networks with Mathematica®. Menlo Park, CA: Addison-Wesley.

Gagne, R. (1962). The acquisition of knowledge. Psychological Review, 69(4), 355-365.

Gagne, R. M. (1984). Learning outcomes and their effects. American Psychologist, 39(4), 377-385.

Gagne, R. M. (1985). The conditions of learning and theory of instruction (4th ed.). Toronto, ON: Holt, Rinehart and Winston.

Gagne, R. M., & Briggs, L. J. (1979). Principles of instructional design (2nd ed.). New York: Holt, Rinehart and Winston.

Gallant, S. I. (1993). Neural network learning and expert systems. Cambridge, MA: MIT Press.

Gallmo, O., & Carlstrom, J. (1995). Some experiments using extra output learning to hint multi layer perceptrons. In L. F. Niklasson & M. B. Boden (Ed.), Current trends in connectionism - Proceedings of the 1995 Swedish conference on connectionism, (pp. 179-190). Lawrence Erlbaum.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. Journal of Experimental Psychology: General, 117(3), 227-247.

Hecht-Nielsen (1990). Neurocomputing. New York: Addison-Wesley.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. Neural Networks, 2, 77-84.

Horon, P., & Lynn, D. (1980). Learning hierarchies research. Evaluation in Education, 4, 82-83.

Keil, F. (1987). Conceptual development and category structure. In U. Neisser (Eds.), Concepts and conceptual development Cambridge University Press.

Klahr, D. (1992). Information-processing approaches to cognitive development. In M. H. Bornstein & M. E. Lamb (Eds.), Developmental psychology: An advanced textbook (pp. 273-335). Hillsdale, NJ: Lawrence Erlbaum Associates.

Klausmeier, H. J. (1992). Concept learning and concept teaching. Educational Psychologist, 27, 267-286.

McClelland, J. L., & Rumelhart, D. E. (1989). Explorations in parallel distributed processing: A handbook of models, programs and exercises. Cambridge, MA: MIT Press.

Medin, D. L. (1989). Concepts and conceptual structure. American Psychologist, 44(12), 1469-1481.

Newell, A., & Simon, H. A. (1972). Human problem solving. Englewood Cliffs, N.J.: Prentice-Hall.

Plunkett, K., & Sinha, C. (1992). Connectionism and development psychology. British Journal of Developmental Psychology, 10, 209-254.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. Journal of Experimental Psychology, 77, 353-363.

Posner, M. I., & Keele, S. W. (1970). Retention of abstract ideas. Journal of Experimental Psychology, 83, 304-308.

Quinlan, P. (1991). Connectionism and psychology: A psychological perspective on new connectionist research. Chicago, IL.: The University of Chicago Press.

Rosch, E. (1975). Cognitive representations of semantic categories. Journal of Experimental Psychology, 104, 192-233.

Rosch, E. (1988). Principles of categorization. In A. Collins & E. E. Smith (Eds.), Readings in Cognitive Science (pp. 312-323). Palo Alto, CA: Morgan Kaufmann.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumlehart & J. L. McClelland (Eds.), Parallel distributed processing: Explorations in the microstructure of cognition Cambridge, MA: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (Ed.). (1986). Parallel distributed processing: Explorations in the microstructure of cognition. Cambridge, MA: MIT Press.

Schneider, W., & Graham, D. J. (1992). Introduction to connectionist modeling in education. Educational Psychologist, 27(4), 513-530.

Schunk, D. H. (1996). Learning theories: An educational perspective (2nd ed.). Toronto: Prentice Hall.

Schyns, P. G. (1991). A modular neural network model of concept acquisition. Cognitive Science, 15, 461-508.

Siegler, R. S. (1991). Children's thinking (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Siegler, R. S. (1995). How does change occur: A microgenetic study of number conservation. Cognitive Psychology, 28, 225-273.

Simon, T. J., & Halford, G. S. (1996a). Computational models and cognitive change. In T. J. Simon & G. S. Halford (Eds.), Developing cognitive competence: New approaches to process modeling (pp. 1-29). Hillsdale, NJ: Lawrence Erlbaum Associates.

Simon, T. J., & Halford, G. S. (Ed.). (1996b). Developing cognitive competence: New approaches to process modeling. Hillsdale, NJ: Lawrence Erlbaum Associates.

Smith, E. E. (1988). Concepts and thought. In R. J. Sternberg & E. E. Smith (Eds.), The psychology of human thought (pp. 19-47). Cambridge, MA: Cambridge University Press.

Smith, E. E., & Medin, D. L. (1981). Categories and concepts. Cambridge, MA: Harvard Unviversity Press.

Suddarth, S. C., Sutton, S. A., & Holden, A. D. C. (1988). A symbolic-neural method for solving control problems. In IEEE International Conference on Neural Networks: Vol. 1. (pp. 515-523). San Diego, CA:

van der Veer, R., & Valsiner, J. (1991). Understanding Vygotsky: A Quest for Synthesis.

Vygotsky, L. (1934/1986). Thought and language (newly revised and edited by Alex Kozulin, Trans.). Cambridge, MA: MIT Press.

Wolfram, S. (1991). Mathematica® : A system for doing mathematics by computer. Redwood City, CA: Addison-Wesley.

Yu, Y. H., & Simmons, R. F. (1990). Extra output biased learning. In Proceedings of the International Joint Conference on Neural Networks (IJCNN-90), 3 (pp. 161-166). San Diego, CA:

4. Network Analysis

4.1 Introduction

There appear to be a number of alluring reasons for using the connectionist approach to model cognitive processes. The most obvious allure is the notion that such models are neurally plausible (Rumelhart & McClelland, 1986). A second, and one which is closely linked to issues of development, is that such models form internal representations based on learned experiences. That is, over some period of time, a connectionist model will develop a representation which map inputs (real word experiences) to outputs (actions taken with respect to real world experiences). The observed behavior of such a model may appear, on the surface at least, to be remarkably similar to that of the actual observed cognitive behavior. But on what grounds can such an isomorphic relationship be substantiated? Herein lies a major area of concern for many connectionist modelers. Essentially, this concern amounts to trying to understand the network's internally formed representation both during its formation and in its final form (Robinson, 1992).

The problem of network interpretation stems from the fact that knowledge is distributed across weights associated with cells and from the nonlinear properties used to form this weight space over time (epochs). Furthermore, such a system forms a knowledge representation structure based on the training set (input) and the starting state of the weights; however, both of these may vary in their initial form. Currently, there is no generally agreed upon form of network analysis. Thus research into understanding the internal workings of the network is exploratory and must be tailored specifically to the task being modeled.

A number of techniques have been used to make some sense of the network's internal structure formed during learning. An important factor that can contribute to our interpretation of a network is to create a visual depiction of the internal network representation (Hunka & Carbonaro, to appear). Hinton (1986) suggested one could infer certain facts about weight data by visualizing the data in a diagrammatic fashion, often referred to as "Hinton Diagrams." One underlying problem with this approach is that a one-to-one mapping must be known between input exemplar and the associated cell's weight vector. For simple feed forward networks with two layers the mapping between exemplar and weights is obvious. Unfortunately, this mapping becomes more complex with the addition of hidden layers. This is because the hidden layer calculates a response vector (that is mapped against the next higher level set of weights) and it becomes difficult to determine the interpretation between this hidden layer activation vector and the original input exemplar.

Another analytical technique which appears to be effective in understanding the solution structure of the network is a statistical analysis of the activation patterns for hidden response cells and the internal weight space (Hanson & Burr, 1990; Shultz & Elman, 1993). Hanson and Burr (1990) suggest two reasons for analyzing the network's representation using statistical techniques:

> First, there is a similarity between the multivariate analysis of data (or psychometric modeling) and the learning/representation relationship in nets. An understanding of the net's internal representation could make a contribution to psychological theory too. Using nets that learn is like doing exploratory data analysis (Tukey, 1977). Moreover, a judicious and critical analysis of the net's representation can reveal what it has discovered through exposure to data. The second reason for analyzing what nets learn is that learning deserves study in its own right. Analyzing what the net computes after learning provides a simple empirical way of exploring the intimate relationship between learning and representation. (p. 482)

Hanson and Burr outline specific types of multivariate analysis (cluster analysis) and more exploratory type techniques such as factor analysis or multidimensional scaling for network interpretation.

One qualifier when using linear modeling statistical methods to analyze the networks representation is that such representations are often formed nonlinearly. For example, in a simple two layer type network statistical techniques may be quite suitable for the interpretation of the representation because decision boundaries are linear (Lippmann, 1987). In such cases the functions to be learned are said to be linearly separable (Minsky & Papert, 1969). More complex three layer networks might require nonlinear decision regions and involve nonlinear interrelationships among hidden cell values (McClelland & Rumelhart, 1989). These nonlinear interrelationships are not well suited for multivariate analysis which assumes linear relations among variables. For example, "if we associate variables with [network] cells in a multivariate analysis such as factor analysis, it is quite possible that the interaction between the weights or activation of two cells will be nonlinear" (Hanson & Burr, 1990, p. 488).

More recently, Berkeley et al. (1995) have attempted to overcome some of the limitations and difficulties associated with multivariate analysis by using banded density plots. Banded density plots are graphical representations that reflect the hidden cell activation levels for the patterns used to train the network. The bands can be mapped back to the input patterns in such a way that specific features of these patterns can be defined with respect to the type of bands they produce. The banding technique thus provides a simple means of interpreting the network's representational validity in the context of both architectural and environmental constraints.[20]

According to Berkeley et al. (1995) the use of bands to interpret trained network structure has two advantages over the multivariate approach: (a) it incorporates an aspect of nonlinearity because it works with the hidden cell activation levels, and (b) it avoids the artifact type assumptions (e.g., factor rotation) and instead relies on the natural clustering produced by the network.

In all of the above cases, the analysis was performed on the network's final state. In contrast, McClelland and Jenkins (1991) used diagrams to plot the dynamic learning performance of the network at various stages during the development of the network's representation. A key part of their visual analysis is the plotting of graphs that show the epoch by epoch performance. Their intent was to demonstrate the model's stage-like conceptual performance on the Piagetian balance scale problem. Thus an understanding of dynamic network learning becomes important when the model concerns issues of development.

In many network interpretations there is an assumption that networks develop their own rule-based knowledge. In this sense rules are inherently developed during the network learning process and are embodied in the weights and nonlinearity used to modify the weights over time. In contrast, the traditional symbolic AI approach used rules which were explicitly defined for the model (Jackson, 1990). Defining the isomorphic relationship between the connectionist cognitive model and the theoretically defined conceptual task (such as Bruner et al.'s 1956) involves a process of making these inherent rules more explicit. This chapter attempts to interpret the inherently developed rules for both constrained and unconstrained networks used to model Bruner et al.'s concept attainment task.

This chapter is divided into the following sections: Section two examines the dynamic learning aspect of the network model. Section three analyzes the final state of the

[20] The banding interpretation technique appears to perform better when networks are defined using the value unit architecture (Dawson & Schopflocher, 1992). Such networks replace the logistic activation function with a non-monotonic activation function.

network. The final section provides a summary of the conclusions derived from these analyses.

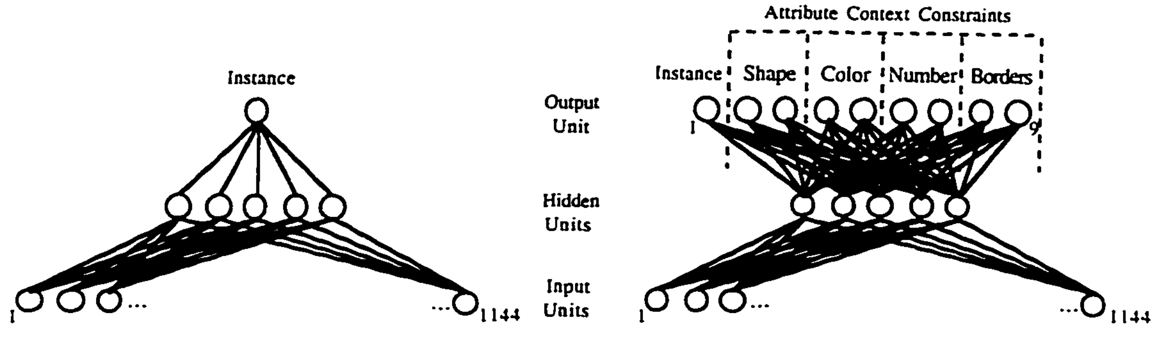


Figure 36: The two networks used to model Bruner's et al. (1956) concept attainment task.

4.2 The Bruner concept attainment task

Bruner, Goodnow, and Austin's (1956) research on concept attainment was re-examined from a connectionist perspective. Bruner et al. (p. 42) used a set of cards to study concept attainment consisting of four attributes, each of which varied on three values: shape (cross, circle, square); color (green, black, red); number of objects (one, two, or three); number of borders (one, two, or three). Each card instance combines one value of each of the four attributes. A category or a concept (conjunctive, disjunctive, and relational) is defined with respect to a subset of cards that share a common set of attribute values.

Two different PDP network models of the Bruner et al.'s (1956) concept attainment task were develop to compare the use of attribute context constraints.[21] These constraints are based on Gagne's (1985) theory of concept learning and require a learner to make discriminations to identify distinctive features of stimulus objects. Essentially these constraints resulted in the addition of extra output cells that help guide the network in constructing a better representational model than those models in which the constraints were absent. Figure 36 shows these two models. The structure of input and hidden layers are the same for each model. The output layers are different. The left model is unconstrained and contains only one output cell, the right model contains an additional eight outputs that are used to define the attribute context constraints. In both models network learning requires the card to be classified as either a positive or negative concept instance but the constrained network is also required to simultaneously learn to make the

[21] See Rumelhart and McClelland (1986) for a detailed explanation of PDP networks.

necessary discriminations among the attribute values associated with the card instances. The network models were trained on the set of card instances shown in Figure 37 using the standard back propagation error correction algorithm (Rumelhart, Hinton, & Williams, 1986). The same initial set of starting weights was use in both cases.

A conjunctive concept, to be attained by the model, is identified in this card set as two and crosses. Only the conjunctive concept type was explored in this analysis. In Figure 37 the plus (+) designates the card is a positive concept instance and the minus (-) designates the card is a negative concept instance. Network input was constructed using a one-to-one mapping between each card's visual representation and a 2-dimensional numeric array. This array had 1144 cells, where colored pixels green, red and black were assigned the values of 0.25, 0.5, and 0.75 respectively (white pixel values were 0.0).

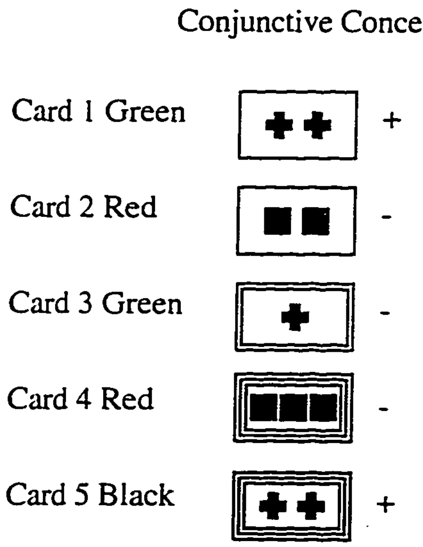


Figure 37: The set of card instances used as input data for the models.

The unconstrained network reached an acceptable level of convergence after 200 epochs whereas the constrained network reached convergence after only 140 epochs. The following sets of data were saved for network analysis: (a) the set of responses at both hidden and output layer for each exemplar during all epochs, and (b) the set of weights for hidden and output layers after each epoch. As a result, a large data set is generated for each epoch during network training. This was especially the case for the set of weights at the hidden layer given that there are 1144 input cells that resulted in a set of 5720 individual weights that needed to be saved after each epoch. Thus, for 200 epochs more than approximately one million numbers were saved.

4.3 Interpretation of the networks dynamic learning

The interpretation of dynamic learning focused on the changes of the cellular responses at both hidden and output layer. In this context dynamic learning refers to the changes in cellular response values that occurred during the epoch by epoch performance of the network with respect to the input exemplars.

4.3.1 Hidden layer responses

The responses produced at the hidden layer are shown in the set of graphs in Figure 38. The left side shows the responses given for the five hidden cells in the unconstrained network. The right side shows the five hidden cell responses for the constrained network. The y-axis of each graph indicates the response level of the hidden cell. Not all graphs show the same scale of response values but all graphs must have responses within the range of 0.0 to 1.0. The x-axis of each graph indicates the number of epochs. In Figure 38 Each colored line designates one exemplar in the set of 5 cards for the conjunctive concept. Table 9 shows the mapping scheme between colored lines and the cards in Figure 37.

Table 9: Mapping between colored lines and card exemplars.

Colored lines	Card exemplar
Black	1
Blue	2
Green	3
Sky Blue	4
Red	5

The end points of each colored line in the graph also show the exemplar number. The cells generally show faster convergence for the constrained network than for the unconstrained network with the exception of cell 4 in the constrained network in which the response for exemplar 4 took more than 120 epochs before it converged. Cell 2 in the constrained network converged almost completely after 30 epochs whereas in the unconstrained network convergence did not begin until 180 epochs. At that point exemplars 1, 2, and 5 began to separate themselves out. Also note that exemplar 3, in cell 5 of the constrained network, converged faster and achieved a higher response value than it did in hidden cell 5 of the unconstrained network.

One way to view the output response at the hidden layer is as a set of rules that was developed during network learning. An individual hidden cell will respond (produce output) in accordance with its own rule when the cell is presented with an exemplar. For example, the response pattern for hidden cell 4 in the unconstrained network (see Figure 38) can be rewritten in rule form as:

> **IF** (exemplar one **OR** exemplar five) **THEN** cell response is approximately 1.0
> **ELSE** cell response is approximately 0.0

Viewed in this context all cells can be said to have developed their own rule-based response pattern to network input. One problem with understanding network functionality at this level is that cell rules are developed both: (a) nonlinearly, and (b) in parallel with all other cells. Thus it is the total combination of all cells that forms an N-dimensional space that reflects the nonlinearity of these inherently formed rules. A 3-D plot can be used to demonstrate this nonlinear parallel formation of rules between cells. The three axes are labeled after each hidden cell they represent. Each axis indicates the range of response values for its respective hidden cell. Figure 39 shows the joint responses of three cells (2, 4 and 5) for each exemplar over 200 epochs. Clearly these responses to exemplars form nonlinear patterns. Exemplar 1 and 5 converge in the same general area of the 3-space and represent positive instances of the concept. Exemplars 2, 3, and 4 appear closer together diagonally across and represent negative concept instances. Contrasting Figure 39 with Figure 40 shows the 3-D plot for hidden cells 2, 3, and 4. In this case the negative concept instances appear clearly separate from each other but the positive instances converge on almost the same spot.

The reason for selecting these two sets (unconstrained and constrained) of three cells for plotting is based on a qualitative assessment of the set of the unconstrained network graphs show in Figure 38. For example, in the case of the unconstrained networks both cells 1 and 3 appear to output values close to 0.0 and seem to respond similarly to all exemplars. Therefore they were deemed to be of less significance in the separation of the exemplars for network convergence.

4.3.2 Output layer responses

Figure 41 shows an early convergence of the responses for the nine output cells in the constrained network. Figure 42 shows that exemplar 2 required 180 epochs to reach convergence. The unconstrained network cell in Figure 42 and cell 1 of the constrained network in Figure 41 are both required to make the same separation between exemplars. In the constrained network exemplar 2 begins convergence at 20 epochs as opposed to the unconstrained network where it requires 180 epochs to reach the same state. The problem the unconstrained network had was to classify card 2 (negative instance), which differed in color and shape only from the card 1 (positive instance). In other words, the unconstrained network which is not required to demonstrate learning of color, border, etc., encounters difficulty classifying closely related exemplars of differing instance.

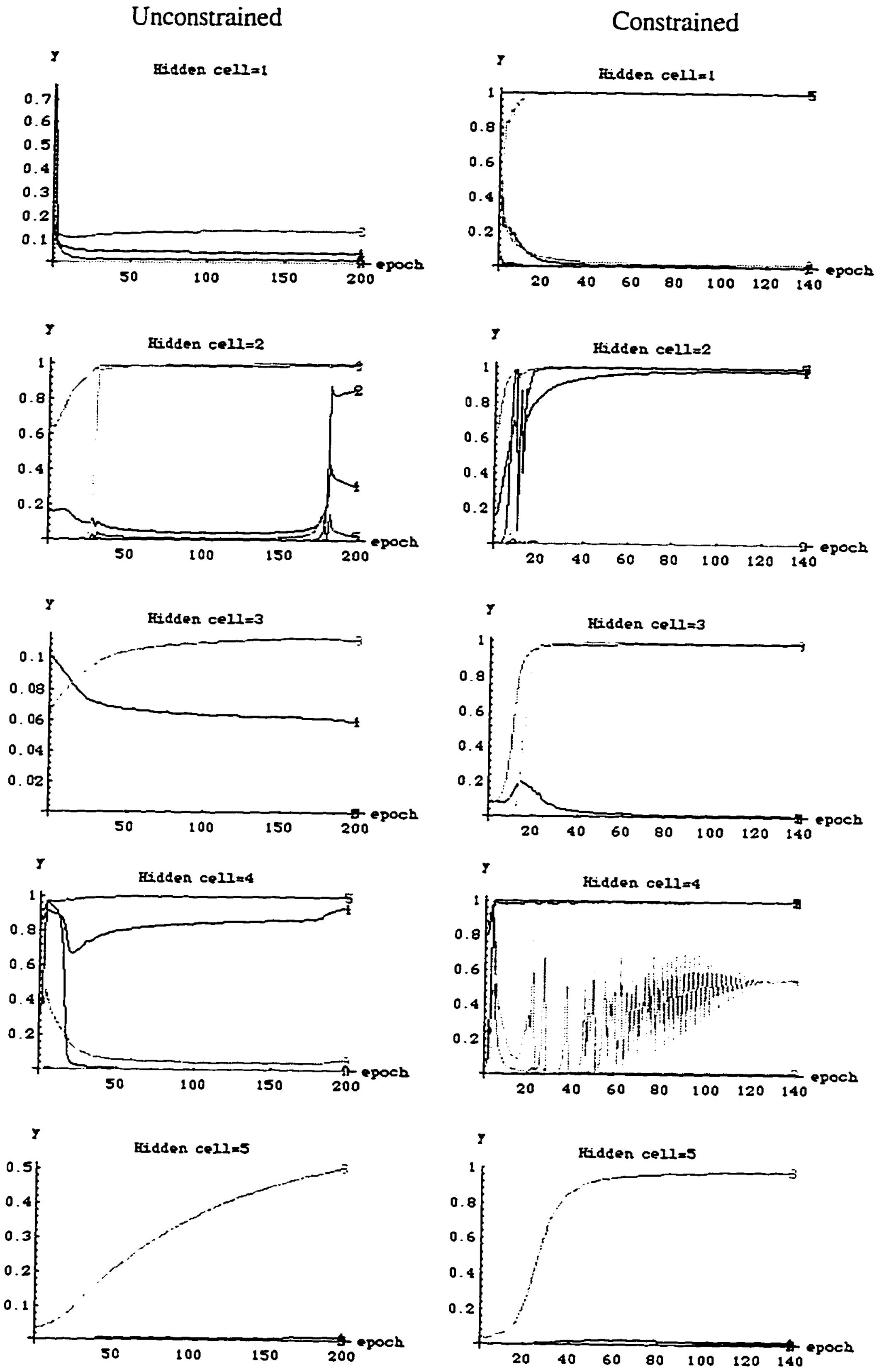


Figure 38: Response migration for hidden cells in constrained and unconstrained networks.

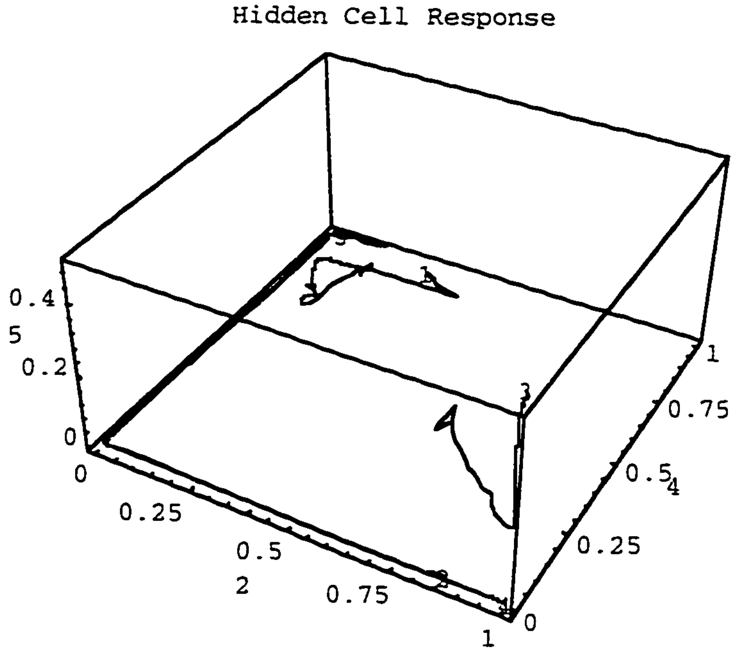


Figure 39. Hidden cells 2, 3 and 5 for responses in the unconstrained network.

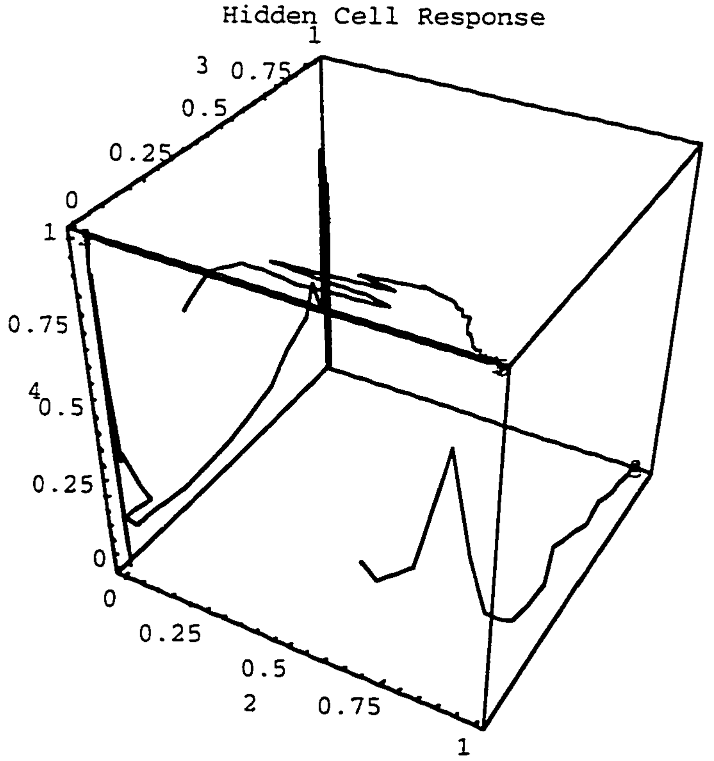


Figure 40. Hidden cells 2, 3 and 4 responses the constrained network.

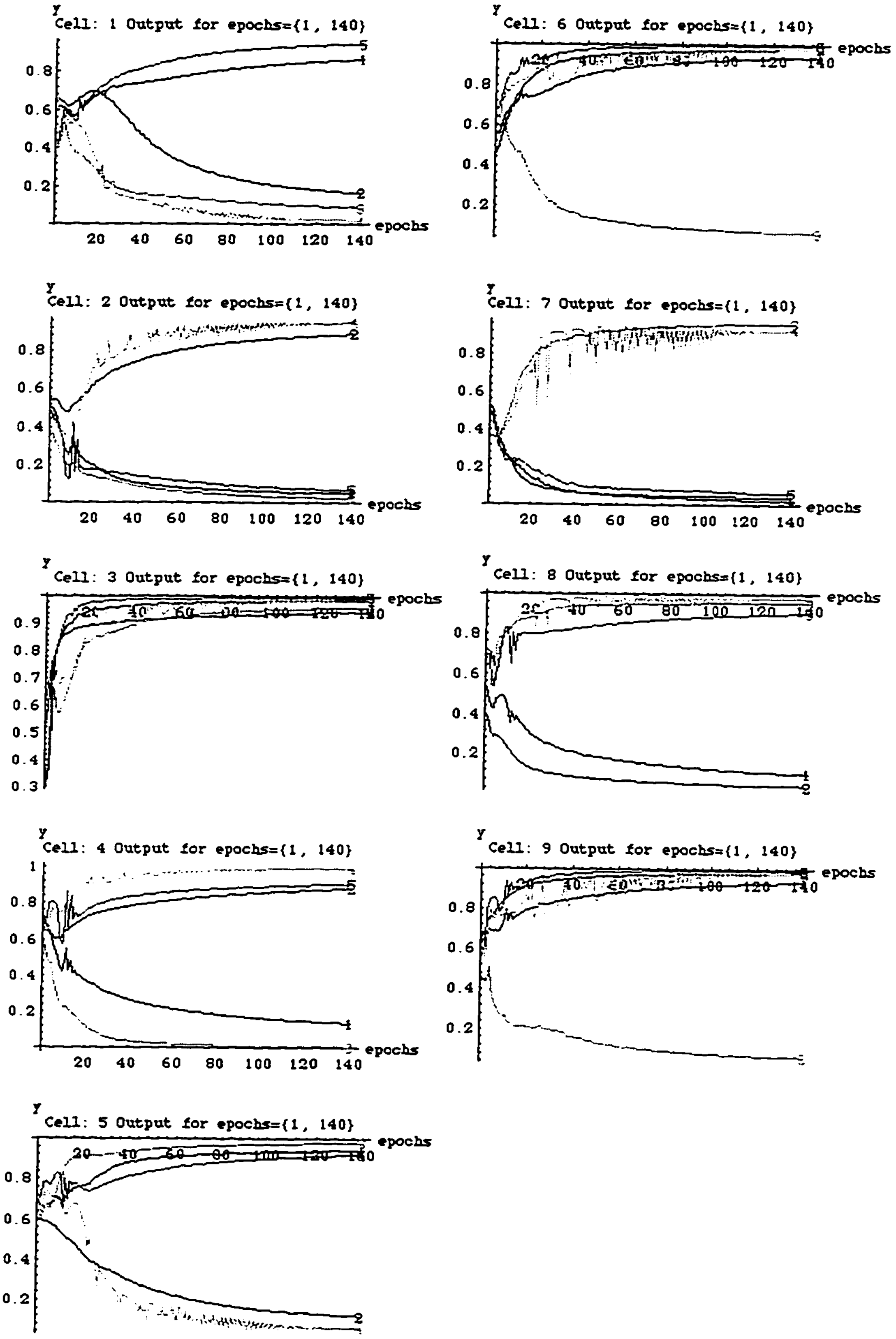


Figure 41: Response migration for output cells in constrained network.

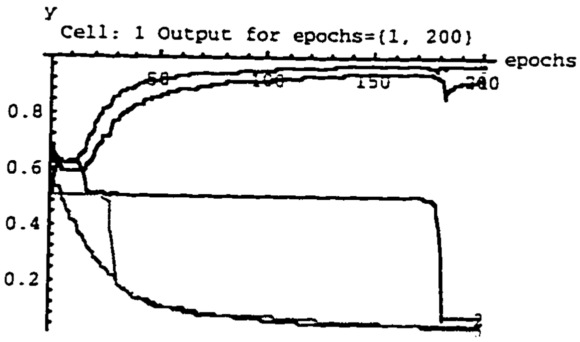


Figure 42: Response migration for output cell in unconstrained network.

4.4 Network's final state

The network's final state is reflected in the total set of weights and the final set of responses produced by each exemplar of the training set. The analysis of the network's final state was restricted to the set of response values produced at the hidden layer on the last epoch.

The analysis technique used here was to perform a singular value decomposition on a data matrix of hidden cell response patterns by exemplars. For the conjunctive concept in Figure 37, the matrix of order (5,5) was resolved into its components (eigenvalues and eigenvectors) by:

$$D=U\Gamma W$$

where D is of order N hidden responses and n exemplars, U are the left eigenvectors of order (N,n), Γ is a diagonal matrix of eigenvalues of order (n,n) and W are right eigenvectors of order (n,n). The left eigenvectors are parameters descriptive of the hidden cells while the right eigenvectors describe the exemplars. The resolution of the data matrix into additive components can be represented by the following:

$$D= U_1\Gamma_1W_1 + U_2\Gamma_2W_2 + U_3\Gamma_3W_3 + \ldots + U_N\Gamma_nW_n$$

Thus matrix D is the sum of an orthogonal set of components. Examination of the eigenvalues can provide information on which components are most significant in accounting for the best least squares estimate of D, i.e., which subset of orthogonal components of U and W can closely estimate D. Given the final state of responses in the unconstrained network, Figure 43 shows the result of plotting the largest three principal components of W (i.e., $\Gamma^{1/2}W$) based on the largest three eigenvalues (1.74, 1.32, and 0.42) and indicates the dimensions by which the hidden layer discriminates among the exemplars.

The three axes are labeled for each of the principal components and indicate the range of component values. Most notably, exemplars 2, 3, and 4 are clustered to the left and define those instances that are negative, while exemplars 1 and 5 are clustered to the right and define instances that are positive. Exemplars 2 and 4 cluster in the bottom right hand corner, both contain square shapes and are colored red. Exemplar 3 is in the upper left quadrant and contains two borders and one cross.

Figure 44 shows the Varimax rotation of the principal components. In this case the negative exemplars (2, 3, and 4) are clearly separate at opposite ends from the positive exemplars (1, and 5). Note that exemplar 3 remains quite distinct after Varimax and exemplar 2 and 4 remain close together.

Figure 45 shows the results of plotting the largest three orthogonal components based on the largest three eigenvalues (2.56, 1.54, and 1.27) for the constrained network. As was the case in the unconstrained network it indicates the dimensions by which the hidden layer discriminates among the exemplars. In this case the three negative exemplars are more widely dispersed in comparison to the unconstrained network, whereas the positive exemplars (1 and 5) are close together as was the case for the unconstrained network. After Varimax rotation, exemplars 1 and 5 are distinguishable from exemplars 2, 3, and 4 on the first dimension with the former having high values (approximately 0.8) and the latter values closer to 0.0. On the second dimension, negative exemplars 2, 3, and 4 receive values in the range 0.4 to 0.8, while the positive exemplars are close to 0.0. The third dimension identifies most markedly the negative exemplar 4 from the remaining exemplars.

This analysis suggest quite clearly that the constrained network provides a greater degree of differentiation among the negative exemplars, with less separation between the positive exemplars. For example, exemplar 4 has the attributes red, three borders, and three squares, and is the most different of the negative exemplars. However, the positive exemplars are collectively distinguishable from the negative exemplars. Thus singular value decomposition proved to be a valuable technique for interpreting hidden cell responses with respect their input exemplars.

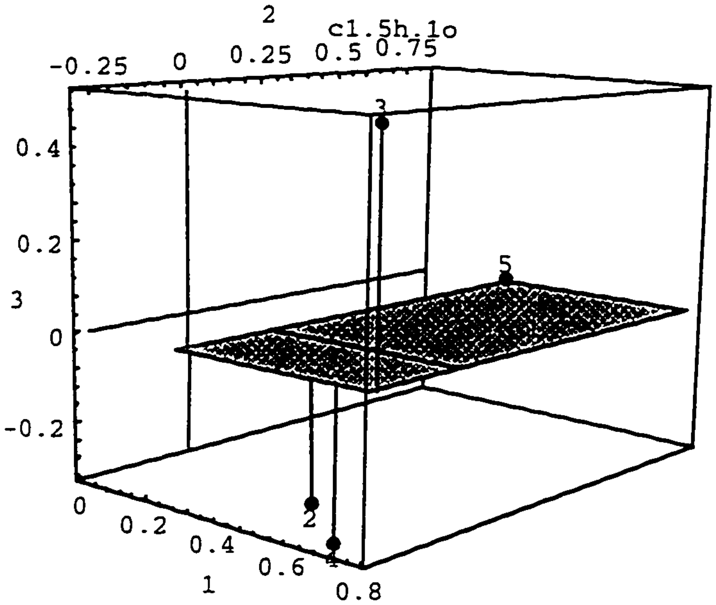


Figure 43. First 3 principal components of hidden responses for unconstrained network.

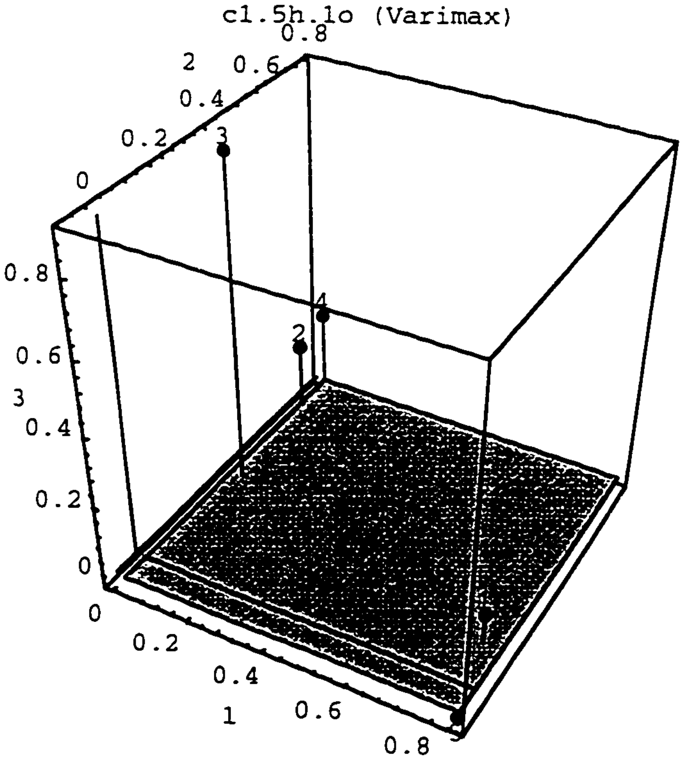


Figure 44. Varimax rotated principal components of hidden responses for unconstrained network.

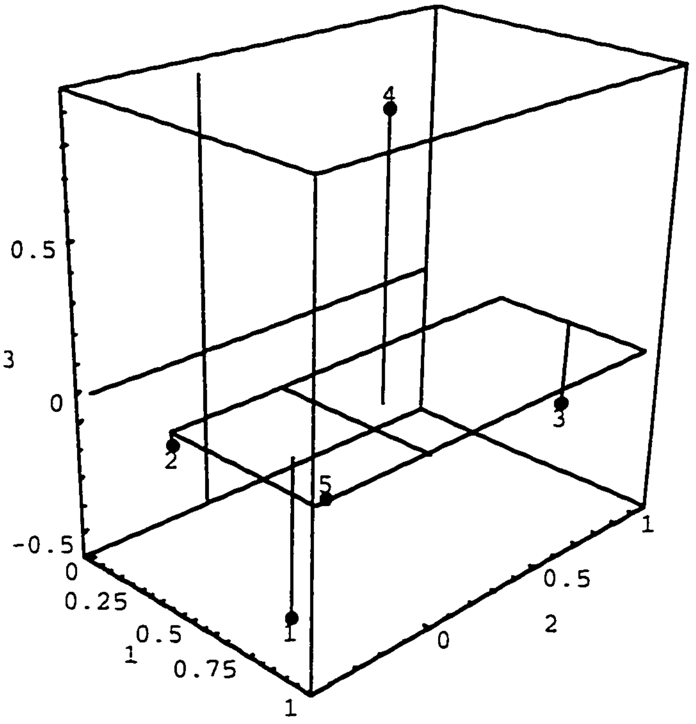


Figure 45. First 3 principal components of hidden responses for constrained network.

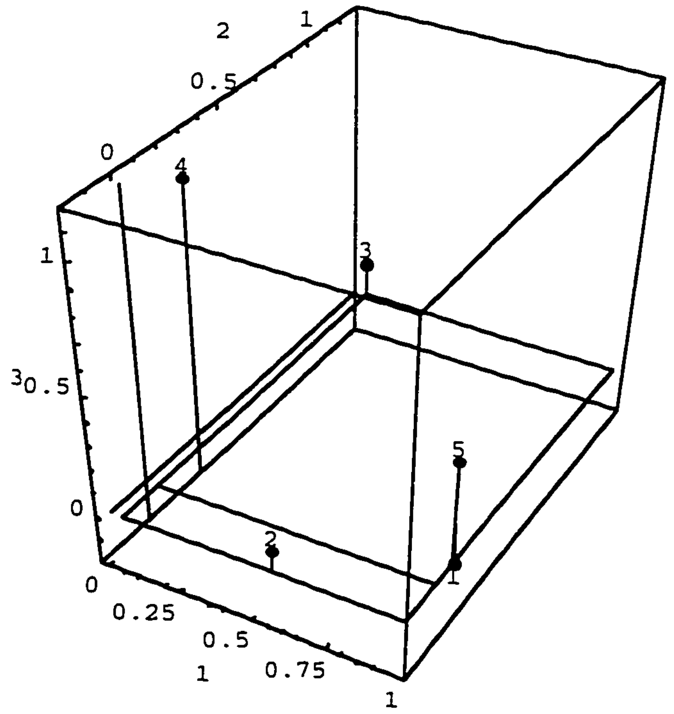


Figure 46. Varimax rotated principal components of hidden responses for constrained network.

4.5 Discussion and conclusion

This chapter began by alluding to the difficulties of substantiating an isomorphic relationship between observed behavior and the connectionist network used to model these behaviors. The difficulty in establishing this isomorphic relationship is based on three assumptions: (a) the physical structure of the model, (b) the learning mechanisms and the representations these mechanisms construct, and (c) the environmental input. Understanding any of these three assumptions is an exploratory process. The work in this Chapter offers some possibilities of understanding network representation from both a dynamic and final state perspective. This investigation is grounded in the task being modeled, in this case the Bruner et al.'s (1956) concept attainment task. A major outcome of modeling this task proved to be the structural difference between constrained and unconstrained networks. Thus the analysis presented in this chapter centered on the internal representational differences between the two structurally different models.

The environmental conditions, in this case the input exemplars and the presentation of these exemplars, closely resemble their respective real world counterparts. From a dynamic perspective it appears the network learned faster with constraints because the constraints helped it to separate the exemplars and construct a better internal categorization process. Analysis of both hidden and output cellular responses support this conclusion. The hidden layer also developed its own unique response rule-space in which to classify the exemplars. The 3-D plots for dynamic learning demonstrate the highly complex nonlinear nature of rule formation. These plots also show that constrained networks are better able to clearly distinguish between negative and positive instances and define a sounder conceptual definition. This supported empirical results that demonstrated constrained networks performed better on generalization tasks.

It appears that the rules at the hidden layer for constrained networks show a much clearer separation of the exemplars than is the case for the unconstrained networks. Thus it was not the case that both networks converged on the same set of rules at the hidden layer even though the initial starting weights were the same for both networks. Both networks reached convergence at different points in time but more importantly it appears the constrained network did so because it better developed a representational model which reflected the association between instance values and the attributes used to define these instance values. This separation was clearly visible from the singular value decomposition of hidden layer responses that showed the constrained network had greater discriminatory power between exemplars without any loss to the positive concept. More importantly this supports the argument in the previous chapter that extra output cells (attribute context constraints), having a functional relationship to the input, can direct the network to form a better internal representation (see section 3.5). In other words the rule space developed for the constrained network produced a functional approximation (a global minima) that better reflected the concept attainment task.

Overall the analysis presented in this chapter focused entirely on the cellular responses given at both the hidden and output layers. This type of analysis is only one of many possible interpretations and based on supporting the use of attribute context constraints for modeling concept attainment tasks. The area that is not discussed is the analysis of the weight space. As mentioned previously, the larger size of the input array produced a hidden-layer weight matrix which proved difficult to analyze and presents a definite area for future research.

4.6 References

Berkeley, I. S. N., Dawson, M. D. R., Medler, D. A., Schopflocher, D. P., & Hornsby, L. (1995). Density plots of hidden value unit activations reveal interpretable bands. Connection Science, 7(2), 167-186.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). A study of thinking. New York: John Wiley & Sons.

Dawson, M. R. W., & Schopflocher, D. P. (1992). Modifying the generalized delta rule to train networks on nonmonotonic processors for pattern classification. Connection Science, 4, 19-31.

Gagne, R. (1962). The acquisition of knowledge. Psychological Review, 69(4), 355-365.

Gagne, R. M. (1984). Learning outcomes and their effects. American Psychologist, 39(4), 377-385.

Gagne, R. M. (1985). The conditions of learning and theory of instruction (4th ed.). Toronto, ON: Holt, Rinehart and Winston.

Hanson, S. J., & Burr, D. J. (1990). What connectionist models learn: Learning and representation in connectionist networks. Behavioral and Brain Sciences, 13, 471-518.

Hinton, G. E. (1986). Learning distributed representations of concepts. In Proceedings of the 8th Annual Metting of the Cognitive Science Society, (pp. 1-12). Hillsdale, N. J. : Lawrence Erlbaum Associates.

Hunka, S., & Carbonaro, M. (to appear). Understanding the back propagated neural network using Mathematica Graphics. Mathematica Education and Research.

Jackson, P. J. (1990). Introduction to expert systems (Second ed.). Don Mills, ON: Addison Wesley.

Lippmann, R. P. (1987, April). An introduction to computing with neural nets. IEEE ASSP Magazine, p. 4-22.

McClelland, J. L., & Jenkins, E. (1991). Nature, nurture and connections: Implications of connectionist models for cognitive development. In K. VanLehn (Eds.), Architectures for intelligence: Twenty-second Carnegie symposium on cognition (pp. 41-73). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

McClelland, J. L., & Rumelhart, D. E. (1989). Explorations in parallel distributed processing: A handbook of models, programs and exercises. Cambridge, MA: MIT Press.

Minsky, M. L., & Papert, S. A. (1969). Perceptrons: An introduction to computational geometry. Cambridge, MA: MIT Press.

Robinson, D. A. (1992). Implications of neural networks for how we think about brain function. Behavioral and Brain Sciences, 15, 644-655.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumlehart & J. L. McClelland (Eds.), Parallel distributed processing: Explorations in the microstructure of cognition Cambridge, MA: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (Ed.). (1986). Parallel distributed processing: Explorations in the microstructure of cognition. Cambridge, MA: MIT Press.

Shultz, T. R., & Elman, J. L. (1993). Analyzing cross connected networks. In J. D. Cowan, G. Tesauro, & J. Alspector (Eds.), Advances in Neural Information Processing Systems 6 (pp. 1117-1124). San Francisco, CA: Morgan Kaufmann.

Chapter 5

5. Overview

This closing chapter is meant to tie together the three major pieces of work presented in this thesis. The work in Chapter 2 was designed to show the implications computational modeling can have for work in the field of Education. Chapter 3 re-examined Bruner, Goodnow and Austin's (1956) classic research on concept attainment from the connectionist perspective. In Chapter 4 the research focused on developing interpretations of the connectionist models presented in Chapter 3. The obvious link between the work in these chapters is the use of computational modeling as a vehicle for furthering our understanding of the cognitive process under investigation.

Essentially computational modeling offers the opportunity to investigate cognitive phenomena not available in the more traditional forms of quantitative and qualitative research. The development of computational cognitive modeling has evolved during the past 40 years out of the advances made in computer technology, psychology, and neuroscience. Two general forms of computational modeling, symbolic artificial intelligence and connectionism, were discussed in Chapter 2 and it was argued that connectionist modeling is the more reductionist of the two. In this sense, connectionist modeling appeals to those researchers who would like to constrain their models with information provided by neuroscience. The discussion, at the conclusion Chapter 2, of the connectionist model built to study Inhelder and Piaget's (1958) balance scale task, is a good example of how connectionism can be used to investigate a cognitive phenomenon.

In Chapter 3 Bruner et al.'s (1956) research concept attainment was selected for connectionist modeling because of the influence this work has had on the area of learning theory. In particular, Bruner's pedagogical strategies of spiral curriculum and discovery learning have had a profound effect on the field of education (Schunk, 1996). More recently it has been pointed out that Bruner's (1985) theory of discovery learning (derived from his early research) is very similar to the theories of constructivism and situated cognition (Brown, Collins, & Duguid, 1989). In all these theories the learner discovers concepts, builds representations of the world, and implicity invents rules. The research presented in Chapter 3 used connectionist modeling to provide further insight into the Bruner et al. (1956) task. Results indicate that the learning of attributes plays a significant role in affecting the speed at which concepts can be attained and the degree to which these learned concepts can be generalized. To support this claim the argument is made that attribute context constraints (expressed in the form of extra output units) can guide the network to construct a more robust conceptual representation. Empirical results substantiated this argument.

Chapter 4 explored the network's representational "response space" by comparing the responses produced by two structurally different connectionist models. Both of these models were derived from the research work done in Chapter 3. The results from this chapter demonstrated that a response space forms a unique set of rules and that these rules reflect the importance of how additional constraints at the output level can significantly improve the performance of the connectionist model with respect to the task being investigated.

5.1 Final thoughts

In essence the work described in thesis, while emphasizing the important implications computational modeling can have for researching cognitive processes, demonstrates the difficulties of modeling and understanding a complex process such as concept attainment. For example, the work of Bruner et al. (1956) is multifaceted and results discussed in this thesis only provide an initial step toward enlightening our understanding of their work through the use of connectionist modeling.

In a recent book entitled Rethinking Innateness: A Connectionist Perspective on Development (Elman, Bates, Johnson, Karmiloff-Smith, Parisi, & Plunkett, 1996) the authors express three reasons why connectionist modeling is important. First, computer models of this type impose a degree of precision which can reveal logical discrepancies in the theory we are attempting to model. A verbal description of the hypotheses lacks the rigorous definition obtained in a connectionist model. Second, the non-linear nature and distributed representations offered by connectionist models can provide empirical results, regarding the hypotheses that would otherwise be unattainable. Third, such models construct representations at a level of detail that is normally not available for investigation. Such accessibility is not usually possible when studying humans. According to Elman et al. (1996), "with humans, we can usually only guess at the nature of the mechanism responsible for a behavior by inference, but with the simulation we can directly inspect the network in order to understand the solution" (p. 45). In the final analysis, computational modeling affords researchers of cognition the opportunity to move beyond the limits defined by qualitative and quantitative procedures.

5.2 References

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. Educational Researcher, 18, 32-42.

Bruner, J. S. (1985). Models of learner. Educational Researcher, 14, 5-8.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). A study of thinking. New York: John Wiley & Sons.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, Parisi, D., & Plunkett, K. (1996). Rethinking innateness: A connectionist perspective on development. Cambridge, MA: MIT Press.

Inhelder, B., & Piaget, J. (1958). The growth of logical thinking from childhood to adolescence. New York: Basic Books.

Schunk, D. H. (1996). Learning theories: An educational perspective (2nd ed.). Toronto: Prentice Hall.