THE UNIVERSITY OF CALGARY

Connectionist Network Models of Attention in Human Learning

by

John Begoray

A DISSERTATION

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY

CALGARY, ALBERTA

SEPTEMBER, 1993

ISBN 0-315-93859-5

Canada

Name JOHN BEGOLAY

Dissertation Abstracts International is arranged by broad, general subject categories. Please select the one subject which most nearly describes the content of your dissertation. Enter the corresponding four-digit code in the spaces provided.

COGNITIVE SCIENCE — SUBJECT TERM

0525 — SUBJECT CODE

U·M·I

Subject Categories

THE HUMANITIES AND SOCIAL SCIENCES

COMMUNICATIONS AND THE ARTS
Architecture 0729
Art History 0377
Cinema 0900
Dance 0378
Fine Arts 0357
Information Science 0723
Journalism 0391
Library Science 0399
Mass Communications 0708
Music 0413
Speech Communication 0459
Theater 0465

EDUCATION
General 0515
Administration 0514
Adult and Continuing 0516
Agricultural 0517
Art 0273
Bilingual and Multicultural 0282
Business 0688
Community College 0275
Curriculum and Instruction 0727
Early Childhood 0518
Elementary 0524
Finance 0277
Guidance and Counseling 0519
Health 0680
Higher 0745
History of 0520
Home Economics 0278
Industrial 0521
Language and Literature 0279
Mathematics 0280
Music 0522
Philosophy of 0998
Physical 0523
Psychology 0525
Reading 0535
Religious 0527
Sciences 0714
Secondary 0533
Social Sciences 0534
Sociology of 0340
Special 0529
Teacher Training 0530
Technology 0710
Tests and Measurements 0288
Vocational 0747

LANGUAGE, LITERATURE AND LINGUISTICS
Language
- General 0679
- Ancient 0289
- Linguistics 0290
- Modern 0291

Literature
- General 0401
- Classical 0294
- Comparative 0295
- Medieval 0297
- Modern 0298
- African 0316
- American 0591
- Asian 0305
- Canadian (English) 0352
- Canadian (French) 0355
- English 0593
- Germanic 0311
- Latin American 0312
- Middle Eastern 0315
- Romance 0313
- Slavic and East European 0314

PHILOSOPHY, RELIGION AND THEOLOGY
Philosophy 0422
Religion
- General 0318
- Biblical Studies 0321
- Clergy 0319
- History of 0320
- Philosophy of 0322

Theology 0469

SOCIAL SCIENCES
American Studies 0323
Anthropology
- Archaeology 0324
- Cultural 0326
- Physical 0327

Business Administration
- General 0310
- Accounting 0272
- Banking 0770
- Management 0454
- Marketing 0338

Canadian Studies 0385
Economics
- General 0501
- Agricultural 0503
- Commerce-Business 0505
- Finance 0508
- History 0509
- Labor 0510
- Theory 0511

Folklore 0358
Geography 0366
Gerontology 0351
History
- General 0578
- Ancient 0579
- Medieval 0581
- Modern 0582
- Black 0328
- African 0331
- Asia, Australia and Oceania 0332
- Canadian 0334
- European 0335
- Latin American 0336
- Middle Eastern 0333
- United States 0337

History of Science 0585
Law 0398
Political Science
- General 0615
- International Law and Relations 0616
- Public Administration 0617

Recreation 0814
Social Work 0452
Sociology
- General 0626
- Criminology and Penology 0627
- Demography 0938
- Ethnic and Racial Studies 0631
- Individual and Family Studies 0628
- Industrial and Labor Relations 0629
- Public and Social Welfare 0630
- Social Structure and Development 0700
- Theory and Methods 0344

Transportation 0709
Urban and Regional Planning 0999
Women's Studies 0453

THE SCIENCES AND ENGINEERING

BIOLOGICAL SCIENCES
Agriculture
- General 0473
- Agronomy 0285
- Animal Culture and Nutrition 0475
- Animal Pathology 0476
- Food Science and Technology 0359
- Forestry and Wildlife 0478
- Plant Culture 0479
- Plant Pathology 0480
- Plant Physiology 0817
- Range Management 0777
- Wood Technology 0746

Biology
- General 0306
- Anatomy 0287
- Biostatistics 0308
- Botany 0309
- Cell 0379
- Ecology 0329
- Entomology 0353
- Genetics 0369
- Limnology 0793
- Microbiology 0410
- Molecular 0307
- Neuroscience 0317
- Oceanography 0416
- Physiology 0433
- Radiation 0821
- Veterinary Science 0778
- Zoology 0472

Biophysics
- General 0786
- Medical 0760

EARTH SCIENCES
Biogeochemistry 0425
Geochemistry 0996
Geodesy 0370
Geology 0372
Geophysics 0373
Hydrology 0388
Mineralogy 0411
Paleobotany 0345
Paleoecology 0426
Paleontology 0418
Paleozoology 0985
Palynology 0427
Physical Geography 0368
Physical Oceanography 0415

HEALTH AND ENVIRONMENTAL SCIENCES
Environmental Sciences 0768
Health Sciences
- General 0566
- Audiology 0300
- Chemotherapy 0992
- Dentistry 0567
- Education 0350
- Hospital Management 0769
- Human Development 0758
- Immunology 0982
- Medicine and Surgery 0564
- Mental Health 0347
- Nursing 0569
- Nutrition 0570
- Obstetrics and Gynecology 0380
- Occupational Health and Therapy 0354
- Ophthalmology 0381
- Pathology 0571
- Pharmacology 0419
- Pharmacy 0572
- Physical Therapy 0382
- Public Health 0573
- Radiology 0574
- Recreation 0575
- Speech Pathology 0460
- Toxicology 0383

Home Economics 0386

PHYSICAL SCIENCES
Pure Sciences
Chemistry
- General 0485
- Agricultural 0749
- Analytical 0486
- Biochemistry 0487
- Inorganic 0488
- Nuclear 0738
- Organic 0490
- Pharmaceutical 0491
- Physical 0494
- Polymer 0495
- Radiation 0754

Mathematics 0405
Physics
- General 0605
- Acoustics 0986
- Astronomy and Astrophysics 0606
- Atmospheric Science 0608
- Atomic 0748
- Electronics and Electricity 0607
- Elementary Particles and High Energy 0798
- Fluid and Plasma 0759
- Molecular 0609
- Nuclear 0610
- Optics 0752
- Radiation 0756
- Solid State 0611

Statistics 0463

Applied Sciences
Applied Mechanics 0346
Computer Science 0984
Engineering
- General 0537
- Aerospace 0538
- Agricultural 0539
- Automotive 0540
- Biomedical 0541
- Chemical 0542
- Civil 0543
- Electronics and Electrical 0544
- Heat and Thermodynamics 0348
- Hydraulic 0545
- Industrial 0546
- Marine 0547
- Materials Science 0794
- Mechanical 0548
- Metallurgy 0743
- Mining 0551
- Nuclear 0552
- Packaging 0549
- Petroleum 0765
- Sanitary and Municipal 0554
- System Science 0790

Geotechnology 0428
Operations Research 0796
Plastics Technology 0795
Textile Technology 0994

PSYCHOLOGY
General 0621
Behavioral 0384
Clinical 0622
Developmental 0620
Experimental 0623
Industrial 0624
Personality 0625
Physiological 0989
Psychobiology 0349
Psychometrics 0632
Social 0451

Nom _____

Dissertation Abstracts International est organisé en catégories de sujets. Veuillez s.v.p. choisir le sujet qui décrit le mieux votre thèse et inscrivez le code numérique approprié dans l'espace réservé ci-dessous.

_____ ☐☐☐ U·M·I

SUJET CODE DE SUJET

Catégories par sujets

HUMANITÉS ET SCIENCES SOCIALES

COMMUNICATIONS ET LES ARTS
Architecture 0729
Beaux-arts 0357
Bibliothéconomie 0399
Cinéma 0900
Communication verbale 0459
Communications 0708
Danse 0378
Histoire de l'art 0377
Journalisme 0391
Musique 0413
Sciences de l'information 0723
Théâtre 0465

ÉDUCATION
Généralités 515
Administration 0514
Art 0273
Collèges communautaires 0275
Commerce 0688
Économie domestique 0278
Éducation permanente 0516
Éducation préscolaire 0518
Éducation sanitaire 0680
Enseignement agricole 0517
Enseignement bilingue et multiculturel 0282
Enseignement industriel 0521
Enseignement primaire. 0524
Enseignement professionnel 0747
Enseignement religieux 0527
Enseignement secondaire 0533
Enseignement spécial 0529
Enseignement supérieur 0745
Évaluation 0288
Finances 0277
Formation des enseignants 0530
Histoire de l'éducation 0520
Langues et littérature 0279
Lecture 0535
Mathématiques 0280
Musique 0522
Orientation et consultation 0519
Philosophie de l'éducation 0998
Physique 0523
Programmes d'études et enseignement 0727
Psychologie 0525
Sciences 0714
Sciences sociales 0534
Sociologie de l'éducation 0340
Technologie 0710

LANGUE, LITTÉRATURE ET LINGUISTIQUE
Langues
- Généralités 0679
- Anciennes 0289
- Linguistique 0290
- Modernes 0291

Littérature
- Généralités 0401
- Anciennes 0294
- Comparée 0295
- Médiévale 0297
- Moderne 0298
- Africaine 0316
- Américaine 0591
- Anglaise 0593
- Asiatique 0305
- Canadienne (Anglaise) 0352
- Canadienne (Française) 0355
- Germanique 0311
- Latino-américaine 0312
- Moyen-orientale 0315
- Romane 0313
- Slave et est-européenne 0314

PHILOSOPHIE, RELIGION ET THÉOLOGIE
Philosophie 0422
Religion
- Généralités 0318
- Clergé 0319
- Études bibliques 0321
- Histoire des religions 0320
- Philosophie de la religion 0322

Théologie 0469

SCIENCES SOCIALES
Anthropologie
- Archéologie 0324
- Culturelle 0326
- Physique 0327

Droit 0398
Économie
- Généralités 0501
- Commerce-Affaires 0505
- Économie agricole 0503
- Économie du travail 0510
- Finances 0508
- Histoire 0509
- Théorie 0511

Études américaines 0323
Études canadiennes 0385
Études féministes 0453
Folklore 0358
Géographie 0366
Gérontologie 0351
Gestion des affaires
- Généralités 0310
- Administration 0454
- Banques 0770
- Comptabilité 0272
- Marketing 0338

Histoire
- Histoire générale 0578
- Ancienne 0579
- Médiévale 0581
- Moderne 0582
- Histoire des noirs 0328
- Africaine 0331
- Canadienne 0334
- États-Unis 0337
- Européenne 0335
- Moyen-orientale 0333
- Latino-américaine 0336
- Asie, Australie et Océanie 0332

Histoire des sciences 0585
Loisirs 0814
Planification urbaine et régionale 0999
Science politique
- Généralités 0615
- Administration publique 0617
- Droit et relations internationales 0616

Sociologie
- Généralités 0626
- Aide et bien-àtre social 0630
- Criminologie et établissements pénitentiaires 0627
- Démographie 0938
- Études de l' individu et de la famille 0628
- Études des relations interethniques et des relations raciales 0631
- Structure et développement social 0700
- Théorie et méthodes. 0344
- Travail et relations industrielles 0629

Transports 0709
Travail social 0452

SCIENCES ET INGÉNIERIE

SCIENCES BIOLOGIQUES
Agriculture
- Généralités 0473
- Agronomie. 0285
- Alimentation et technologie alimentaire 0359
- Culture 0479
- Élevage et alimentation 0475
- Exploitation des pâturages 0777
- Pathologie animale 0476
- Pathologie végétale 0480
- Physiologie végétale 0817
- Sylviculture et faune 0478
- Technologie du bois 0746

Biologie
- Généralités 0306
- Anatomie 0287
- Biologie (Statistiques) 0308
- Biologie moléculaire 0307
- Botanique 0309
- Cellule 0379
- Écologie 0329
- Entomologie 0353
- Génétique 0369
- Limnologie 0793
- Microbiologie 0410
- Neurologie 0317
- Océanographie 0416
- Physiologie 0433
- Radiation 0821
- Science vétérinaire 0778
- Zoologie 0472

Biophysique
- Généralités 0786
- Medicale 0760

SCIENCES DE LA TERRE
Biogéochimie 0425
Géochimie 0996
Géodésie 0370
Géographie physique 0368
Géologie 0372
Géophysique 0373
Hydrologie 0388
Minéralogie 0411
Océanographie physique 0415
Paléobotanique 0345
Paléoécologie 0426
Paléontologie 0418
Paléozoologie 0985
Palynologie 0427

SCIENCES DE LA SANTÉ ET DE L'ENVIRONNEMENT
Économie domestique 0386
Sciences de l'environnement 0768
Sciences de la santé
- Généralités 0566
- Administration des hipitaux 0769
- Alimentation et nutrition 0570
- Audiologie 0300
- Chimiothérapie 0992
- Dentisterie 0567
- Développement humain 0758
- Enseignement 0350
- Immunologie 0982
- Loisirs 0575
- Médecine du travail et thérapie 0354
- Médecine et chirurgie 0564
- Obstétrique et gynécologie 0380
- Ophtalmologie 0381
- Orthophonie 0460
- Pathologie 0571
- Pharmacie 0572
- Pharmacologie 0419
- Physiothérapie 0382
- Radiologie 0574
- Santé mentale 0347
- Santé publique 0573
- Soins infirmiers 0569
- Toxicologie 0383

SCIENCES PHYSIQUES
Sciences Pures

Chimie
- Généralités 0485
- Biochimie 487
- Chimie agricole 0749
- Chimie analytique 0486
- Chimie minérale 0488
- Chimie nucléaire 0738
- Chimie organique 0490
- Chimie pharmaceutique 0491
- Physique 0494
- PolymÇres 0495
- Radiation 0754

Mathématiques 0405
Physique
- Généralités 0605
- Acoustique 0986
- Astronomie et astrophysique 0606
- Electronique et électricité 0607
- Fluides et plasma 0759
- Météorologie 0608
- Optique 0752
- Particules (Physique nucléaire) 0798
- Physique atomique 0748
- Physique de l'état solide 0611
- Physique moléculaire 0609
- Physique nucléaire 0610
- Radiation 0756

Statistiques 0463

Sciences Appliqués Et Technologie

Informatique 0984
Ingénierie
- Généralités 0537
- Agricole 0539
- Automobile 0540
- Biomédicale 0541
- Chaleur et ther modynamique 0348
- Conditionnement (Emballage) 0549
- Génie aérospatial 0538
- Génie chimique 0542
- Génie civil 0543
- Génie électronique et électrique 0544
- Génie industriel 0546
- Génie mécanique 0548
- Génie nucléaire 0552
- Ingénierie des systèmes 0790
- Mécanique navale 0547
- Métallurgie 0743
- Science des matériaux 0794
- Technique du pétrole 0765
- Technique minière 0551
- Techniques sanitaires et municipales 0554
- Technologie hydraulique 0545

Mécanique appliquée 0346
Géotechnologie 0428
Matières plastiques (Technologie) 0795
Recherche opérationnelle 0796
Textiles et tissus (Technologie) 0794

PSYCHOLOGIE
Généralités 0621
Personnalité 0625
Psychobiologie 0349
Psychologie clinique 0622
Psychologie du comportement 0384
Psychologie du développement 0620
Psychologie expérimentale 0623
Psychologie industrielle 0624
Psychologie physiologique 0989
Psychologie sociale 0451
Psychométrie 0632

THE UNIVERSITY OF CALGARY

FACULTY OF GRADUATE STUDIES

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies for acceptance, a dissertation entitled "Connectionist Network Models of Attention in Human Learning" submitted by John Begoray in partial fulfilment of the requirements for the degree of Doctor of Philosophy:

John Mueller

Supervisor, Dr. J. Mueller,
Educational Psychology

Dr. J. Eggermont, Psychology

Dr. W. Hunter, Teaching and Supervision

Mildred L. G. Shaw

Dr. M. Shaw, Computing Science

A. Marini

Dr. A. Marini, Teaching and Supervision

S. Hunka

External Examiner, Dr. S. Hunka
Educational Psychology, University of Alberta

Jan. 25, 1994

Date

ABSTRACT

Computer-based connectionist networks are seeing increasing use as models of human cognitive processes. Because these processes are complex, a multi-layer network is required. Because humans learn, the network model must also learn. The most common learning rule for a multi-layer network is back-propagation. The variability of the output signal produced by the back-propagation learning rule is usually fixed. If, instead, a parameter is added which gradually decreases the variability, the network should learn faster. The results of a simulation based on a network with such a parameter confirmed this.

When parameters are added to a network to improve performance, the resulting network is useful as a model of human cognition only if the parameter itself models some aspect of the human performance. The literature pertaining to human cognition within the context of a limited attentional resource suggests that attention declines as learning increases. A human learning experiment, based on a paired-associate learning task, confirmed the decline in attention and suggested several possible mathematical descriptions of the change with a view to modelling this change by gradually decreasing the variability of output signals in a back-propagation connectionist network.

A series of simulations was implemented to determine which of the possibilities suggested by the human learning experiment represented the best model of the human learning situation. The simulation results suggested that the fastest learning performance is achieved with a linear decreasing function, but a more faithful model of human learning is achieved with a sigmoidal function.

Acknowledgements

I would like to thank the members of my committee, listed here in close approximation of the chronological order in which they became involved in this project:

Dr. M. Shaw for exposing me to the difficulties associated with computer-based knowledge representation and for giving me the opportunity to discover connectionism.

Dr. J. Eggermont for helping me understand the implementation details of connectionist networks and discover the possibility of introducing a temperature parameter.

Dr. W. Hunter for invaluable assistance at several important stages and for helping me maintain an educational perspective.

Dr. J Mueller for providing me with an opportunity to sort out the human cognition issues, and especially for shepherding me through an overly protracted, and sometimes difficult, process.

I would also like to thank the Calgary Public School Board and Lester B. Pearson High School for giving me an opportunity to collect data. I also appreciate the significant financial support provided by a Doctoral Fellowship through the Social Sciences and Humanities Research Council.

And, in the end, my warmest and most profuse thanks go to my wife Deborah for giving me the chance to do this at all.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

This dissertation addresses topics from a number of disciplines including Psychology, Education, and Computing Science. To assist readers with a less technical background in some of these areas, this preface will present an informal overview of the entire research project. The more formal dissertation will begin in the next chapter.

When I first began to study "cognition" in the early seventies it wasn't called that, and behaviourism was beginning to give way to associationism. The main-stream of psychological research was concerned with the cause and effect of cognition, not the mechanism.

When I returned to the study of cognition in the late eighties, the current model which seemed to best account for the complexities of behaviour in general and learning in particular was a cognitive model -- the kind with short-term memory, long-term memory, and a central executive. Some attention was now being focused on internal representations and processing mechanisms, especially with semantic networks.

This perspective seemed very mechanistic and overly complicated, and it had some problems with some specific areas of cognition in which I was particularly interested. For example, localized representation: the notion that one location in the net represented one thought or idea (bird,

canary, yellow, etc.), didn't seem right in light of evidence which suggested a more distributed representation. Further, the models seemed to rely on some, ill-defined central control mechanism to process input and activate the node representing a conscious thought. And furthermore, the control process seemed to rely on rules, while human behaviour mostly does not.

Perhaps the components of these models were not to be taken so literally. The old Pandemonium model didn't mean there were actual demons yelling in the head and semantic network models probably didn't mean there were 'yellow', 'canary', and 'bird' neurons in the brain either. Perhaps rules just described behaviour instead of controlling it. But if this was the case, there should be a deeper, more fundamental level where the components of the model were neurologically plausible. And if this was the case, the higher, more abstract levels were largely allegorical, and perhaps less useful once a deeper understanding was achieved.

At least part of the appeal of the more mechanistic, semantic network models is that the model (or parts of it) can be implemented using fairly straight-forward computer programming techniques. I was interested in computer-based models of cognition and these models seemed interesting at least in part because they were implementable.

Shortly after this, I encountered connectionist models of cognition. These models specified a distributed representation of knowledge and no specific processing mechanism -- or, more exactly, an automatic, distributed mechanism.

Computer-based models of connectionist representations were anything but straight-forward. Mostly they relied on the stochastic nature of non-linear equations to respond to specific inputs with a non-deterministic output. Parallel Distributed Processing (PDP) seemed to be the most appropriate implementation technology for connectionist models of human cognition because of its sophisticated learning rules involving back-propagation.

One of the terms in the set of non-linear equations which defined back-propagation did not seem to have a theoretical justification. It was just fixed at a value which worked "best" for any given implementation. The effect of this term on processing in a PDP network seemed to be similar to that of the temperature term in a Boltzman machine in that when it was high, learning proceeded rapidly and was characterized by a high level of activity and by "divergent" processing but the error rate was also high. When this "temperature" term was set low, learning was slower but proceeded in a more regular fashion and the error rate was low.

It seemed to me that optimal learning would occur if the temperature was high when learning began and low as mastery was approached. I developed a simulation of a learning task based on a connectionist network using a back-propagation learning rule where I gradually decreased the temperature at a fixed rate as learning progressed.

The simulation with declining temperature was able to learn significantly faster than simulations with a fixed temperature, but I was more interested in modeling human cognition than developing a fast-learning neural network. A temperature factor would be useful in a connectionist model of human cognition only if there were an equivalent factor in human learning: something which started high then declined during a novel learning situation, but was relatively low once the material was mastered.

I liked one part of the more mechanistic cognitive models -- the idea of limited attentional resources. But without the special purpose processing mechanisms of the cognitive models, how could a connectionist model represent these effects? The behaviour of the simulation with the declining temperature term seemed to have some of the characteristics of human attention as it relates to the idea of a limited attentional resource. By adding a representation of attention, I hoped to show that a simulation with a variable temperature term was a "better" model of human cognition than one without such a term. To

do this I needed to compare human performance on a learning task with the simulation's performance on a similar task.

Human attention certainly seems to decline as a student masters knowledge and skills, but what is the rate of decline and what factors influence it? The initial simulation reduced temperature at a fixed rate and by an arbitrary amount. This first approximation showed the benefit of a temperature term but it didn't seem likely that the relationship between learning and attention was that simple. More likely, attention declines in some non-linear way as a function of either amount of learning or elapsed time or some even more complex interaction of both.

To get a more accurate picture of just how human attention decreases, I conducted a human learning experiment which measured attention and learning over a fixed number of trials of a simple paired-associate learning task. I hoped the results would provide a mathematical description of the change in attention versus learning and/or time. I could then use this mathematical description to vary the temperature term in a connectionist network to provide a more accurate model of human learning.

The results of the human experiment definitely showed a decline in attention as learning proceeded and as time passed. However, the exact mathematical relationship between these was not entirely clear. In the end, I developed seven network simulations: the conventional PDP

network with fixed temperature and a two-by-three matrix of simulations based on either learning or time and using one of three functions (linear, quadratic, and sigmoidal) to determine temperature.

In terms of network performance, the original simulation which decreased temperature in a linear fashion with time was the clear winner. Learning was faster and more consistent than in any of the other simulations. However, maximizing network performance was not the objective. Finding a more appropriate way to model human learning was.

No one claims that human learning is either fast or consistent. There are many circumstances where computer performance is much faster. So, almost by definition, the computer-based network with the best performance is not likely to be the best model of human learning. And, in fact, this research suggests that the best way to model human learning in a connectionist network is to decrease temperature sigmoidally as learning increases.

CHAPTER ONE: CONNECTIONISM

The term "connectionism" has been applied to various aspects of human cognition for nearly a century and, although still not entirely accepted as a major theory of human cognition, connectionism has gained considerable support over the last decade. As yet, there is no clearly accepted delineation of the connectionist realm. This chapter will build a specific perspective on the term connectionism and will establish a context for the rest of this dissertation.

Computer-based connectionist networks are used by researchers in both cognitive science and artificial intelligence. Although the implementation details of all of these networks may be similar, the objectives often are not. Most artificial intelligence research tries to develop computer-based solutions to the kinds of problems which traditionally have been solved best by humans. Sometimes what is known about human cognition facilitates this development but the objective is to solve the problem in the best way possible. On the other hand, cognitive science research is more concerned with developing a computer application which faithfully replicates some aspect of human behavior. In general, artificial intelligence applications attempt to maximize the performance of a network while cognitive science research may sacrifice the absolute

performance of a network to more get a more human-like response. It is this later perspective which is most appropriate for this dissertation.

All connectionist models of cognition share a general set of features. Details of how these general features behave and interact place constraints on a model's *architecture*. A number of implications arise from the interaction of these general features within a model. The first three sections of this chapter will describe the general features of connectionist models, identify some architectural constraints, and discuss the implications these constraints hold for connectionist models.

Not all cognitive scientists accept connectionism as an appropriate psychological model. The fourth section identifies and discusses some of the criticisms leveled against connectionism and the last section discusses the relationship between connectionism and some of the more conventional psychological paradigms. The chapter will close with some conclusions about the use of connectionist models to explore human cognition.

Basic Common Features

All connectionist models have, as their basis, a network of interconnected *nodes* or *neurons*. One such node is illustrated in Figure 1.1. Each node has one or more *input* connections which come either from the external

environment or from other nodes in the network. Each node also has a single *output* connection, but that output can branch to send its signal to more than one other node or to the outside environment. These connections carry *signals* from node to node in one direction only. Each connection has a *weight* associated with it. This weight modifies the strength of any signal passing through the connection.

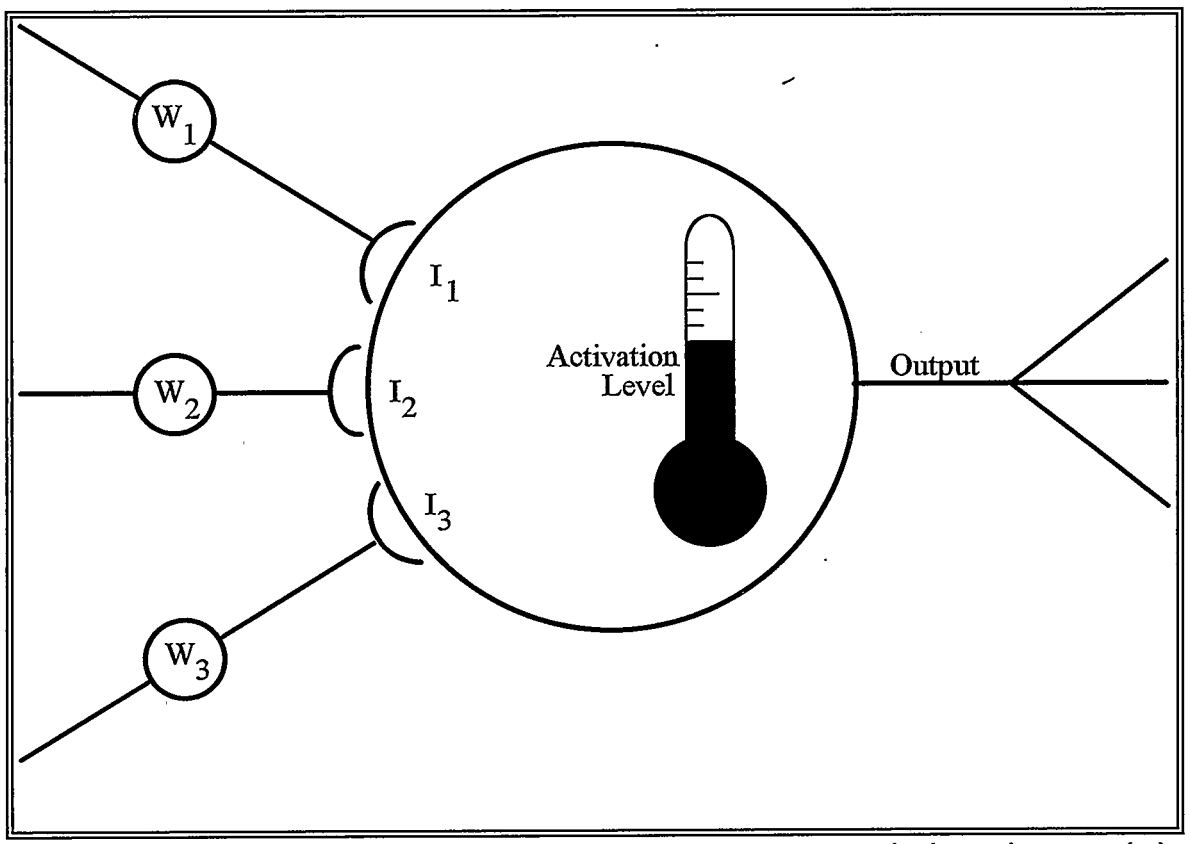

Figure 1.1 A connectionist network node receiving input (I) modified by weights (W) on connections from three other nodes and producing output which branches to three other nodes.

Each node has an *activation level* that changes with time. An *activation function* determines this activation

level based in part on the node's previous activation level and in part on the strengths of all of the weighted input signals the node receives. All of the nodes in a network use the same activation function to determine their activation levels.

The activation level, in turn, serves as the basis for an *output function* which determines the strength of the node's output signal. Again, all of the nodes in the network use the same output function.

The process of evaluating the weighted input signals, calculating a new activation level, and producing an output signal is an entirely local event analogous to the firing of a neuron in a biological neural network. There are two aspects of this calculation which are characteristic of connectionist networks. First, there is no central controlling mechanism -- all processing is done by the individual nodes and all nodes perform their processing at the same time (in parallel). Second, the transformation of input signal to output signal must be non-linear and, in fact, some would suggest sigmoidal (Kosko, 1987).

A connectionist network reacts to signals received from the environment. For convenience, these signals are usually represented as coming from an *input node* whose activation level is fixed at a specific value which represents the signal from the environment. A network will usually have more than one input node so that more than one environmental

stimulus can be represented. Instead of allocating one node to each stimulus, most networks represent each stimulus as a pattern of values across all input nodes. In this latter case, the input to the network is said to be *non-orthographic* since two or more different stimuli may present the same value to one or more of the input nodes.

To be useful, a network must produce a response to an environmental stimulus. To accomplish this, one or more nodes in a network are designated as *output nodes*. The signal coming from an output node (or the pattern of signals coming from all output nodes) represents a response from the network.

A network may consist solely of input and output nodes or it may include *hidden nodes* as well. Hidden nodes have no direct connection to the environment. Instead, they accept signals from other nodes, adjust their own activation levels accordingly, and send a commensurate output signal to other nodes.

The processing which occurs when a network responds to a stimulus might proceed as follows:

A stimulus is presented to the network by setting the activation level of all input nodes to a pattern of specific values and fixing them there. All of the input nodes then send out signals whose strength is determined by these fixed activation levels. These signals are sent along connections to other specific nodes. Each connection has a weight

attached to it and this weight modifies the strength of the signal passing through the connection.

Any one node receiving a signal from an input node will likely be receiving signals from other nodes as well. It combines all of these signals and uses the result to adjust its activation level. The node then emits its own signal based on its new activation level.

In this way, signals are propagated through the entire network until all output nodes are producing an output signal. The pattern of these signals across all output nodes is the network's response to the specific stimulus "presented" to the input nodes. If a second, different stimulus is now presented (by setting the input nodes to a new pattern of activation levels), the network will respond with a different response (pattern of signals) at the output nodes.

Architectural Constraints

The general features mentioned in the previous section are common to all connectionist models. Any one specific model, however, will implement these basic features in its own, unique way resulting in a specific network architecture. The details of this specific architecture will place constraints on the behavior of the specific network and, indirectly, on the connectionist model the network represents.

Interconnection and Layers

The amount of interconnection between nodes in a network is determined by the number of other nodes to which any one node may be connected. In a *fully* interconnected network each node is connected to all of the other nodes. In a *partially* interconnected network each node is connected to only some of the other nodes. Most connectionist networks, especially those with hidden nodes, are usually partially interconnected.

In a partially interconnected network, nodes are usually grouped into layers according to the nature of their interconnectedness and according to the function of the node (input, hidden, or output).

One of the oldest connectionist networks consisted of a single layer of input nodes and a single layer of output nodes, with all possible connections established between layers but not within layers. This type of network is very simple but it is quite good at modeling human perception, and is consequently often called a *perceptron* (Rosenblatt, 1962).

Although a perceptron is quite powerful in some situations, it is unable to address many important processing requirements (Minsky and Papert, 1969). More sophisticated models allow for a third layer of hidden nodes between the input and output nodes. Some architectures allow these hidden nodes to be connected to any combination

of input nodes, output nodes, and other hidden nodes. Other architectures restrict connections in some way. For example, in a bottom-up architecture with a bottom layer of input nodes and a top layer of output nodes, input nodes must be connected to either hidden or output nodes and hidden nodes cannot be connected down to input nodes (Rumelhart, Hinton, & Williams, 1986).

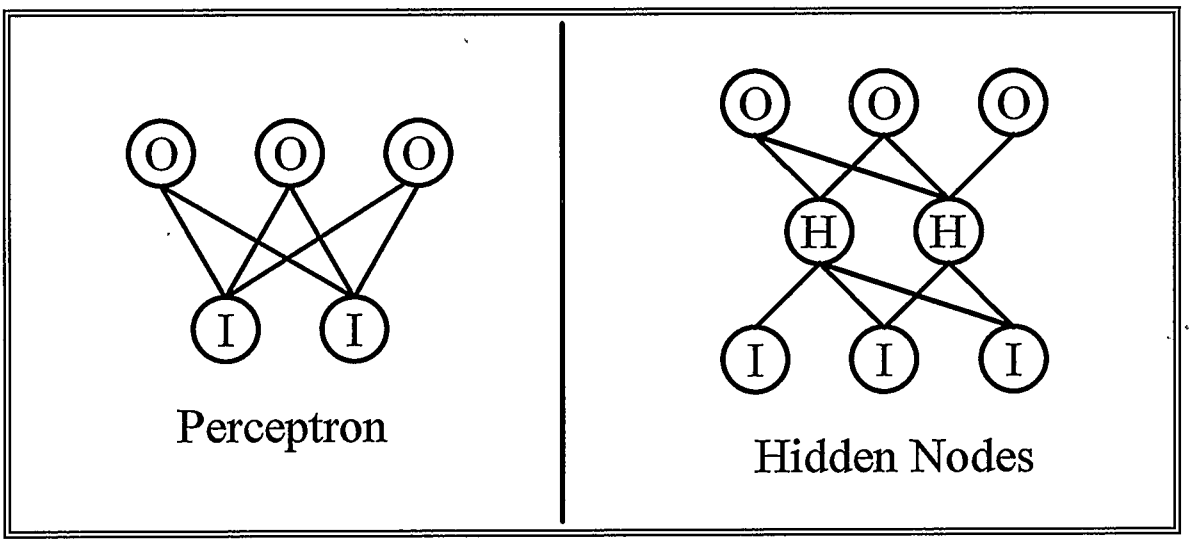

Figure 1.2 Examples of two network configurations showing different numbers of layers and different amounts of connectivity.

Signal Propagation

For most connectionist models, the update of signals within each node occurs discretely, rather than continuously (Obermeier & Barron, 1989). This means that all of the nodes in a network fire and then "rest" for a moment before firing again. In a network with no hidden nodes this may not seem important because the single firing will carry all of the input signals directly to the output nodes, but this

distinction has much more significance when hidden nodes are introduced.

If there is a layer of hidden nodes between the input nodes and the output nodes, a single firing of the network will not be sufficient for the activation levels of the input nodes to affect the output nodes. Instead, the input activation will spread only as far as the hidden nodes, and a second firing will be required before the final output is generated. In networks with multiple layers of hidden nodes (i.e., networks where hidden nodes can be connected to other hidden nodes) multiple firings will be required before the input activation has spread throughout the network and all nodes again come to rest in a stable state.

In a richly interconnected network, it is possible for a pattern of connections to exist such that the output signal from a node eventually comes back as an input signal for that same node. This is called a *feedback loop*. A signal entering such a loop could resonate indefinitely and prevent the network from ever reaching a stable state. To prevent this, Kaplan, Weaver, and French (1990) specify mechanisms of inhibition and fatigue to dampen activation between and within nodes respectively. The resulting architecture can "provide the system with the means of having internal, semi-autonomous, activatible representations of reality that do not rely uniquely on the sensory interface".

Training and Learning

The output pattern produced by a network in response to a specific input pattern depends on the activation level of the nodes in the network when the stimulus is presented and on the weights of the connections between nodes. The activation levels automatically change each time a new set of signals is propagated through the network, but the weights stay the same. It is the weights, therefore, which represent the fixed 'knowledge' which allows a network to produce the same (or almost the same) response any time it is presented with the same stimulus pattern.

To establish the connection weights which will allow a network to provide an appropriate response to each stimulus, a network must be trained. This can either be done manually by an outside agent or automatically by the network itself. If a network adjusts its own connection weights, the network is said to be able to learn. The function the network uses to make these adjustments is called the network's *learning rule*.

Hebbian Learning Rule

The basis for most connectionist network learning rules is the Hebbian learning rule "which holds that associations are built up between things that occur together" (Zeidenberg, 1987, p. 240). Any time two connected nodes have high levels of activation, this rule increases the

weight of the connection between them. In most cases, a more elaborate variant of this learning rule is used.

Delta Learning Rule

The application of this rule assumes a fully interconnected network with no hidden nodes. Initially the weights of all connections are set to small random values. The network is trained by repeatedly presenting training pairs consisting of both the input and the desired output. Each time the network receives the input it produces some output. For each output node, the delta rule calculates the difference between the actual and the expected output and adjusts the weight of the connection to each node accordingly. The amount of adjustment depends not only on how "wrong" the output node was but on the strength of the input to that node (Jones & Hoskins, 1987).

Back Propagation

For networks with hidden nodes, a more complicated learning rule is required. The difference between the actual and the expected output can only be used to update those nodes directly connected to the output nodes. However, the amount by which these nodes are updated can serve as a basis for updating the nodes to which they are connected. In this way the output error can be propagated back through the network all the way to the input nodes. This is much like the spread of activation forward through the network, except that it spreads back from the output

nodes to the input nodes and it is the connection weights not the activation levels that are adjusted. This back-propagation learning rule is somewhat restricted as well in that it only works for a bottom-up network topology (Rumelhart, Hinton, & Williams, 1986).

Back propagation is not an entirely new concept. Thorndike proposed a teaching rule by which the positive outcome strengthened connections between an immediately preceding behavior and stimulus input present at the time, the so-called "Law of Effect" (Walker, 1990, p. 25). This rule has much in common with back propagation.

Competitive Learning Rule

Networks with massively interconnected hidden nodes can learn without being specifically trained. One example of this involves the competitive learning rule. In such a network, clusters of hidden nodes are structured in such a way that: each node in the cluster is connected to all input nodes; the weights of the input connections are initially random; input connections to one node in the cluster inhibit similar connections to other nodes in the cluster; and a Hebbian learning rule is applied. When a competitive learning network stabilizes, each cluster will come to represent a general feature or characteristic of the input. If a similar layer of hidden nodes takes its input from the clusters in the first layer, this second layer will come to represent more complex features of the input.

> Given a threshold number of connections between a set of simple neurons, a form of self-organization takes place, and from this organization collective computational properties emerge, such as association, generalization, differentiation, preferential learning, optimization, and fault tolerance (Josin, 1987, p. 184).

Characteristics of Learning in Networks

A single connectionist network can be trained to make appropriate responses to several different stimului. The simplest case involves mutually exclusive stimuli where the signal strength at every input node is different for each stimulus. An extreme example of this would be where each stimulus sets a different input node to a high value. As long as there are at least as many input nodes as stimuli, the network receives unambiguous input and can learn to make consistent, reliable responses. Even if the different stimuli are represented as patterns of high values at several input nodes, the network will learn to differentiate between stimuli as long as each stimulus uses a different set of input nodes for its pattern.

On the other hand, it is possible to have redundancy in the input. This happens when several different stimuli set some of the same input nodes to similar values. Connectionist networks have some characteristic ways of responding to this kind of redundant input and the responses

are often similar to the way humans respond in similar circumstances. This section will discuss several of these characteristic responses:

Distributed Representation

Connectionist networks can resolve ambiguous input because they use a distributed representation. In contrast, semantic networks use local representations in that a single concept in such a network is represented in a single node. For example the concept "grandmother" would be represented by a single "grandmother" node. In a connectionist network, however, a single concept is represented by a unique pattern of connection weights distributed across the entire network (McClelland, 1988). Walker (1990) uses a piano analogy to describe distributed representation. The keys of the piano represent all of the nodes in the network. Any one sound coming from the piano is analogous to a single concept in the connectionist model. This sound is "represented" by the keys being pressed at any one instant. If a piano with 100 keys were played with one hand (5 keys pressed) there are 75 million possible sounds which could be produced.

In a connectionist network the proportion of nodes active (keys being played) at one time is more likely to be half of the nodes in the network which is usually considerably more than five. In a network with only one hundred nodes, fifty of them might have significant activation levels at one time. If five pianists were

playing the piano at the same time (10 hands = 50 keys), reporting the number of possible sounds would require a thirty digit number. While the piano only has a few dozen keys, the human brain has billions of neurons, even though the human brain is not fully interconnected, the resulting possible combinations of distributed representations should be more than sufficient to represent all of the concepts any one human mind could hold.

Automatic Generalization

In a distributed representation, similar concepts have similar patterns of activation or, to put it another way, if large areas of two patterns of activation are the same, those two patterns represent similar concepts. This feature of connectionist models leads to automatic generalization. Zeidenberg (1987) presents the following example of automatic generalization. The concepts "gorilla" and "chimp" are related. This means that many of the most highly activated nodes in the gorilla pattern are also highly activated in the chimp pattern. If the concept 'hairy' comes to be associated with gorilla, the weights between the highly activated nodes in both the hairy pattern and the gorilla pattern are increased or strengthened. Since these nodes in the gorilla pattern are mostly the same nodes as in the chimp pattern, hairy becomes associated with chimp automatically. Similar automatic generalization is a well established feature of human learning (Baddeley, 1990).

Graceful Degradation

If a connectionist network is trained to respond to one input pattern and receives a slightly distorted version of that input it will probably still produce the proper response. A conventional rule-based system, on the other hand, will just fail to produce a match for the input and will have no suitable response at all. If the input is distorted even more, the connectionist network will not fail, it will just become increasingly more likely to provide an incorrect response. The performance of the connectionist network degrades gracefully as the quality of the input decreases. As with automatic generalization, graceful degradation is a common characteristic of human performance (Norman & Bobrow, 1975).

Goals

When a connectionist network is learning, it adjusts its connections' weights to reduce what it "perceives" to be incongruities between the input and a desired output. The desired output can be viewed as a "goal" and the learning process as goal satisfaction. "Connectionist models offer for the first time a convenient way of incorporating goals into the dynamics of information processing systems" (Estes, 1988).

Content Addressable Associative Recall

The sharing of common elements by similar concepts provides an automatic mechanism whereby an attempt to

retrieve one representation automatically activates similar representations and also activates the set of common elements which constitute the generalization of the other representations. The retrieval cue need contain only part of the input that was learned. The retrieval cue can even include incorrect information and the network will recall at least an approximation of the original material. This results in the instant, or at least very fast, retrieval of information the network has "learned".

Fault Tolerance

Distributed representation makes a connectionist network less sensitive to damage. The loss of a few nodes which are important to a concept may make the concept a little fuzzy but the concept can still be recalled. In contrast, with the kind of local representation found in a semantic network, loss of a node involves loss of an entire concept. The fact that brain damage in humans does not lead to the loss of discrete concepts suggests that the human brain also incorporates distributed representations.

Knowledge Structures

Learning in connectionist networks is accomplished through the automatic application of simple mathematical expressions and results in adjustments to connection weights. Knowledge, in such a network, is nothing more than the set of all connection weights across the entire network.

Many cognitive psychologists consider knowledge to be stored in relatively complex structures variously called frames (Minksy, 1977), scripts (Shank & Abelson, 1977), or schema (Rumelhart & Norman, 1987). "Such knowledge structures are assumed to be the basis of comprehension." (McClelland, Rumelhart & Hinton, 1986, p. 9). However, these complex constructs are only approximations of the actual underlying structure of knowledge as represented by the connectionist model. Although usually associated with semantic networks, these knowledge structures can be implemented in connectionist networks as well.

> In the network, you don't explicitly define the schemata; you only set the associations between pairs of descriptors. The schema emerges out of the network as a natural consequence of its behavior. Thus, the schemata are not explicitly represented in the network, but rather are simply patterns of activation across a set of descriptors (Zeidenberg, 1987, p. 237).

Such a network nicely accounts for such human cognitive behavior as activation of schema on incomplete information (associative recall) and the formation of overlapping schema (automatic generalization).

Introspection

Knowledge in connectionist networks is embedded inextricably in the machinery of processing. Consequently, this knowledge is completely inaccessible to introspection or report. "However, it should be noted that while the

connection changes themselves are not accessible, the patterns of activation [the connections] make it possible to construct can be accessible to other parts of the processing system" (McClelland, 1988, 112).

Criticisms

Although connectionist networks describe an interesting processing mechanism which seems to have much in common with human performance, such networks are not unconditionally accepted as suitable models of human cognition. Criticisms of the various connectionist models have ranged from dissatisfaction with the specifics of current implementations to outright rejection of connectionism.

The Perceptron Controversy: Limits to Learning

The perceptron mentioned above was developed three decades ago (Rosenblatt, 1962). This simple network consists of input nodes and output nodes but no hidden nodes. The result is a very simple network suitable for simulating (among other things) some aspects of human perception. This was the first connectionist network and was actually built into the computer hardware. Connectionist networks of this type are seldom used today (Hecht-Nielsen, 1988), but early criticisms of perceptrons still hinder acceptance of modern connectionist models.

Perceptrons could learn but they could not simulate complex performance. Minsky and Papert's (1969) much

publicized criticism of simple, connectionist networks involved a series of mathematical proofs which showed that a perceptron-like network was incapable of performing a number of elementary logical processes. However, a minor change in architecture, the addition of hidden nodes, allows connectionist networks to perform much more complex processes. Minsky and Papert were aware of the value of hidden nodes, but networks with significant numbers of hidden nodes are difficult to analyze in the formal manner they used to discredit simple networks. Their book was instrumental in discouraging research into neural networks even though "little attention was paid to the fact that they directed their criticism at a very simple system, the single-layer perceptron" (Zeidenberg, 1987).

One aspect of Minsky and Papert's criticism did address networks with hidden nodes. At that time no algorithm was known which would allow networks with hidden nodes to learn. They did, however, suggest that sometime in the future "perhaps some powerful convergence theorem will be discovered, or some profound reason for the failure to produce an interesting 'learning theorem' for the multilayered machine will be found" (Minsky & Papert, 1969, p. 232).

This represented a serious limitation of connectionist networks. Without hidden nodes, a network must rely solely on distinctions which already exist in the input data. If

the data is "discontinuous or nonlinearly separable" these distinctions may not be sufficient to permit the network to learn. Adding hidden nodes allows the network to develop its own, internal representations, allowing it to learn patterns which are not linearly separable (Caudill, 1988). Much human cognition involves even less discrete input, so it would seem that to be effective, connectionist networks must have hidden nodes. But, if connectionist networks require hidden nodes to simulate complex human performance, then their value is strictly limited in the absence of the ability of hidden units to learn. This limitation has recently been redressed by the back-propagation learning rule.

In an opening talk at the 1988 IEEE International Conference on Connectionist Networks, Marvin Minsky acknowledged that:

> Given a threshold number of connections between a set of simple neurons, a form of self-organization takes place, and from this organization collective computational properties emerge, such as association, generalization, differentiation, preferential learning, optimization, and fault tolerance. (Josin, 1987, 184).

In a clever and amusing account of this controversy, Papert (1988) has also all but retracted his earlier criticisms:

> Once upon a time two daughter sciences were born to the new science of cybernetics. One

sister was natural, with features inherited from the study of the brain, from the way nature does things. The other was artificial, related from the beginning to the use of computers. Each of the sister sciences tried to build models of intelligence, but from very different materials. The natural sister built models (called neural networks) out of mathematically purified neurones. The artificial sister built her models out of computer programs.

In their first bloom of youth the two were equally successful and equally pursued by suitors from other fields of knowledge. They got on very well together. Their relationship changed in the early sixties when a new monarch appeared, one with the largest coffers ever seen in the kingdom of the sciences: Lord DARPA, the Defense Department's Advanced Research Projects Agency. The artificial sister grew jealous and was determined to keep for herself the access to lord DARPA's research funds. The natural sister would have to be slain.

The bloody work was attempted by two staunch followers of the artificial sister, Marvin Minsky and Seymour Papert, cast in the role of the huntsman sent to slay Snow White and bring back her heart as proof of the deed. Their weapon was not the dagger but the mightier pen, from which came a book -- Perceptrons -- purporting to prove that neural nets could never fill their promise of building models of mind: *only computer programs could do this*. Victory seemed assured for the artificial sister. And indeed, for the next decade all the rewards of the kingdom came to her

progeny, of which the family of expert systems did best in fame and fortune.

But Snow White was not dead. What Minsky and Papert had shown the world as proof was not the heart of the princess; it was the heart of a pig. To be more literal: their book was read as proving that the neural net approach to building models of mind was dead. But a closer look reveals that they really demonstrated something much less than this. The book did indeed point out very serious limitations of a certain class of nets (nowadays known as one-layer perceptrons) but was misleading in its suggestion that this class of nets was the heart of connectionism . . .

Connectionist writings present the story as having a happy ending. The natural sister was quietly nurtured in the laboratories of a few ardent researchers who kept the faith, even when the world at large let itself be convinced that the enterprise was futile. . . But for the moment suffice it to note that the princess has emerged from relative rags and obscurity to win the admiration of all except a few of her sister's disgruntled hangers-on. (Papert, 1988)

Level of Analysis

Some critics of connectionist models (e.g., Fodor & Pylyshyn, 1988) would agree that neural networks are a reasonable representation of brain functioning, but they feel that psychological theories should be grounded in higher-order structures. "Accepting the higher-level regularities in symbol processes means accepting the charge that the lower-level interactions are sometimes

implementational. It does not mean accepting that they are always or 'merely' implementational." (Walker, 1990, p. 35)

Although more physiological than cognitive, Marr's (1982) framework of three levels of analysis is often used to evaluate computer-based cognitive models (Sejnowski & Churchland, 1989; Pylyshyn, 1989). This framework distinguishes between abstract, procedural (algorithmic), and implementational (architectural) levels of analysis, and it suggests that analysis at one level can proceed in the absence of understanding at lower levels. In the past, some cognitive scientists applied this "doctrine of independence" to study the mind at the level of symbolic algorithms at a time when little was known about the architecture of the brain. Unfortunately, this doctrine has also been misapplied in that today some researchers feel that what is now known about the implementation level of cognition has nothing to contribute to analysis at the algorithmic level. "In contrast to the doctrine of independence, current [connectionist] research suggests that considerations of implementation are vital in the kinds of algorithms that are devised and the kind of computational insight available to the scientist" (Sejnowski & Churchland, 1989, p. 303).

Many connectionist researchers argue that the architecture of a model places strong constraints on the types of algorithms the model can support. When most cognitive models were implemented on similar serial, symbol

processing computers, all of the models were subject to the same constraints, and thus disregard for the lower level of analysis may have been justifiable. Connectionist models assume a parallel architecture (real or simulated) and thus are subject to some different constraints than most symbol processing models. Many connectionists argue that this architecture is more biologically plausible and that analysis at the architectural level is as important as that at other levels.

> It is the architecture that determines which kinds of algorithms are most easily carried out on the machine in question. It is the architecture of the machine that determines the essential nature of the program itself. It is thus reasonable that we would begin by asking what we know about the architecture of the brain and how it might shape the algorithms underlying biological intelligence and human mental life (Rumelhart, 1989).

Another aspect of the level at which analysis proceeds has to do with the power of a theory. Meaningful theories must be testable. For some theorists, this means that theories should occasionally fail. Some critics of connectionism (Estes, 1988) feel that connectionist models "are too powerful to be susceptible to direct empirical test". Massaro (1988) presents a mathematical analysis of the ability of a multi-layered, fully interconnected connectionist network to process linguistic information, and

concludes that such a model is too powerful to be of any theoretical value.

In a traditional, empirical assessment of the performance of competing models, this may pose a problem but there are other ways of evaluating connectionist models. Chapter 6 argues for a more qualitative methodology instead of looking at performance based on the power of the model.

Relationship to Other Paradigms

Some cognitive scientists suggest that connectionism may represent the start of a paradigm shift for psychology (Schneider, 1987). I feel that while connectionism contradicts the premises of some psychological paradigms, it represents a synthesis of the common ground among others. This section discusses some of that common ground. The strength of any new theoretical perspective can be seen in its ability to subsume, or at least co-exist with, existing competing theories.

Constructivist

An important characteristic of connectionist models is that input patterns are not stored. Instead, connection strengths are modified so that, at a later date, the input pattern can be recreated. This constructivist view suggests that learning is a matter of "finding the right connection strengths so that the right patterns of activation will be

produced under the right circumstances" (McClelland, Rumelhart & Hinton, 1986, p. 32).

Gestalt

A number of characteristics of Gestalt theories are consistent with a connectionist perspective. They both attempt to "explain the organization of the perceptual world, not its relationship with the environment" and these organizing processes are relatively automatic (Epstein, 1988). With both Gestalt and Connectionist theories, biologically plausible mechanisms are preferred, and in both processing is distributed rather than under the control of a central executive process (van Leeuwen, 1989).

Empiricist/Nativist

Although the emphasis within connectionism is on systems which learn by adjusting connection weights in response to input (empiricism), networks start with non-zero connection weights. These are often small random values, but in theory, they could be significant values which predispose a network towards some initial reaction (nativism).

> A subsidiary role for [connectionism] could be to inject some empiricist realism into post-Chomskyean theories of human cognition . . . a test of the degree to which any connectionism is merely a neural kind of materialism (mind depends on the brain, and the brain is a connectionist machine) is whether it makes any predictions, both

about experimental results and desirable social interventions" (Walker, 1990, 34).

Cognitive

Cognitive models are usually based on explicit rules and provide a sophisticated mechanism for selecting and applying the rules. Connectionist models provide a very simple, local mechanism that does nothing more than adjust connection strengths which allows a "network of simple nodes to act as though it knew the rules" (McClelland, Rumelhart & Hinton, 1986, p. 32).

The lack of a formal logic mechanism is not seen as a limitation of connectionist models. Since logic is "a system that was invented as a corrective for human thought [it] constitutes an improbable candidate for being the basis for thought" (Kaplan, Weaver, & French, 1990, p. 68).

Connectionist models "learn" how to behave. The dynamic representation of knowledge in the ever-changing connection weights of the network are at the heart of a connectionist model. Cognitive models are mostly concerned with a static representation of what is and have little concern for how it got that way.

Behaviorist/Associationist

The lack of higher-order executive processes and a concern for learning suggest that connectionist models are at least associationistic if not behavioristic. Connectionist models "recapture the associationist's and

behaviorist's interest in learning which cognitivists largely gave up in their search for mechanisms of the mind that were often taken to be innate" (Bechtel, 1985, 56). Papert (1988) has even suggested that connectionism now "promises a vindication of behaviorism". However, Kaplan, Weaver, and French (1990) suggest that to characterize connectionism as a "computerized revival of behaviorism" is appropriate only for the more simplistic architectures. Unlike behaviorism, however, connectionism is interested in the cognitive mechanism which mediates responses to stimuli.

Conclusion

This chapter has discussed connectionism from a theoretical perspective -- defining some basic terms and setting a context for discussion. Now that the stage is set, the next chapter will look at some specific connectionist research.

CHAPTER TWO: CONNECTIONIST RESEARCH

Connectionist theories and networks are increasingly being used as a tool for research into human cognition. Some connectionist research paradigms resemble the conventional, empirical research performed by most experimental psychologists over the last few decades. Other connectionist research has more of a computing-science flavour where the intent is to build a computer simulation of human behaviour. Some connectionist research combines these two approaches and uses the behaviour of a simulation to "predict" the responses of subjects in a subsequent experiment. This section presents some examples of connectionist research, in part to address specific theoretical and methodological issues, and in part to establish an approach for the research project associated with this dissertation.

Some connectionist research focuses on the theoretical implications of a connectionist model of human cognition. For such theoretical models to be accepted as valid they must account for what is already known about human behaviour. To be accepted as useful, connectionist theories must add something new to existing accounts of human behaviour. The first few examples of research presented here address specific issues or implications arising out of

the application of connectionist theories to human performance.

Other connectionist research actually implements computer-based connectionist models which perform in a manner similar to human performance with little regard for maximizing artificial-intelligence-like, problem-solving performance. Several examples of this type of research are presented as well.

Research To Support Theories

The research presented in this section was conducted to provide evidence in support of specific aspects of connectionist theories. Many of these specific aspects were discussed in the previous chapter and this section is organized to parallel that discussion.

Content Addressability

Most cognitive information processing models claim that recall processes are mediated by the manipulation of symbols. These models can easily handle arbitrary, symbol based memory retrieval but have problems with content addressable memory retrieval. The empirical data suggests that human cognition is just the opposite. Human recall of content addressable information is relatively easy and natural, while any significant amount of arbitrary information presents a major memory task (Oaksford, Charter & Stenning, 1990).

Starting with the assumption that the meaning of a concept can be decomposed into its semantic elements or features, in what way does the number of features (complexity) of a concept affect memory load and processing time? Within the context of the decomposition assumption, most semantic memory models would predict an increased processing time for more complex concepts. Klimesch (1987) takes the fact that empirical observation does not confirm this as evidence against the decomposition assumption. Instead, he suggests that it is the semantic model of memory itself which is discounted by this evidence and presents a connectivity hypothesis with distributed representations of concepts to account for this.

Automatic Generalization

Organization theory suggests that in a free recall experiment, as the subjects study the to-be-recalled words, they chunk the words into subjective categories using a process similar to automatic generalization. When they recall the words, imperfect recall represents forgetting of entire categories rather than just some of the words within each category. Recall of lists of explicitly related groups of words can be facilitated by cueing the subjects with the name of the organizational categories. This helps because it makes sure that the subjects do not overlook any of the categories. Cued recall of unrelated lists of words is less straight-forward.

Penny (1988) was able to facilitate recall of unrelated items learned incidentally during a sorting task based on subjective categorization. On an unexpected recall text following the sorting task, the subjects were presented one item from each of the categories they had established during the sorting task and were asked to recall the others. The results were not only consistent with organizational theory, they also support the connectionist notion of automatic generalization.

Distributed Representation

John, Tang, Brill, Young, and Ono (1986) mapped the levels of activation in cat brains performing detection of previous learned visual stimuli. They found that, depending on how extensively the cat had been trained to the stimulus, between 5 million and 100 million neurons were activated by presentation of the visual cues. More important, they found that the activated neurons were widely distributed throughout the brain. This finding is "compatible with prior evidence of a distributed memory system" and "difficult to reconcile with theories in which individual neurons are dedicated to specific memories".

Learning Rules

As mentioned above, one of the constraints which distinguish different connectionist architectures is the learning rule each uses. Gluck and Bower (1988) discuss a series of experiments which test the appropriateness of a

least-mean-squares learning rule (similar to the delta learning rule). They compared the performance of an appropriately configured connectionist network to that of human subjects on a category learning task involving simulated medical diagnosis. "The results of these three experiments provide preliminary converging evidence that the LMS rule is more general than formerly believed" (Gluck & Bower, 1988) and is able to serve as the basis for a model of human category learning.

Research To Model Performance

This section provides some examples of research based on connectionist models and uses these examples to make an argument for the suitability of one specific class of connectionist models for the research project associated with this dissertation.

Gluck and Bower (1988) present two general methodologies for using connectionist networks to investigate human cognition. One involves selecting some aspect of human performance and constructing a network to perform the same task in a manner such that the "major regularities and salient phenomena" are preserved. The second methodology focuses on a specific experimental paradigm and then builds a network whose performance will predict human performance within that paradigm. The studies presented here show a progression from simple to more

complex models. Taken as a whole, these studies suggest that a multi-layer network using a back-propagation network is likely to be the most appropriate configuration for a network to investigate the aspects of human cognition addressed by this dissertation.

Simple Connectionist Network

Gluck and Bower (1988) present a series of three experiments in which they simulate "human category learning" within the Rescorla-Wagner associative learning paradigm. They used a categorization task which consisted of diagnosing one of two mutually exclusive diseases based on the presence or absence of four symptoms. They used a simple, two-layer network with four input nodes reflecting the symptoms and a single output node reflecting the binary categorization. They used a least-mean-squares (LMS) learning rule (also known as the Wodrow-Hoff rule or the delta rule) to train the network. The LMS rule implements the Rescorla-Wagner paradigm. Their objective was to discover the extent to which the LMS rule (and indirectly the Rescorla-Wagner paradigm) provides "an empirically accurate account of how people learn".

The three experiments varied the training set with respect to: the frequency of the two diseases, the predictive value of symptoms, and the extent to which the absence of a symptom implies the presence of its converse. The methodology consisted of determining the performance

predicted by the LMS model then comparing these predictions to the performance of human subjects. In general, the performance of the subjects was consistent with the predictions of the LMS rule.

Although single-layer networks are able to simulate human categorization tasks, they are unable to perform some very simple discrimination tasks. The most commonly discussed short-coming of single-layer networks is their inability to solve the exclusive-or (XOR) problem. This problem requires the network to respond positively if either of two inputs is present and negatively if neither is present or if both are present. Although single-layer networks cannot learn to solve this problem, networks with one or more layer of hidden nodes can.

Multi-Layer Networks

Kehoe (1989) characterizes stimulus discrimination learning within a classical conditioning paradigm as a special case of this XOR problem. In this case, the subject is trained to respond only when either but not both of two conditioned stimuli are present. The performance of a properly configured connectionist network is extremely similar to animals conditioned to respond to either a tone (CST) or a light (CSL).

The topology of the network used by Kehoe is just slightly more complicated than the minimum usually required to solve the XOR problem in order to account for the

presence of the unconditioned stimulus (UCS). Three input nodes representing the UCS, CST and CSL are fully interconnected with a hidden layer of two nodes. The two hidden nodes are connected to a single output node representing the presence or absence of the response (R).

Connections in multi-layer networks are usually restricted to adjacent layers, but Kehoe has added a connection between the UCS node and the R node. Although this is not typical of multi-layer topologies, it is certainly consistent with the classical conditioning paradigm.

Another manner in which Kehoe diverges from a "standard" multi-layer configuration is in the assignment of different output thresholds to the two hidden units. In theory, all of the "knowledge" represented in a connectionist network resides in the connections. Occasionally, some implementations hard-wire these connections to represent specific knowledge in the domain within which the network is expected to perform, but in most cases connectionist networks are expected to "learn" what the connection weights should be and this learning is an important part of the simulation. Hard-wired output thresholds constitute another form of outside knowledge. The theoretical basis for connectionist networks does not postulate a mechanism whereby these thresholds could be learned.

If the purpose of a connectionist network is to simulate human behaviour for some practical purpose (screening loan applicants, interpreting sonar signals, detecting bombs, etc.), then it probably does not matter where knowledge resides in the network or where it came from. On the other hand, if the purpose of the simulation is to increase our theoretical understanding of human learning and behaviour, then any form of hard-wired knowledge requires some theoretical explanation for its inclusion. Kehoe provides no such justification for the existence of the thresholds on the hidden-layer nodes or for the values to which they are set. This diminishes the usefulness of Kehoe's investigations in terms of a full understanding of how classical conditioning proceeds.

Kehoe does, however, show that although Gluck and Bower (1988) were able to simulate some aspects of human behaviour using a simple, one-layer network, simulating other behaviour will likely require a multi-layer network. Further arguments in favour of multi-layer networks are provided by Klimesch (1987).

Klimesch (1987) compared connectionist networks to semantic networks with respect to their ability to predict the human processing requirements of complex versus simple stimuli. Connectionist and semantic models both assume that complex stimuli can be decomposed into properties or features and that processing a stimulus requires processing

the individual features rather than the stimulus as a whole. This decomposition assumption predicts that complex stimuli (those with more features) require more processing than less complex stimuli. Since semantic networks are essentially serial processing models, they predict that human reaction times should be greater when processing complex stimuli. Connectionist networks, on the other hand, are essentially parallel and predict equal reaction times for both simple and complex stimuli.

Klimesch goes even further and presents a connectionist model which predicts reduced reaction times for complex stimuli. This model assumes a richly interconnected topology in which signals reverberate between feature detector nodes and response nodes until the response node becomes sufficiently activated. In theory, a single persistent feature will eventually produce the response, but each additional active feature adds signal strength and causes the activation level of the response node to increase at a greater rate. The more features which are present, the sooner the response node will fire. Klimesch presents the results of a series of experiments on human subjects which support the predictions of this model. Most implementations of connectionist networks are much less richly interconnected than this model, but Klimesch does provide additional arguments indicating that connectionist models of human behaviour require multi-layer networks.

To facilitate understanding of human cognition, a connectionist model should perform in a manner similar to human behaviour, but it should also learn to perform that way under conditions similar to human knowledge acquisition. MacWhinney, Leinbach, Taraban and McDonald (1989) conducted a series of simulation experiments to explore learning in a multi-layer connectionist network. The task was to determine the proper definite article for a series of German language nouns.

The network consisted of four layers of nodes: an input layer representing features of the input noun; a hidden layer with two pools of nodes representing the gender, case, and number (singular or plural); a second hidden layer representing no predetermined generalizations; and an output layer representing each of the six German definite articles. The different experiments varied the nature of the features represented by the input layer. In all, the number of nodes in the network was approximately 100 for each experiment.

Since this was a multi-layer network, the back-propagation learning rule was used. The 305 word training set consisted of 102 different nouns repeated between 1 and 17 times according to their approximate frequency of occurrence in the German language. The network was trained to a criterion of either 100% performance or 200 times through the training set (200 epochs). This was repeated for twenty simulated subjects. In thirteen of the

"subjects" the learning set was mastered. The network was still making errors on one or two words when the remaining seven "subjects" reached 200 epochs.

The model was successful on three counts. It performed in a manner similar to that of human subjects with respect to errors produced, progression of learning, and processing of novel input. Second, the model generated some clear predictions about performance in areas in which research on human subjects had not yet been conducted. Finally, "the success of the current model for this particularly difficult problem in language learning would seem to indicate that claims regarding the insufficiency of connectionist accounts for language learning . . . are, to say the least, premature" (MacWhinney, Leinbach, Taraban and McDonald, 1989, p. 275).

This model is notable in several instances. Including more occurrences of common words in the training set exposed the network to the words in a manner which more closely approximates natural human learning conditions. Adding a second layer of hidden nodes creates a more interesting network, even though the pools were pre-established by restricting the connections from the input layer. Finally, this model shows that a multi-layer network using the back-propagation learning rule exhibits behaviour similar to that of human subjects.

Another example of learning in back-propagation networks is provided by Norris (1990). Some idiot savants are able to determine the day of the week for almost any day, month, and year. Norris attempted to construct a connectionist network to model this behaviour.

The initial configuration of this network consisted of a simple back-propagation network with a single layer of hidden nodes. The input layer consisted of 31 day nodes, 12 month nodes, 5 decade nodes, and 10 year nodes; the hidden layer consisted of 50 nodes; and the output layer consisted of 7 day-of-week nodes. The training set consisted of one fifth of the dates randomly selected from the period 1950 to 1999. "After 1000 iterations through the training set the net performed reasonably well on dates on which it had been trained. However, on new dates the net's performance was little better than chance." (Norris, 1990, p. 280).

Norris concluded that, in order to generalize to novel input, the network required a second layer of hidden nodes to represent the "rules" required to perform the calculation. In addition, the training set was restructured to present the dates in order, and a more elaborate training procedure was used where the network first learned to process the dates in a single month and then a single year before being exposed to the rest of the set. After each of these stages, the associated weights were fixed to prevent interference from future learning. The new configuration

learned relatively quickly and reached 90% accuracy on novel dates. In addition, most errors were for dates in the first two months of leap years; an error profile which Norris reports as also being common to the human idiot savants.

Although the network required help to learn, "there is really nothing magical about the form of that help. What we have done is to make up for some of the deficiencies in currently available connectionist learning algorithms" (Norris, 1990, p. 286). The implication is that a more sophisticated learning algorithm would automatically consolidate previous learning and would not require this help.

The restructuring of the input to this network is similar to the approach taken by MacWhinney, Leinbach, Taraban and McDonald (1989) in constructing their training set. Both sets of input more closely approximate the way a human subject would naturally encounter the items.

The more interesting aspects of human cognition involve something more complex than simple pattern matching. The recent connectionist studies presented here show a move towards multi-layer networks to capture this complexity. In some cases, however, researchers have also moved away from the simple processing mechanism associated with most other connectionist networks.

Models which have been arbitrarily configured and re-configured until their performance meets certain

expectations may well say more about a programmer's ability to "hack" a solution than they say about human cognition. This is not to say that these models are not useful analogies of human behaviour, but they are less likely to capture any real understanding of human cognition than a network which learns an appropriate configuration with little or no help. Although back-propagation is not a perfect learning rule, it is capable of producing learning in a multi-layer network and, as such, is probably the most appropriate learning rule for connectionist networks which attempt to model human cognition.

Conclusions

There are several conclusions which can be reached from these samples of connectionist research. First, connectionist models certainly seem to be valid and useful tools for studying human cognition. Second, an accurate model of human learning would likely require a multi-layer network with a sophisticated learning rule. And third, the most appropriate learning rule for multi-layer networks would seem to be some form of back-propagation. All of the simulations presented later in this dissertation attempt to model human cognition using a multi-layer connectionist network with a back-propagation learning rule.

Connectionist networks represent a general class of models rather than one specific modeling technology. Any

one network implementation reflects only a specific instance of one theoretical model. When modeling human cognition in a connectionist network, it is important to identify the specific features of the network implementation and to ensure that those features model specific attributes of the human behaviour of interest. The next chapter describes back-propagation networks in more detail with a focus on their implementation details, and presents the results of a research project based on a back-propagation network.

CHAPTER THREE: BACK-PROPAGATION

Learning rules for two layer perceptrons are relatively easy to specify but such networks can't model complex cognitive processes. Multi-layer networks (those with hidden nodes) can model complex processes, but a suitable learning rule is more difficult to specify. Although multi-layer connectionist networks have been in use for some time, suitable learning rules for these networks have been available for less than a decade. Most such learning rules are variations of the back-propagation procedure presented by Rumelhart, Hinton and Williams (1986) and Rumelhart (1989).

This chapter presents a general description of how back-propagation networks learn, then discusses how a back-propagation network addresses the specific implementation details discussed in Chapter 1. Consideration of one specific detail of the output function leads to suggestions for a modification of that function. The results of several learning simulations (with and without the modification) are presented and implications of the modification are discussed.

The Learning Rule

To understand how back-propagation works in a connectionist network with hidden nodes, consider a network

which learns to generate a series of outputs in response to a finite set of inputs. The network is presented with a number of pairs of patterns. One member of each pair is the input pattern and the other is the output pattern the network is expected to produce. The input pattern is presented and the network generates some (initially random) output. The difference between this output and the desired output (the amount of error) is used to adjust the weights so that a more appropriate output will be generated next time. This procedure is repeated for each of the pairs of patterns a large number of times until the total error across all patterns has been reduced to some acceptably small value.

The use of an error value (the difference between actual and expected output) is the basis for the *delta learning rule* in a connectionist network. If a network has no hidden nodes, then the delta rule is easy to apply because all of the error resides in the output nodes and the amount of error can be determined by simply comparing the actual and expected output values. However, in a multi-layer network the output nodes only account for some of the total error. The rest of the error comes from the hidden nodes. This makes application of the delta rule difficult because there is no explicit expected output value associated with a hidden node. Each node which acts as a source of activation for another node (i.e., all but the

output nodes) is at least partially responsible for the error associated with that destination node. Back-propagation is a technique for taking some of the error at the destination node and allocating it back to all of the source nodes.

The error value at a source node is based on the total error at the nodes to which it sends output, and also on the extent to which it is responsible for that total error. The total error is the sum of each destination node's error multiplied by the weight which connects that node to the source node. The responsibility of the source node for that total error is a function of its output value. Once the amount of error at a source node has been determined, a generalized delta rule can be used to adjust the weight between the source node and each of its destination nodes.

A new pattern can now be presented and the entire process repeated. The back-propagation network continues to feed input patterns forward through the network and propagate error back until the total amount of error for all of the nodes in the network becomes sufficiently small that the network provides the correct response to each stimulus almost all of the time.

The amount of error in the network at any given time has a negative impact on the performance of the network and can thus be used as a measure (actually an inverse measure) of the amount the network has learned to that point. In

fact, learning in back-propagation networks is often defined as *error reduction through gradient descent.*

Error reduction through gradient descent can be illustrated using a topographic metaphor. If all of the potential error states that a network could ever be in can be visualized as a multi-dimensional landscape, the current error state can be represented by a point on that landscape. If the error point is given substance it becomes a spherical object. Changing the error in the system would then be analogous to rolling the sphere across the landscape. High points in the landscape represent large amounts of error and low points less error. Untrained networks have a large amount of error so the sphere starts out on a high point on the landscape. The objective of learning is to roll the sphere across the landscape until it comes to rest in a low spot: gradient descent.

In a simple network with no hidden nodes, the landscape becomes a bowl-shaped valley. Initially, the sphere starts somewhere up one side of the valley to indicate some amount of error greater than the minimum error at the bottom. The objective of learning is to minimize error so, as learning proceeds, the sphere will be moved downhill. When the sphere reaches the bottom, a move in any direction increases the error so learning stops. The learning rule for such a network is simple: always move the sphere downhill, and when you can't you're done.

Adding hidden nodes to a network complicates the situation immensely. Each additional node actually adds another dimension to the landscape but the effect is easier to visualize if, instead, you think of additional nodes as making the bowl-shaped valley more irregular. The result is that depressions (*local minima*) can form at higher elevations. If the sphere rolls into one of these depressions it must roll uphill for a while before it can continue down into the valley.

The difficulty lies in distinguishing between these local minima and the bottom of the valley. If the sphere fails to make this distinction it may either become stuck in a local minima and thus stay at a high error level, or it may try to roll up out of the bottom thus continuing to try to learn when it was already at the lowest level possible. In either case, learning will not proceed to an optimal solution. Optimizing the gradient descent in this error-space is the objective of many of the variations on the basic back-propagation learning rule.

Implementation Details

So far, this discussion of back-propagation has been very general but specific back-propagation networks often contain variations designed to optimize the performance of the network. These variations can be classified by how they

respond to the architectural constraints discussed in Chapter 1.

Number of Layers

In theory, the back-propagation algorithm can be applied to a network with any number of layers but, in practice, the number of layers is usually quite small. The amount of error in the network's response to a given input can only be accurately determined at the output layer. The error values propagated back through layers of hidden nodes only approximate the effect of that layer on the final error. "Every time the error from the output layer is back-propagated to a previous one, it becomes less and less meaningful" (Caudill, 1991, p. 59).

Maren, Jones, and Franklin (1990) cite several mathematical proofs that suggest that no more than two layers of hidden nodes will ever be required. In addition, they suggest that empirical tests of back-propagation networks show no significant advantage to having more than one hidden layer, especially when each possible outcome is represented by a single output node.

Number of Nodes at Each Layer

Separate considerations apply to the number of nodes in the input, output, and hidden layers in a network. Usually the number of input and output nodes is strongly dictated by the nature of the task, but design considerations can affect them to an extent.

If each possible output of the network can be represented by a discrete output node, the network will likely only require one hidden layer (see above), but a large number of output nodes may be required. If the output is a continuous value or is encoded in a pattern of binary nodes, fewer output nodes may be required but the "use of encoding patterns forces additional work onto the hidden nodes, which may require an additional hidden layer" (Maren, Jones, and Franklin, 1990). Depending on the purpose of the network, encoding input as patterns is often just what is wanted since many of the more interesting phenomena of connectionist networks (automatic generalization, fault tolerance, etc.) only apply when input is represented as a pattern across several input nodes.

Determining the optimal number of hidden nodes is much more difficult. Maren, Jones, and Franklin (1990) suggest that the maximum number of hidden nodes should be less than the number of input patterns (to avoid the formation of "grandmother" nodes) but more than the number of significant features in the input (so the network needn't come up with exactly the right representation). Although the exact number of hidden nodes is not a critical parameter, networks with an excessively large or small number of hidden nodes will train more slowly than ones with approximately the right number.

> Adding a few more connections creates extra dimensions in weight-space and these dimensions provide paths around the barriers that create poor local minima in the lower dimensional subspaces (Rumelhart, Hinton & Williams, 1986, p. 535).

In any case, Caudill (1991) suggests that when back-propagation networks are simulated in software running on serial rather than parallel hardware, the total size of the network should not exceed 200 to 300 nodes.

Connectivity

Connectivity specifies how the nodes in a layer can be connected to other nodes in the same or other layers. Most back-propagation networks have restrictions on the connections which can exist. Notably, "connections within a layer or from higher to lower layers are forbidden, but connections can skip intermediate layers" (Rumelhart, Hinton and Williams, 1986, p. 533).

Some models attempt to influence the performance of the network through specific configurations of connections. To the extent that this configuration represents "knowledge" imposed on the network, the value of the model may be compromised. Part of the value of connectionist models is that they are able to learn. If, instead of learning, a network is "hard-wired" to perform a task, then its performance says little about how a human might learn to perform a similar task. This would argue in favour of a network with a uniform connectivity configuration.

Starting Weights

The weights in a connectionist network represent the knowledge already in the network. In theory, one network could learn to perform many unrelated tasks. The weights learned by mastering a previous task which is totally unrelated to the current task would have a random effect on the current task. On the other hand, if previous learning is not totally unrelated to the current task, it could interfere with the current task and this would appear as non-random, potentially disadvantageous, weights. In practice, most networks are implemented to learn and perform one task only, so starting weights are usually set to random values.

If, by chance, a large random starting weight at one connection is significantly different than the optimal value for the current learning situation, it will take the network some time to "unlearn" that weight before settling into a more appropriate value. If, on the other hand, the random starting weights are restricted to relatively small values, the network will never have to unlearn a significantly inappropriate weight. Rumelhart, Hinton and Williams (1986) suggest that, in general, connectionist networks using back-propagation should start with all weights set to small random values.

Activation Function

The activation level at a specific node represents the amount of input that node has received from other nodes in the network. It is usually just the sum of the output of these other nodes multiplied by the weight of the connection to the specific node. This makes a node's activation a linear function of the inputs.

Output Function

One of the characteristic features of connectionist networks is that the output of a node is a non-linear reflection of its input. If the node's activation level is a linear reflection of its input, some non-linear transformation must be applied to the activation level to determine the node's output. There is no one transformation function which **must** be used with a back-propagation network but the function used must have a bounded derivative (Rumelhart, Hinton and Williams, 1986, p. 534). A popular choice is a logistic function which yields a sigmoidal distribution of output values over the range of possible activation values. For example:

$$\text{output} = 1 / (1 + e^{(-1 * \text{activation})})$$

Error Calculation Function

Back-propagation requires that an error value be calculated for each node in the network which receives input from other nodes (i.e., hidden and output nodes). For output nodes this is just the difference between the actual

and the expected output. For hidden nodes the error value is based on the net error at the destination nodes to which the hidden node sends output, and on some representation of the level of that output.

The net error at the destination nodes is usually calculated as the sum of each destination node's error value multiplied by the weight of the connection between the destination node and the hidden node.

The representation of the level of output at the hidden node is based on the activation level of the node but the function used is not the output function described above. Instead, the derivative of that output function is used.

> Applying the derivative serves two purposes. First, it contributes to the stability of the network since it ensures that, as the outputs approach 0 and 1, only very small changes can occur. Second, it helps compensate for excessive blame attached to the [hidden node] (Caudill, 1988).

For example, if the network's output function is the sigmoidal function described above, the error value for a hidden node would be:

error = (activation)(1 - activation) * (net error)

Momentum

The learning function adjusts the connection weights by applying a generalized delta rule to the output level of the node at the source of the connection and the error value of

the node at the destination of the connection. A momentum term is often added to this basic calculation. If a weight needs to be adjusted to reduce error, not all of that adjustment should be made on a single learning cycle, otherwise the network may over-react. If some smaller adjustment is made on one trial, then it is reasonable to suggest that further adjustments in the same direction will be required on subsequent trials. Many networks add some fraction of any previous weight adjustment to the current adjustment to preserve the momentum of learning at that weight.

> Applying momentum to a back-propagation network is almost certainly the single easiest thing you can do to make your network train faster - sometimes by orders of magnitude (Caudill, 1991, p. 59).

Temperature

The actual network implementations presented below evaluate an unusual variation of the output function in a back-propagation network. Boltzmann machines are neural networks which use a learning rule which is quite different from back-propagation (Hinton & Sejnowski, 1986). As these networks learn, they avoid local minima by gradually decreasing the overall activation level or "temperature" of the network. A similar application of temperature might improve the performance of a back-propagation network

As mentioned above, the output function is usually sigmoidal and has the following general form:

$$\text{output} = 1 / (1 + e^{(-1 * \text{activation})})$$

This general form may not be directly applicable to any one specific implementation because the average level of activation will depend as much on the various implementation details as on the input patterns to be learned. For this reason, in a specific network implementation, the actual activation value may be scaled by a constant value. In fact, this was the case in the network-based learning simulation presented by Caudill (1988). The formula used was:

$$\text{output} = 1 / (1 + \exp(-1 * \text{activation} / \text{constant}))$$

where $\exp(X) = e^X$. In this case, the constant had a value of 0.2, and Caudill's justification for that specific value was simply that it seemed to work best.

If the objective of a specific network implementation is to maximize performance in one problem-solving domain, anything which improves the performance of that network is appropriate, and it is generally useful as long as it can be successfully applied to other network implementations. It is not the value of the constant which determines the usefulness of this implementation detail, but the inclusion of a constant at all. To evaluate the usefulness of such a constant, it is necessary to consider its effect on network performance.

If, on average, the output signal is generally high for most nodes and across most input patterns, the network will react strongly to each input pattern and may over-react to the point where each new pattern obliterates much of what was learned from the previous pattern and overall learning will be difficult. On the other hand, if the output signals are generally small, the network will react very little to each input pattern and learning will be slow. If an appropriate value is chosen for the constant, some intermediate, and hopefully optimal, general level of output will result.

For a more graphic illustration, consider again the metaphor of error reduction through gradient descent as a point or sphere rolling across a landscape in error-space. Only now, add an element of liveliness or bounce to the sphere representing the error. As it moves through error space, a lively sphere will bounce across local minima and thus avoid the greatest problem with gradient descent. The problem with a lively sphere is that even when it reaches the bottom of the valley it will continue to bounce and the network's performance will be erratic. A less lively sphere will stay in the bottom once it gets there but is still likely to get stuck in especially deep local minima. If, however, the liveliness of the sphere starts high and gradually decreases, it will avoid local minima on early

learning trials, but later it will settle down at the optimum low point.

This would suggest that optimal learning would be achieved if the constant in the output function was replaced by a variable which started with a relatively high value and then gradually decreased as the network learned. The formula would now be:

output = 1 / (1 + exp(-1 * activation / temperature))

where temperature is a variable which starts high and then decreases. I use the term temperature here because it was the description by Hinton and Sejnowski (1986) of the annealing process in Boltzmann machines which first caused me to consider applying a similar process to a back-propagation network.

Another way of describing the effects of this temperature variable is by considering the distribution of values produced by this output function. As mentioned above, this distribution is sigmoidal. Changes in the temperature parameter affect the slope of the mid-range of the function which, in turn, affects the variability of the output.

To test the effect on learning of temperature in a back-propagation network, I ran several simulations with and without such a variable. The specific implementation details and the results are presented below.

The Simulations

A series of network-based simulations using the back-propagation learning rule were run to compare the relative effect of using either a constant or a variable factor (temperature) to scale the activation value used to calculate the output signal from the nodes in the network. One of the dangers of scaling the activation at all is that the reaction of the network to each new pattern is dampened. The use of momentum in the learning rule would counteract that dampening effect, but it is conceivable that the use of both momentum and a declining temperature variable might produce undesirable interactions. Since momentum is a highly-regarded and useful implementation detail, additional simulations were run to look for interaction effects as well.

Method

The simulations were implemented using a connectionist network with a back-propagation learning rule. The network consisted of three layers with 35 input nodes, 4 hidden nodes, and 8 output nodes. The input and output nodes were determined by the material to be learned (see below). The number of hidden nodes in the original Caudill (1988) network seemed appropriate and was retained to facilitate comparison with that work.

The network was trained using ten pairs of patterns based on the first ten letters of the alphabet. Each input

pattern (stimulus) consisted of a five-by-seven bit-map of each letter. The output patterns (response) were the eight-bit binary representation of the ASCII value for each letter.

Training proceeded using a standard back-propagation technique. To begin with, all weights in the network were set to small random values and then the network began to cycle through a series of training trials or epochs. Each single epoch proceeded as follows:

- the input nodes were set to the first input pattern
- the input values were propagated through the hidden nodes to the output nodes
- the error value for each output node was calculated as the difference between the output value of the node and the value (zero or one) of the corresponding bit in the expected output pattern
- the error value for the entire pattern was calculated as the sum of the absolute values of the error at each of the output nodes
- the error at each of the output nodes was propagated back through the network and used to adjust the weights
- the above procedure was repeated for the next nine pairs of patterns.

The simulation continued to cycle through one epoch after another until the network learned to reliably produce the appropriate response.

In all, four "conditions" were simulated -- a two-by-two matrix with and without momentum and temperature. For the two conditions with momentum, each time a weight was adjusted one half of the previous adjustment for that pattern was added as well. For the two conditions with temperature, the scaling factor was initially set at 0.7 and reduced by 0.005 at the end of each epoch. To ensure that the initial random starting weights did not bias the results, each simulation was run twelve times to simulate twelve "subjects" in each condition.

Results

At the end of each epoch, the amount of error for each pattern was examined to see how well the patterns had been learned. If the error for any one of the ten patterns was greater than 10%, training continued, otherwise the simulation stopped and recorded the number of epochs required to reach this level of learning. If any simulation ran for over 500 epochs, the simulation stopped even though the learning criteria had not been met. This happened for only one "subject" in each of the first three "conditions". Since there is no way to know how many trials it might have taken these simulations to reach criteria (or if they ever would), these measures must be considered conservative estimates of how long it might have taken those simulations to learn. The number of epochs required to reach criterion

for the twelve simulation runs in each of the four conditions are presented in Table 3.1 in increasing order.

Table 3.1
Number of epochs required to learn for: a basic network, a network with just momentum, a network with just temperature, and a network with both momentum and temperature.

Basic Net	Momentum	Temperature	Mom. & Temp
64	43	55	45
65	60	78	51
106	73	86	52
113	84	88	56
131	84	93	59
136	85	98	60
149	96	99	60
231	111	112	61
253	115	120	70
322	173	162	89
467	273	282	89
501*	501*	501*	305
$\bar{X}$ 211.5	**141.5**	**147.8**	**83.1**
sd 142.5	**123.2**	**120.3**	**68.2**

* conservative estimates

When using a network with a back-propagation learning rule, there will always be a few simulations which will take considerably longer to learn, and there will even, occasionally, be some which will never learn. This means

that the distribution of possible values for the number of epochs required to learn will be very positively skewed. In addition, the theoretical effects of both temperature and momentum discussed above would suggest that both of these factors would reduce the variance in the distribution of possible values for learning rules including those factors. For these reasons, a parametric analysis of variance was considered inappropriate for this data. Instead, the equivalent randomization test was used to test for differences between the implementations with temperature and those without. The addition of a temperature term to the back-propagation learning rule was found to significantly improve the performance of the network ($p < 0.044$).

Conclusions

The inclusion of a declining temperature term in a back-propagation network certainly seems to decrease the amount of time it takes that network to learn. But when connectionist networks are used to model human cognition, the absolute performance of the network is not the most important consideration. If, on the other hand, some aspect of human performance behaves in a manner which is similar to a declining temperature, then it is appropriate to include temperature in a network model of that human performance. If, for example, there were no factor in human learning which continually decreased as learning improved, then it

would not be appropriate to include a declining temperature term in a network-based model of human learning. On the other hand, there is a factor which does decline with at least some forms of human learning and that factor is attention.

With continued practice, some forms of human performance become increasingly automatic. Within the context of a model of cognition which includes a limited attentional resource, such a change in performance can be characterized as a decrease in attention with learning. A network-based model which attempts to replicate those forms of learning might use a declining temperature term to model attention. The next chapter describes the theoretical basis for declining attention in human learning with a view to producing a back-propagation network model of such learning.

CHAPTER FOUR: ATTENTION

The previous chapter described the performance of several connectionist networks which varied according to how they treated a "temperature" term in the output function. Simulations based on the networks which decreased temperature as the network learned showed superior absolute performance in that they took fewer trials to learn. However, Chapter 2 made the argument that networks which are arbitrarily adjusted to produce the desired performance are less useful as models of human cognition compared to networks where such adjustments represent some theoretical construct. This chapter describes the interaction between attention and human learning leading to the conclusion that the effect of attention on human learning could be modelled by the temperature term in a connectionist network.

In order to identify the specific implementation features of such a network, it is important to precisely describe the human performance to be simulated and how that performance is measured. In this case, the performance involves the interaction between learning and attention, specifically the decrease of attention required as an increasingly automatic response is learned.

This chapter will start with a definition of attention as a limited cognitive resource and then will discuss the consequences of insufficient resources on cognitive

processes. The concepts of automatic and controlled processes and multiple concurrent processes will be discussed within the context of this definition of attention. Finally, a methodology for measuring the attentional resources currently in use by a cognitive process will be presented.

A Specific Definition of Attention

Within the context of this dissertation, the term attention will be used as it is usually represented in an information-processing model of cognition with a limited attentional resource. Within such a model, not only does cognitive processing proceed in parallel, much of the processing proceeds without our conscious attention -- it occurs automatically. Despite nativist claims that much of this automatic processing is hard-wired, at least some automatic processes are learned. The decrease in attention associated with learning which I am attempting to model is a reflection of the increasingly automatic nature of a response which results from repeated learning trials.

Automatic cognitive processes can best be understood in contrast to controlled processes. Controlled processes require active attention. Automatic processes proceed without attention. Further, according to Schneider and Shiffrin (1977) "any automatic process requires an appreciable amount of consistent training to develop fully".

This suggests that perhaps learning, or certainly over-learning, can be characterized as a shift from controlled to automatic processing. If learning is defined in this manner, it is important to know whether this shift from controlled to automatic processing is gradual or occurs more suddenly after a certain amount of learning has occurred.

Although it is possible that automaticity is an all-or-none phenomenon, recent research suggests that it may be considered a more continuously varying attribute of a learning situation (Cohen, Dunbar, and McClelland, 1990). These authors present a series of connectionist simulations which model the relationship between automaticity and attention within the context of the Stroop effect. They conclude that it is appropriate to model attention as a continuous variable:

> The mechanisms used in this model show how the principles of continuous processing, expressed in terms of the [connectionist] framework, can be applied to the study of attention (Cohen, Dunbar, and McClelland, 1990, p. 358).

If the human cognitive processing mechanism converts even some highly learned processes from controlled to automatic ones, then there is likely some advantage to doing so. The advantage can be seen if the human mechanism operates within the constraints of a limited attentional resource.

Miller (1956) is generally credited with the notion that individuals have a specific limit to their cognitive processing capacity in the form of a limited attentional resource. He put that limit in the general range of seven, plus or minus two discrete pieces of information. Miller presented several experiments which tested a subject's ability to make absolute judgments about the magnitudes of various aspects of a stimulus (for example, frequency of tones, loudness of tones, saltiness of taste, points on a line, etc.). In general, individuals could accurately distinguish between approximately seven or fewer magnitudes, but began to confuse different magnitudes when the number was increased. This value of 7 ± 2 is often referred to as Miller's magic number, and is the basis for the numerous citations of Miller's work over the years, but several other implications of his work have been more fully developed by those who adopted this assumption of a limited attentional resource.

Implications of a Limited Attentional Resource

If the human cognitive processing mechanism has limitations with regard to attention, then these limitations will affect human performance. Research in a number of areas has investigated these effects and some of that research is discussed here.

Graceful Degradation

The model of cognition being considered here suggests that within an individual there is a finite limit to the resources available at any one time. Depending on the nature of the process being executed, there may not be sufficient resources to meet the demand. What happens to a process which receives a smaller allocation of resources than it demands?

In the entirely mechanistic environment of computers, a process will not execute at all if the resources it requires are not available. The process will either wait until resources become available or it will fail. Norman and Bobrow (1975) suggest that the human information processing system is more flexible than that. If a process does not receive all of the resources it requires, it will attempt to function with the resources it does receive. The consequence of insufficient resources is usually a degradation in the quality of the output of the process. The amount of degradation will be, in some sense, proportional to the size of the short-fall in resources. Only occasionally, and under situations of extreme resource shortages, will a process fail entirely. They refer to this as the "principle of graceful degradation".

As Norman and Bobrow see it, human cognitive processing is similar to computer processing in that it consists of the execution of programs, but the availability of resources

affect the execution of those programs in different ways. In their model of human cognition, several programs usually work in concert as a set to achieve a specific purpose. Such a set of programs taken together represent a single process. Each program in the set requires some measure of attentional resource. Each program also requires some input and generates some output. The output provided by the entire process is likely to be a combination of the outputs of several programs. The output of other programs provide intermediate results which may be combined by other programs into other intermediate results culminating in the final output. A critical question is, in what way is the output of one program made available to a subsequent program and how does the availability of resources affect the exchange of information?

In a computer based process, a program requiring input from another must wait until the other program has entirely finished the processing necessary to generate the required output. Allocating greater or fewer resources to an entire process will only affect the time it takes to provide the final result. Norman and Bobrow suggest with the human information processing mechanism, programs produce "continually available output". From the moment a program begins to execute, it can supply output. Initially the quality of the output will be low and, if the program is starved for resources, it may never get much better. If,

however, the program is receiving even minimal resources, it can work to improve the quality of its output. Increasing the resources available will further increase the quality of the output. This means that a program which requires input from another program can begin executing (and producing output) immediately using whatever quality of input it can get. The quality of its output will depend on the quality of its input and on the resources it is allocated. A general increase in the allocation of resources to the overall process increases the quality of output from that process.

The conclusion to be reached from this is that a decrease in the resources available to a process produces a graceful degradation in the performance of that process rather than an outright failure. Further, a change in the amount of resources allocated to a process can be inferred from a change in the quality of its output.

In addition to quantifying the size of available attentional resources, Miller (1956) suggested some strategies available to cognitive processes which allow them to process larger amounts of information with fewer resources. Chunking is one such strategy. This strategy devotes some of the available resources to combining and encoding the information to be processed so that more will fit into the remaining resources. This additional processing adds overhead to the process, but it can result

in a net gain. For example, in theory, a strategy (requiring one chunk of overhead) to make a series of six binary evaluations (using the remaining chunks) might permit discriminations to be made between as many as 64 (2^6) items instead of the usual 5 to 9. In practice, these more complex information processing strategies are rarely that efficient, but Bereiter and Scardamalia (1987) suggest the net result is that, normally, adults performing "attention-demanding" operations have sufficient remaining capacity to hold five chunks.

The reading and writing of text provides an excellent example of the effects of chunking. When young children are initially exposed to written language, it requires almost all of their attentional resources to process individual letters. With practice, however, they are soon able to deal with text as a series of words instead. Eventually, with even more practice, text is processed as a series of phrases and sentences instead of individual words.

Chunking represents a gradual change in the way information is processed. Specifically, with practice, the strategy of chunking allows a cognitive mechanism to come to process more and more information while still using the same amount of attentional resource. Similarly, a mechanism can come to process the same amount of information using fewer and fewer resources. This is shown in the shift from controlled to automatic processing.

Automatic versus Controlled Processes

Schneider and Shiffrin (1977) view memory as consisting of a large set of inter-associated nodes with the associations established through learning (see also Anderson, 1983). At any one time, most of the nodes are inactive. All of the inactive nodes taken together represent long-term memory (LTM). Also at any one time, a small set of nodes will be active. This set of active nodes constitutes short-term memory (STM). STM, then, consists of nodes which would be part of LTM were they not active and which will return to LTM when they decay from STM.

Various processes exist which influence the activation of nodes and, hence, the flow of information into and out of STM. According to Schneider and Shiffrin, these processes include "decisions of all sorts, rehearsal, coding, and search of short- and long-term [memory]". Processes are, themselves, stored in one or more LTM nodes but they need not enter STM to execute.

Within the context of this model, automatic processes execute (a) in response to a specific input configuration and (b) without the subject's attention. Such a process requires relatively permanent associations in LTM built up through an "appreciable amount of consistent training" and are "difficult to suppress, to modify, or to ignore". An example of an automatic process is a search task which responds to a target (input) by enabling a correct detection

to occur (output). Note that this is in response to a well trained stimulus. This processing proceeds automatically "regardless of concurrent inputs or memory load".

In contrast with automatic processes, controlled processes are activated by attention. In general, only one controlled process executes at a time (although several slow processes can be interleaved) and are subject to STM capacity limitations (7±2). The advantage of controlled processes is that they are easy to set up and can respond to novel situations for which automatic processes have not been learned. An example of a controlled process is a search task involving a target which had not yet been extensively learned. The target is compared against all possible responses until a match is found.

Keele (1972) also starts from the premise that individuals have some limit on their information processing capacity and that multiple concurrent tasks interfere with each other to the extent that the individual's limit is exceeded. Keele further characterizes tasks as consisting of two sub-processes, retrieval of information from memory and operations performed upon that information. Limitations on an individual's ability to perform simultaneous tasks may, theoretically, arise during either sub-process, but Keele argues that if signals irrelevant to a task can be shown to contact memory yet not interfere with that task,

then, under those circumstances, competition for resources does not occur during memory retrieval.

Keele attempted to demonstrate this with a variation on the Stroop effect. The experiment used five types of stimuli: colour words, non-colour words, scrambled letters from the colour words, mixed Gibson forms (letter-like symbols), and pure Gibson forms (a 'word' made up of the same Gibson form repeated several times). Each stimulus could appear printed in one of four colours of ink. The subject's task involved pressing one of four colour-coded keys to identify the colour of ink of the presented stimulus.

Subjects' reaction times were significantly slower to colour-word stimuli, with no difference among the other four types. In particular, reaction time to non-colour words was faster than to colour words. The subjects were obviously discriminating between colour words and non-colour words. This discrimination must have occurred at the semantic level. Keele considers memory retrieval to be prerequisite to a semantic evaluation, and therefore, all stimuli must have contacted memory. The conclusion is that conflict between simultaneous tasks occurs at the operations stage and that "memory retrieval is not attention demanding".

All of this research suggests that even moderately complex cognitive processes may actually consist of several sub-processes executing at the same time. If the human

cognitive processing mechanism can support multiple concurrent processes all of which demand attentional resources, then it is important to consider the way in which the mechanism responds when all demands cannot be met.

Multiple Concurrent Processes

Gopher and Navon (1980) consider the consequences when a cognitive processing mechanism attempts to perform two tasks simultaneously. They also begin with the assumption of one central pool of cognitive resources. These resources are differentially allocated to all tasks being processes depending on task requirements. When the requirements of all of the tasks exceed the capacity of the central pool, the tasks interfere with each other.

The requirements of a task, and hence the resources allocated to that task, are not solely a characteristic of the nature of the task. Intention on the part of the subject can influence the resources allocated to a task, especially when resources are scarce. Subjects can assign priorities to tasks with high priority tasks receiving a greater share of resources.

> If time-shared tasks are assumed to compete for allocation of the same resources, then increasing the priority of one task should result in an increment of its share of resources. This should lead to an improvement in its performance. Simultaneously, the decreased amount of resources allotted to the other task should now lead to a

decrement in its performance (Gopher & Navon, 1980).

The multiple concurrent tasks used by Gopher and Navon consisted of a two dimensional tracking task. The assumption was that tracking in each dimension represented a separate task. The difficulty of each task was varied by changing the velocity of the target and the frequency with which it changed direction. Priorities were manipulated by varying the minimum acceptable level of tracking accuracy.

In a more natural setting, individuals do not usually receive specific instructions about which task they should be attending to. Instead, attentional resources are allocated to multiple concurrent tasks on the basis of the degree to which each task has become automated.

Measuring Attentional Resources

If a certain cognitive task currently requires a specific amount of attentional resource and if learning is claimed to be able to reduce the amount required, then it becomes important to be able to measure the amount of attentional resource a process requires both before and after the learning takes place. The principle of graceful degradation described above provides a way in which this can be done.

One of the conclusions reached by Baddeley, Lewis, Eldridge and Thomson (1984) from a series of experiments on

the relationship between attention and long-term memory is that additional load placed on attentional resources during learning decreases recall performance. This suggests that the decrease in performance could be used as a measure of the amount of resources being used by the additional load.

In a similar vein, Norman and Bobrow (1975) suggested a link between reaction time and accuracy as performance measures. Within the context of their distinction between data-limited and resource-limited processes, paired-associate learning involving simple or familiar stimuli would be a resource-limited process.

> When a process is resource-limited, then we expect reaction time to be directly related to accuracy, because better resulting output is dependent on more processing resources being allocated to the process (Norman and Bobrow, 1975, p. 53).

Bower and Clapper (1989), in a discussion of experimental methods in cognitive science, also suggest a dual task methodology for measuring attention. To the extent that an individual has a limited attentional resource any concurrent tasks in which the individual is engaged must share this resource. If less attention is required by one task, then more is available for another. If performance on one task depends on the amount of attention it receives, then decreased attention on another task will increase performance on the first one.

They suggest that a suitable measure of this type of attention-related performance is reaction time to a probe stimulus. This methodology asks a subject engaged in a primary learning task to also respond to an unrelated stimulus (probe). For example, a subject attempting to learn a list of paired-associate items might also be asked to press a button when a tone sounds. The subject's reaction time to the probe "is presumed to be slower the more absorbing the primary task is at the moment the probe appears" (Bower and Clapper, 1989, p. 288).

In a description of attentional allocation for concurrent tasks, Sperling and Dosher (1986) also suggest that the amount of attentional resource allocated to each of several concurrent tasks determines the quality of performance on that task. Performance on the probe task is measured by reaction time. If this task gets fewer resources because more resources are allocated to learning, performance will decrease and reaction time will go up.

Conclusions

The purpose of discussing the research presented here was to present a particular perspective on the allocation of a limited attentional resource to cognitive processes and to discuss how the amount of resources required by a process changes with learning. Specifically, the points raised suggest that:

- Cognitive processes can be characterized as those which require attention (controlled processes) and those which do not (automatic processes).
- With sufficient practice, it is possible for some types of controlled processes to become automatic.
- Each individual has a fixed amount of attentional resource. When that individual engages in an activity which requires a controlled cognitive process, that controlled process monopolizes a certain amount of attentional resource. Automatic processes require none of the individual's attentional resource.
- There is an optimum amount of attentional resource a controlled process can use, but if the process receives less than that amount it will degrade gracefully rather than fail outright as long as a certain minimum amount of attentional resource is available.
- An individual can engage in a number of concurrent tasks as long as there are sufficient attentional resources to meet the minimum demands of all of the controlled processes.
- One way of determining the relative attentional resource demands of two tasks is to monitor the performance of a third task in the presence of first one then the other of the two tasks.

From the point of view of this dissertation, the question is whether a temperature term which declines as a back-propagation network learns is a suitable model of the decline in attentional resources allocated to a task as a human learns to perform the task with greater and greater fluency. Before such a network could be accurately developed and evaluated, it was necessary to more precisely establish the relationship between learning and attention. The next chapter presents a human learning experiment which uses the methodology discussed above to attempt to more clearly establish that relationship.

CHAPTER FIVE: HUMAN LEARNING EXPERIMENT

Despite its empirical methodology, this experiment was not designed to test specific hypotheses. Instead, it was intended to substantiate expected phenomena in a specific context and gain some qualitative insight into the nature of these phenomena with the intention of replicating the phenomena in a connectionist network and establishing criteria for evaluating its performance. Specifically, the objectives of this experiment were:

- Substantiate the distinction between automatic and controlled processes as a continuous rather than dichotomous one (i.e., some processes may be entirely automatic, but all others are controlled to a greater or lesser extent, depending on the amount of attentional resources they demand).
- Establish that the transition from a controlled process to an automatic one, as the result of practice, is not sudden. Instead, the degree to which a process is controlled gradually declines.
- Identify a mathematical description of the transition from a controlled process to an automatic one by monitoring the decrease in attentional resources required.

Method

The purpose of this experiment was to obtain a mathematical description of the transition of a simple cognitive learning task from a controlled process to an automatic one with "controlled" and "automatic" being

defined in terms of the amount of attentional resource demanded as discussed in Chapter 3. The learning task was accompanied by a series of reaction time probes designed to measure the learner's unallocated attentional resources and, indirectly, the amount of attentional resources required by the learning task. The amount of attentional resources required by the learning process was expected to decline in some regular way for all subjects regardless of their absolute level of performance. It is the nature of this regularity which is of interest here, not the absolute performance of the subjects.

Subjects

Nothing about the nature of this experiment suggested that any one population would be especially appropriate or inappropriate because the results were to be based on within-subject measures. Volunteers were solicited from a single class of grade-ten math students attending Lester B. Pearson High School in Calgary, Alberta. Since this particular high school integrates all three grade-ten math streams, the one class represented a range of academic ability. This school also integrates computer technology extensively across the curriculum, so these students already had considerable experience with computers in a variety of domains, and had specific experience with Microsoft Windows, the graphic user interface used for this experiment (see Materials below). No counterbalancing or random assignment

was required because all subjects participated in the same, computer-based, learning task.

Materials

The materials for this task consisted of twelve arbitrary paired-associate items. The stimulus half of each pair was one of twelve small pictures selected from the icons supplied with Microsoft Visual Basic. Each icon is a 32x32 pixel colour bitmap similar to those illustrated in Figure 5.1. The response half of each pair was one of four keys on a standard computer keyboard, specifically, D, F, J, and K. These keys were chosen to allow comfortable hand positioning and because they can be located by feel (on IBM keyboards, the F and J keys have bumps on them). For each subject, the specific key associated with each picture was assigned by random selection without replacement.

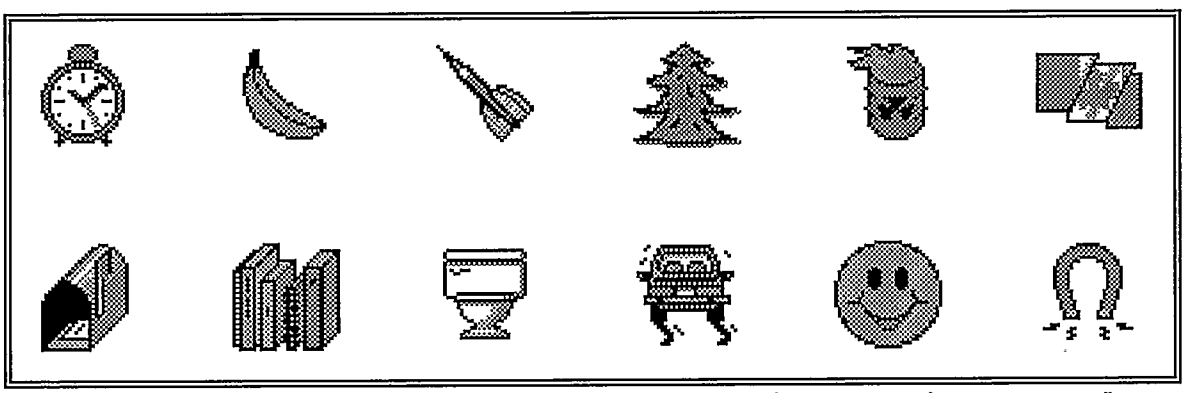

Figure 5.1 Black and white representations of icons used as stimuli.

A Visual Basic program was written to present the stimuli and collect and record the subjects' responses,

response latencies, and reaction times to the attention probe. The program was run under Microsoft Windows 3.0.

Procedure

The main learning task used a variation of the paired-associate paradigm. The subject's task was to learn, through trial and error, which key was associated with each picture. As noted above, the intention was to use the results of this experiment as the basis for a simulation of similar learning in a back-propagation network. This learning task was structured to closely approximate the procedure used to train such a network. With the back-propagation learning rule, a single learning trial consists of presenting a stimulus, allowing the network to generate a response, then presenting the correct response so the network can calculate the amount of error in its response and make appropriate adjustments to the connection weights. To approximate this procedure, in this experiment, the subject was presented with an item, made a response, and then was presented with the correct response. Unlike a traditional paired-associate learning task, there were no learning trials where all of the stimulus-response pairs were presented together. The subject both studied and responded to each item before the next item was presented.

Specifically, the experimental procedure proceeded as follows. At the beginning of each trial, one of the pictures (randomly selected without replacement) was

presented and the system waited until the subject pressed one of the four keys. The system then displayed both the picture and the correct key for a one-second study period. The subject was given no specific feedback about the response they had made. The program simply displayed what their response should have been.

The same procedure was repeated twelve times -- once for each picture. To facilitate comparisons with the connectionist network simulations to be presented later, each such block of twelve presentations will be referred to as one *epoch*. After each epoch, the system immediately began another epoch using the same twelve pictures but in a new random order.

After the first epoch, a reaction time task was interleaved with the learning task. On approximately every third item the study period was interrupted by a reaction time probe. The probe consisted of the entire screen going blank. When the probe occurred, the subject was to press the space bar as quickly as possible. As soon as the space bar was pressed, the screen was restored and the interrupted study period was restarted. Four of twelve items were probed each trial with the four items determined by random selection without replacement.

The system continued from one epoch to another for a total of twenty minutes. The intention was that the subjects would continue with the task until responses become

automatic. Pilot studies conducted during the beta-test phase of the software development suggested that twenty minutes was more than sufficient to take most subjects well past the point of mastery. The twenty minute experimental session was followed by a computer-administered questionnaire to collect each subject's age, gender, handedness, and a self-report of previous academic achievement (grade 9 Math mark).

Results

For each item, the system recorded the latency between the onset of the picture and the subject's key press as a measure of amount of learning for that stimulus. The system also recorded whether the subject's response was correct or incorrect. For each reaction-time probe, the computer recorded the subject's reaction time as a measure of attention during the study period. As mentioned above, only one quarter of the latency measures were accompanied by a reaction time probe measure. The full data set consisted of almost five thousand latency measurements on twelve subjects across as many as three dozen epochs each.

Two measures of the amount of learning were recorded: number of correct responses on each trial and the average latency of responses for each trial. The number of correct responses is a relatively coarse measure of the amount of learning. The latency measures were collected to provide a

more precise measure. The subjects were instructed to respond as quickly as possible. To the extent that subjects did follow these instructions, the average latency for each trial should show a negative correlation with the number of correct responses for that trial. Figure 5.1 shows these correlations for all twelve of the subjects sorted by the magnitude of the correlation.

Table 5.1

Correlation between number of correct responses on a trial and average latency of responses for that trial sorted by magnitude of correlation.

Subject #	Correlation	
119	**-0.83**	*
121	**-0.80**	*
110	**-0.65**	*
112	**-0.61**	*
115	**-0.53**	*
106	**-0.44**	*
100	**-0.34**	*
109	**-0.34**	*
108	-0.22	
104	-0.22	
107	-0.17	
216	**0.45**	*

* $p < 0.05$

For eight of the subjects, there was a significant negative correlation ($p < 0.05$) between the latency measures and the number of correct responses on each of the trials. This indicates that, as these subjects mastered the items, they responded more quickly when tested. Three of the other subjects did not show a significant correlation. This would suggest that either they did not follow the instructions to respond as quickly as possible or that they did not reach a significant level of mastery in the twenty minutes spent on the task. The positive correlation for one subject indicates that they actually began to respond more slowly as they mastered the material. In fact, inspection of the raw data for that subject suggested that this subject was not really attending to the task at all. In any case, this data suggests that latency is an appropriate measure of learning for only eight of the twelve subjects so the remainder of the results presented here are based on just those eight subjects.

For many subjects, both measures of learning (correct responses and latency) began to degrade during the latter part of the task, well after mastery had been reached. This was not entirely unexpected since the procedure for this learning task was designed to take the subject past the point of mastery. Consequently, no criteria were set which would allow a subject to stop once they had learned all of the items. The measures collected from these subjects,

along with feedback from pilot subjects, suggest that the subjects eventually passed some point of persistence and were, consequently, not working as diligently during the latter part of the task.

For individual subjects, the average and the variance of the scores for each trial gave some indication of the point at which persistence began to fade. With most subjects, the variance tended to decline through the first two-thirds of the epochs. After this, it began to vary widely. The epoch where the variance began to dramatically increase was used as a clipping point. With subjects for whom this point was not entirely clear, additional data points were retained to provide as conservative an estimate as possible of the point where persistence began to decline. The amount of data "clipped" in this manner varied from subject to subject. In several cases it amounted to only a few epochs, but in one case just over half of the epochs were clipped. Over all, approximately three-quarters of the epochs were retained. Since the subjects were required to continue with the experiment for twenty minutes even though many of them had reached mastery long before then, the objective of this clipping was just to discard the measurements taken after mastery had been achieved.

The data also showed a small number of extreme outliers at seemingly random points. These may have been due, for example, to the subject not pressing the key hard enough to

register. The few moments it would take to realize that their response had not registered would push the latency measure to an extreme value.

To identify and eliminate the extreme outliers, the largest scores for each subject were examined. In cases where the largest score seemed a "reasonable" amount greater than the second largest score all data was kept. In some cases, however, the largest score was as much as twice the size of the second largest score. In these extreme cases, the abnormally large scores were dropped. In all, nine extreme outliers were discarded. These nine scores represented only about one half of one percent of the scores under consideration and the reduced data set is more coherent and more readily interpreted.

One of the objectives of this experiment was to provide a basis for comparison between human subject performance and the performance of a connectionist simulation on a similar task. In order to be able to suggest that the learning in these two, very different situations is comparable, it is necessary to make qualitative comparisons of the way in which learning progressed for both. As such, summary statistics of the human learning would not be sufficient. Instead, graphical representations of the subjects' performance were developed for later comparison with similar graphs to be based on the performance of the simulations (see Chapter 6).

The sets of graphs presented on the next two pages (Figure 5.2 and Figure 5.3) show two views of the learning for each of the eight subjects individually -- one based on the number of correct responses and the other based on the latency measures. Each graph presents the actual measures for each trial with a broken line and a running average of these values with a solid line. The running average was based on the five values immediately adjacent to each point and is presented to more clearly show the learning trends in the data. Each subject is graphed separately because there were large differences in the magnitude of the learning measures between subjects and it is the qualitative nature of the learning trends which is of interest, not the speed of learning.

The first set of graphs, depicting the number of correct responses per trial (Figure 5.2), clearly indicates an increase in learning for seven of the eight subjects. Most of them show a typical positively decelerating learning curve but a few of them seem to be just slightly sigmoidal in that little learning seems to take place for the first few trials. Because of the "discovery" nature of this task, it is not surprising that some subjects required a few trials to become comfortable with the task.

As one would expect from the correlations presented in Table 5.1, the graphs based on the average latency of responses in each trial (Figure 5.3) mirror those based on

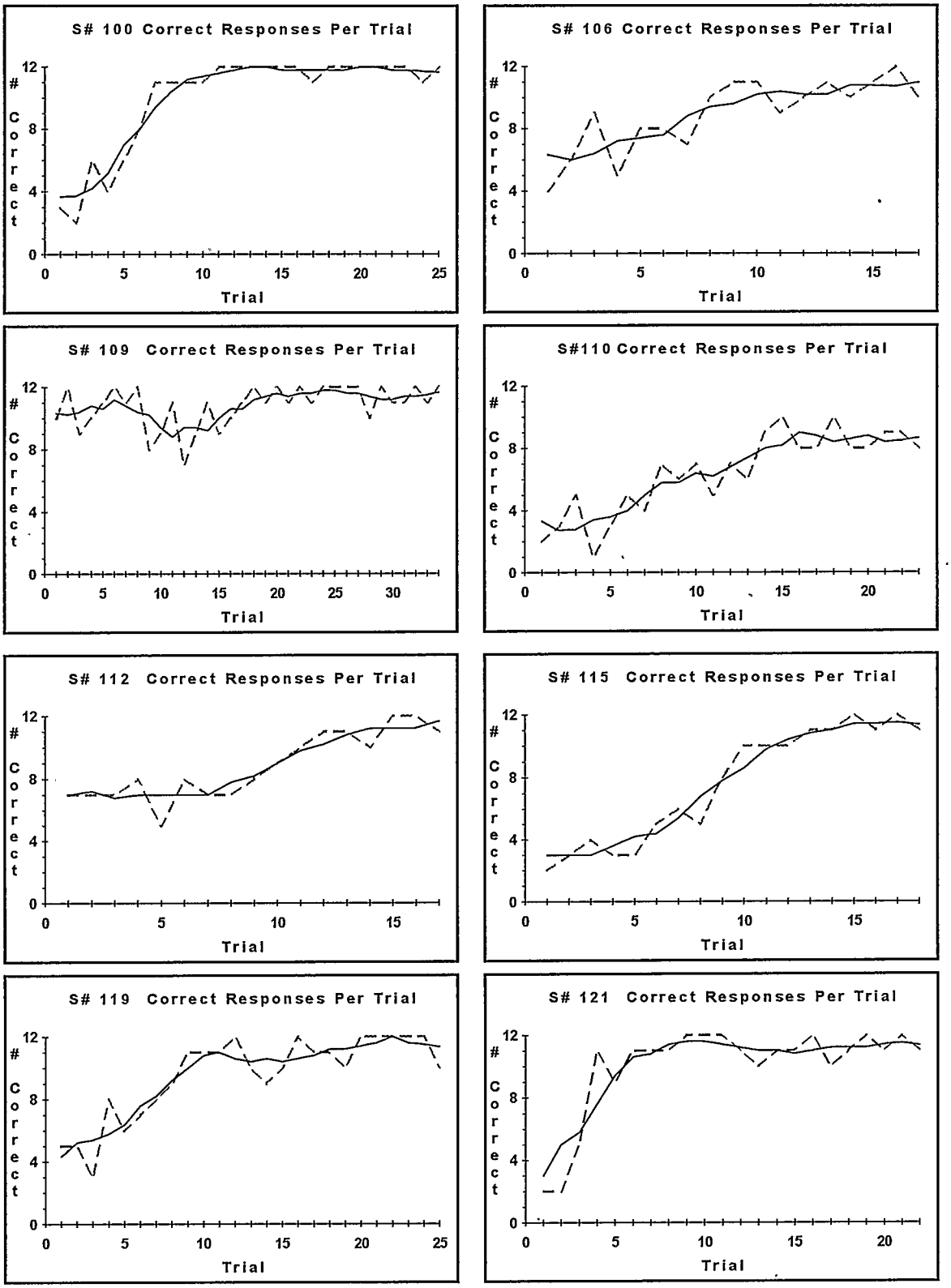

Figure 5.2 Raw (broken) and smoothed (solid) learning curves

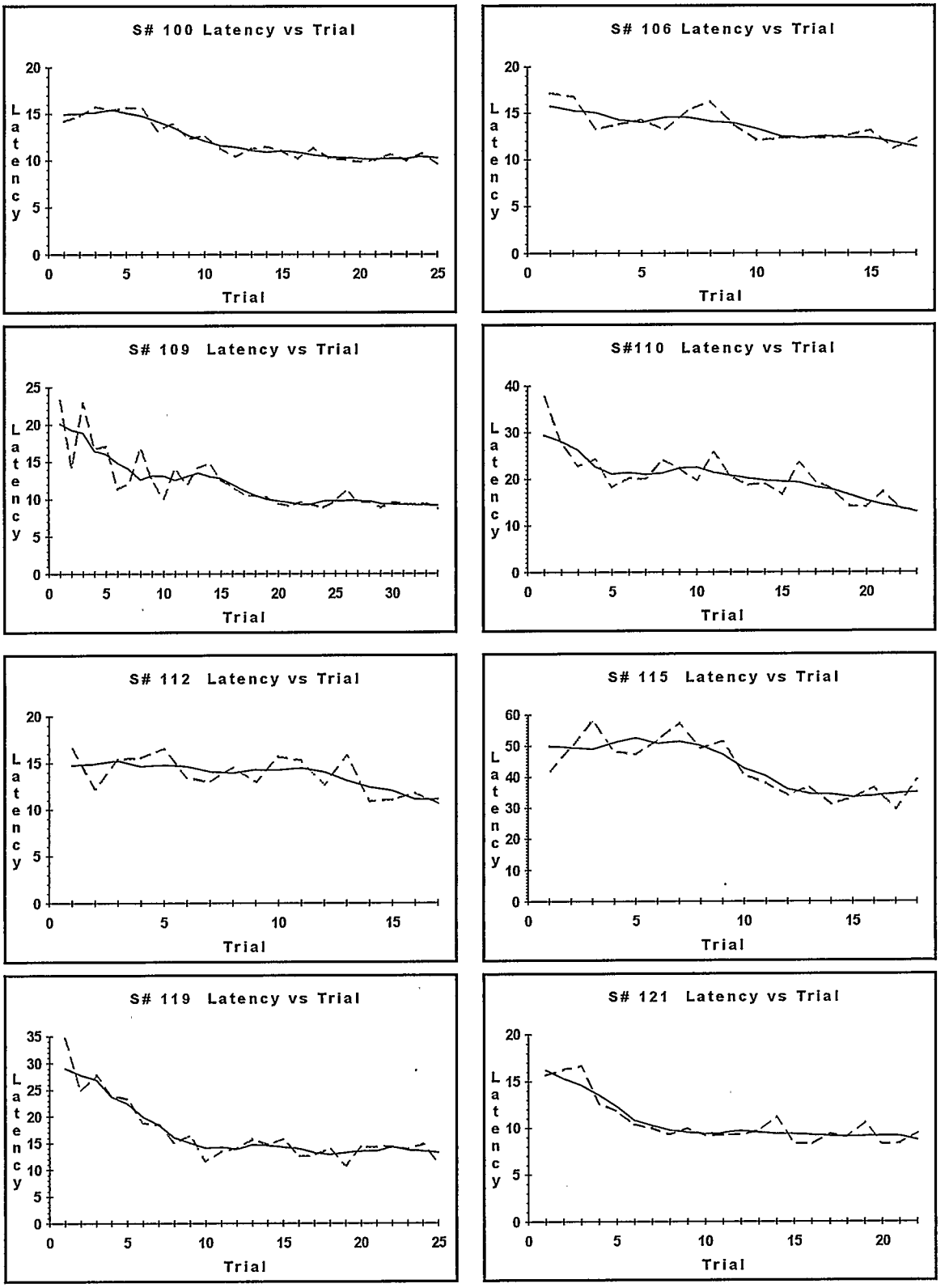

Figure 5.3 Raw (broken) and smoothed (solid) latency curves

correct responses to a large extent. The subjects showed considerable variability in the magnitude of both the latency measures and the number of trials. For subjects with large latency measures, this had the effect of compressing the learning curve into a narrower region of the graph. This makes comparisons between subjects difficult but, at this point, it is the shape of the curve, more than the magnitude of the values, which is of interest. Most of the latency curves are negatively decelerating and, as with the curves of the number of correct responses, some show a sigmoidal tendency.

The two sets of graphs provide evidence that learning has occurred for at least these eight subjects. In addition, the corresponding trends in the two sets of graphs support the correlations in Table 5.1 in suggesting that the latency of responses and the number of correct responses are both appropriate measures of the amount of learning. These graphs will serve as the basis for qualitative comparisons between human learning and the performance of the simulations presented in Chapter 6.

The results presented above give a picture of the nature of the learning taking place for this task. However, a more important issue for this investigation is the way in which attention varied as this learning occurred.

The first possibility investigated here is that attention (as measured by probe reaction times) changes as a

function of amount of learning (as measured by response latencies). Seven of the eight subjects showed a significant positive correlation between latency and probe reaction time on the items to which they responded correctly (Table 5.2). Since the response latencies are a negative measure of learning and the probe reaction times are a positive, if indirect, measure of attention, these results support the suggestion that attention is declining as learning proceeds.

Table 5.2

Correlation between average probe reaction times on a trial and average latency of responses for that trial.

Subject #	Correlation	
119	**0.46**	*
121	**0.75**	*
110	**0.64**	*
112	0.11	
115	**0.52**	*
106	**0.48**	*
100	**0.54**	*
109	**0.45**	*

* $p < 0.05$

These correlations suggest that there is a relationship between learning and attention but, in order to model this relationship, a more precise mathematical description is required. In an attempt to determine this mathematical description, both linear and sigmoidal curves were fitted to the points obtained by plotting the latencies versus the reaction times for each subject. The resulting r values are presented in Table 5.3 below.

Table 5.3

r values of linear and sigmoidal curves fit to plots of the latency versus reaction time measures.

Subject #	Linear	Sigmoidal
100	**0.67** *	**0.91** *
106	**0.58** *	**0.63** *
109	0.45	**0.50** *
110	0.42	0.45
112	0.22	0.41
115	0.36	**0.51** *
119	**0.47** *	**0.53** *
121	**0.58** *	**0.77** *

* $p < 0.05$

As one would expect from the correlations in Table 5.2, many of the linear curve-fits produced significant r values

but, in all cases, the sigmoidal curve produced a larger r value and more of the values were significant. These larger r values suggest that a sigmoidal curve produces a better fit with this data.

Although the above results suggest that attention is a sigmoidal function of amount of learning, a second possibility is that attention simply declines with time on task. Since latency is also decreasing as the subjects learn, this would result in just such a positive correlation between the latency and reaction time measures as appear in Table 5.3 above. Graphs of the change in probe reaction time measures over time are presented in Figure 5.4 below.

As with the learning curves in Figures 5.2 and 5.3, the actual values are presented as broken lines and a running average over five adjacent points is presented as a solid line. Although most of the graphs do seem to suggest a negatively decelerating relationship between probe reaction time and trial number, these curves are not as clear as the learning curves. However, linear regression analyses on each of the sets of data do suggest that, at least, attention is declining. Table 5.4 below presents the slope and the standard error for each of these analyses.

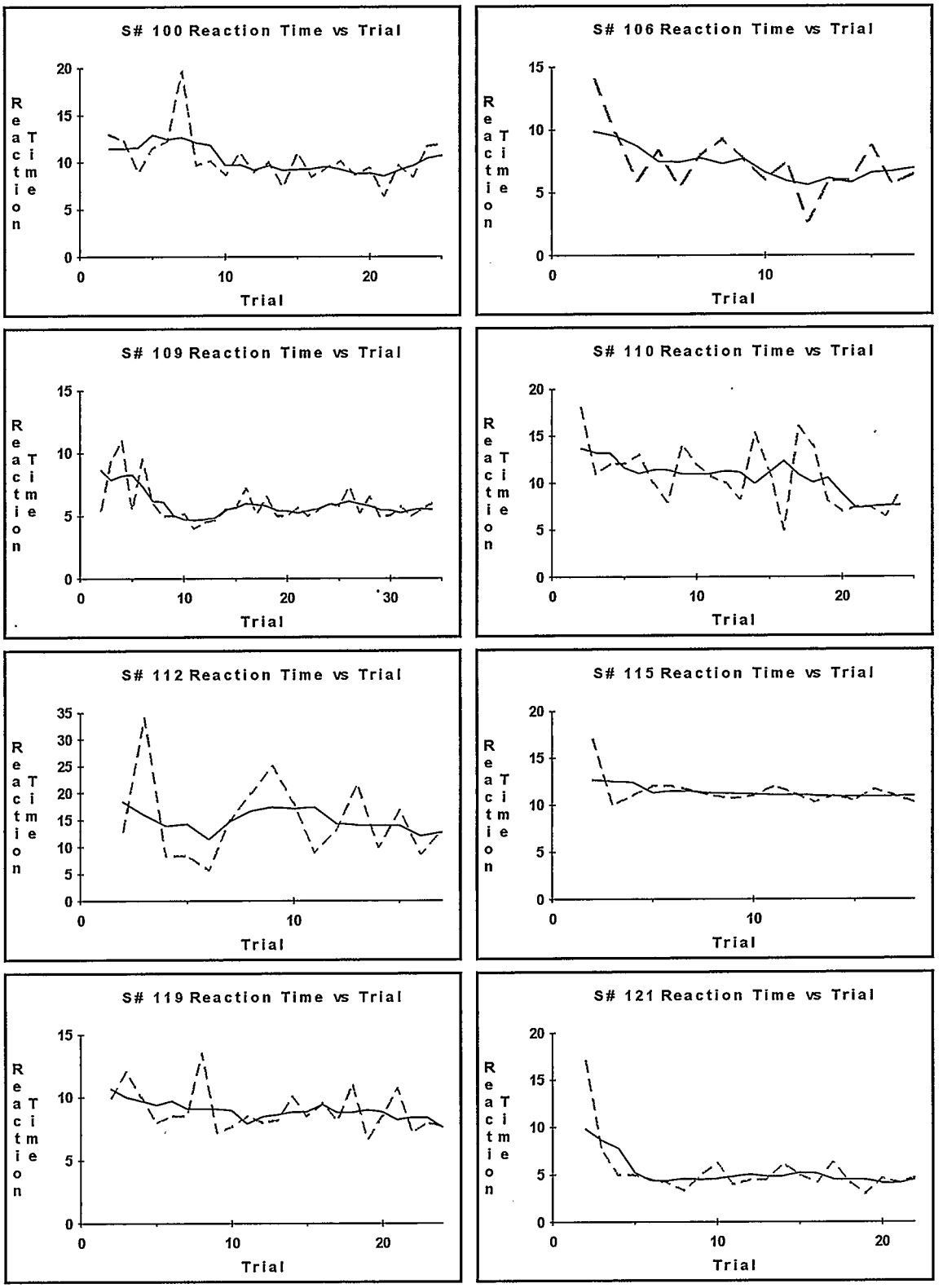

Figure 5.4 Individual reaction time curves.

Table 5.4

Slope of the regression line for the relationship between reaction time versus trial number and the associated standard error.

Subject #	Slope	Standard Error
119	-0.080	5.017
121	-0.202	4.542
110	-0.304	7.134
112	-0.365	17.058
115	-0.136	6.164
106	-0.273	5.132
100	-0.045	8.730
109	-0.209	6.687

The fact that all eight linear regressions produced negative slopes does suggest that the probe reaction times are decreasing over time. However, the large standard errors for most subjects suggests that a simple linear relationship between attention and time is only a rough approximation and that the actual relationship is more complex.

Conclusions

The objectives of this study are mostly concerned with changes in attention as subjects learn. Prerequisite to these considerations is establishing that learning has occurred and presenting a description of the increase in learning as the task proceeded. The two measures of

learning, number of correct responses per trial and average latency of responses for each trial, indicate that, for at least eight of the subjects, learning did occur. Graphical representations of the data describe typical negatively decelerating learning curves with some suggestion of a sigmoidal curve on early trials. In the next chapter, these curves will serve as a basis for qualitative comparisons with the performance of connectionist network simulations of this learning task.

With learning established for the eight subjects, their probe reaction times provided a suitable indirect measure of the amount of attention they were devoting to the learning task. Although these results do not clearly show the exact mathematical relationship between learning and attention, they certainly helps substantiate the distinction between automatic and controlled processes as one of degree of attention and help characterize the transition from automatic to controlled process due to learning as a gradual not a sudden one.

It would have been valuable to establish a more precise mathematical description of this continuously declining relationship between learning and attention. Unfortunately the measures obtained were not sufficiently regular to conclusively establish the mathematical function which best characterizes this relationship. However, the results do suggest two candidates: attention decreases as a sigmoidal

function of learning and attention decreases as some, probably non-linear, function of time on task.

Interpreting the inconclusive results shown here within the theoretical context of limited attentional resources can further constrain these possibilities. If attention is to be represented by a continuously declining function, there are theoretical limits to the nature of that function. To the extent that attention is a limited cognitive resource, it is conceivable that attention could decline to zero, but it is not meaningful to suggest that attention could ever be negative. Declining attention could thus not be appropriately modelled by either a negatively accelerating or a linear declining function. Of the remaining alternatives, parsimony would suggest either a negatively decelerating quadratic function or a sigmoidal function.

The next chapter will present the results of a series of simulations based on this range of possible functions. The objective of the simulations was to determine which of these possibilities results in the most appropriate connectionist network model of the learning situation presented here.

CHAPTER SIX: CONNECTIONIST NETWORK SIMULATIONS

This chapter describes a series of connectionist network simulations of human learning using a back-propagation learning rule. The objective was to explore the use of a temperature term to more accurately model attention in human learning. The temperature term is the one described in Chapter 3, attention is as defined in Chapter 4, and the human learning being modelled is that which took place in the experiment described in Chapter 5. Specifically, the objective is to model the decline in attention which accompanies paired-associate learning.

Modeling Human Cognition

Before discussing the specific performance modelled here, there are a number of general issues associated with the modeling of human cognition which should be addressed. These issues have to do with the relationship between a simulation and a model, assessing qualitative rather than quantitative performance, the granularity of investigation, and alternatives to empirical hypothesis testing.

Simulations and Models

Gluck and Bower (1988) present two general methodologies for using implementations of connectionist networks to model human cognition. One involves selecting some aspect of human performance and constructing a network

to perform the same task in a manner such that the "major regularities and salient phenomena" are preserved. The second methodology focuses on a specific experimental paradigm and builds a network whose performance will predict human performance within that paradigm. The simulations presented here represent the first of these methodologies.

The human learning experiment presented in Chapter 5 clearly indicated that attention declines as learning proceeds. The results did not show precisely what the mathematical relationship was, but it did constrain the possibilities and point out several possible approaches for these simulations. The different simulations presented here each model a different one of these possibilities. The objective was to see which one most faithfully preserved the 'major regularities and salient phenomena' of the human learning experiment.

Performance

Some computer simulations of specific connectionist models represent attempts to solve practical problems in research areas that are usually classified as artificial intelligence. The objective of this type of research is to find an optimal solution to the problem. If human performance suggests refinements to the model, they are useful only if they improve the quantitative magnitude of the simulation's performance in that specific problem

domain, but it is the magnitude of the performance, not its qualitative aspects, which is at issue.

However, other implementations of connectionist models, including those presented here, are more concerned with accurately simulating human performance in a specific domain. Refinements to the model are useful only if they bring the qualitative performance of the simulation closer in line with the human performance. The simulations presented here represent alternate implementations of one specific refinement (adding temperature to a back-propagation network). In evaluating the performance of the various simulations, the most useful simulation will be the one whose performance is qualitatively the most human (i.e., the most like the performance of the subjects in the human learning experiment) regardless of its relative quantitative performance.

Sensitivity Analysis

Although part of the appeal of connectionist models is the simplicity of their processing mechanism, that mechanism includes a large number of parameters (number of nodes, number of layers, learning rule used, momentum, temperature, etc.), each of which is subject to refinement. Although different connectionist models often focus on refinements to one specific parameter, evaluating the effect of a single parameter on even one aspect of the model's performance (e.g., learning) can be very difficult. Schneider (1988)

identifies two general approaches to evaluating a model's performance: "parameter estimation" and "sensitivity analysis", and argues in favour of the latter.

Parameter estimation is the technique most commonly used by psychologists to evaluate traditional cognitive models. This approach tries to find the values for the parameters which will yield the best results: in this case the best quantitative fit between model performance and human performance. Parameter estimation is largely "results oriented" and may be relatively insensitive to interactions between parameters.

Sensitivity analysis evaluates the behaviour of the network across the full range of meaningful values for the parameters (e.g., the use of several different mathematical functions to vary temperature in the simulations presented here), and describes the interactions which result.

> Sensitivity analysis identifies the interactions of variations of parameters to determine where changes in components have a large impact on the system's performance (Schneider, 1988, p. 282).

Schneider presents a number of reasons why sensitivity analysis is especially appropriate for evaluating connectionist models. Connectionism as a modeling technique is still relatively new and many possible parameters have yet to be identified. Connectionist models use nonlinear functions and these can obscure the impact of variations in parameters, especially as they approach boundary conditions.

Parameters which are itemized as features of one model may, in fact, have little effect on its performance, with the result that other models which appear different may be substantially the same. In his conclusion, Schneider argues that:

> Before an author presents an extended discussion of the importance of a parameter, it is his/her responsibility to communicate the sensitivity of the system to that parameter (p. 283).

The simulations presented here show both the qualitative and quantitative sensitivity of the back-propagation learning rule to various implementations of a temperature parameter.

Hypothesis Testing

The traditional, empirical hypothesis-testing method based on statistically significant differences is most appropriate for quantitative comparisons between implementations of a specific, isolated parameter. However, as noted previously, there are arguments in favour of a more qualitative comparison of the overall performance of these simulations.

In the absence of specific, testable hypotheses there is no basis for tests of statistical significance. Instead, a more qualitative assessment of the performance of the various simulations was used. Even where hypothesis-testing might traditionally have been done, Loftus (1993) argues for a graphical approach to the presentation and interpretation of results over the more traditional hypothesis testing

approach based on tests of statistical significance. A similar perspective seemed appropriate here.

These arguments are not intended to suggest that research involving connectionist models should avoid specific hypotheses -- only that the acceptance or rejection of the hypotheses could be based on something other than a more conventional statistical test of significance. The objective behind the simulations presented here was to identify which of several possible implementations of temperature in the back-propagation learning rule serves as the best model of attention in human learning. In the end, comparisons between the different simulations were based on graphical representations rather than statistical tests involving the discrete effect of a single manipulation.

Method

In all, six simulations were compared. The implementation details of all simulations were basically the same as the initial back-propagation network described in Chapter 3 with three exceptions: the length of time they were run, performance criteria measured, and the manner in which temperature was adjusted.

Running Time

The simulation in Chapter 3 was run until the network "mastered" the material by reaching a specific performance criterion. The intention of the human learning experiment

in Chapter 5 was to take the subjects to the point of automaticity -- well past simple mastery. Some characteristics of the human learning did not become evident until long after mastery had been reached. To avoid overlooking similar characteristics with the simulations presented here, they were each run for a specific number of epochs instead of stopping once a specific performance criterion was reached. On average, the simulation in Chapter 3 reached criterion after 83.1 epochs. Each of the simulations presented here were run for 200 epochs.

Performance Recorded

Learning in a back-propagation network is defined as error reduction through gradient descent (see Chapter 3). The most direct measure of learning in such a network is the amount of error still remaining at the end of an epoch. The simulation in Chapter 3 ran until a specific learning criterion was met but this criterion was actually expressed in terms of the amount of error remaining. The performance measure recorded for that initial simulation was the number of epochs required to reach criterion. In contrast, all of the simulations presented here were run for a fixed number of epochs, so a similar measure of learning performance would not be appropriate. Instead, the average amount of error for each epoch was automatically recorded as the simulation ran, and these values used to produce graphs of each simulation's performance.

Adjusting Temperature

The simulation in Chapter 3 decreased temperature by a specific amount each epoch (i.e., temperature was an inverse, linear function of elapsed time). In the human learning experiment in Chapter 5, learning increased with time and attention decreased with time. Within the context of the literature on learning and attention presented in Chapter 4, this human learning data suggests that attention decreases with learning. It is possible, however, that attention merely decreases with time-on-task, and is more or less independent of the amount learned. Three of the simulations presented here continued to reduce temperature based on elapsed time (epochs), but the other three based their temperature on the amount of learning as measured by remaining error.

Although the human data clearly showed a decline in attention it failed to provide a precise mathematical description of the nature of that decline. However, again within the context of the learning and attention literature discussed, it did suggest several possibilities: to model attention, the decline in temperature should be based on either a linear, quadratic, or sigmoidal function.

Linear Function. The simplest possibility is that the decline of attention is linear. This is not a likely possibility because a linear decreasing function will eventually reach zero and continue into negative values,

but it is included here for completeness. There is nothing about human cognition which suggests that negative attention values would be appropriate and, in any case, learning would stop in a back-propagation network once the temperature reached zero. For comparison with the earlier simulation, two of the simulations presented here used a simple linear transformation to decrease temperature based on either elapsed epochs or remaining error respectively. To ensure at least a minimum amount of learning each epoch, a minimum "floor" value was imposed.

Quadratic Function. The data from the human experiment suggested that a negatively decelerating quadratic transformation was a more likely approximation of the human data than a linear function. Because the curve is decelerating it can be set up so that it never reaches zero, and thus there is no need to impose a floor. Two of the simulations used a quadratic transformation, again with one based on time and one on learning.

Sigmoidal Function. The early automaticity literature argued for a threshold function to represent the change in attention. The data from the human learning experiment presented in Chapter 5 does not support this but at least some of the data would be consistent with a sigmoidal transformation. In fact, depending on the scale, offset, and slope, a sigmoidal function could approximate either a linear, quadratic, or threshold functions (see Figure 6.1).

A sigmoidal transformation with a medium slope was used to adjust temperature in the remaining two simulations.

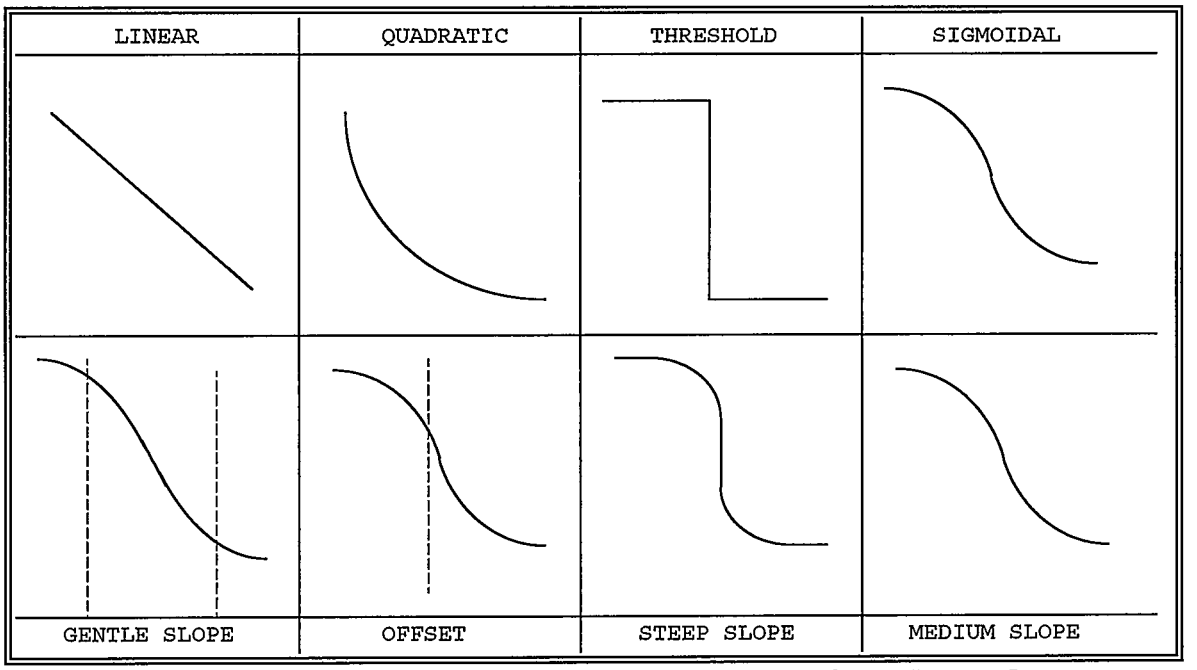

Figure 6.1 Sigmoidal Functions Scaled, Clipped, and Adjusted to Approximate Linear, Quadratic, and Threshold Functions

In all, the six simulated "conditions" for this simulated "experiment" fill a two by three design matrix representing the transformation function used (linear, quadratic, or sigmoidal) and the basis for the transformation (elapsed epochs or remaining error).

Each time any connectionist network simulation is run it produces slightly different results. This is analogous to the variability of responses made by human subjects within the same condition of an experiment. To ensure that the results obtained from each of these simulations

reflected the general characteristics of the specific model which represents temperature in that way, each of the six simulation was run twenty times to produce data for twenty simulated subjects.

Results

Each of the six graphs presented on the next three pages depicts the learning performance of all twenty of the simulated "subjects" in one of the six "conditions". Graphing each subject's individual performance (instead of some measure of central tendency) shows the variability of responses as well as the general trends.

Although 200 epochs were recorded for each subject, after the first 100 epochs there was very little difference in the performance of any of the subjects either within or between conditions. Because of this, only the first 100 epochs are included in the graphs. This accentuates the trends in the early part of the learning curves.

For comparison with the human subject data presented in Chapter 5 and with the theoretical constructs being modelled, the graphs are labelled as representing learning over time. In fact, time was measured in epochs where one epoch represents one presentation of each stimulus (analogous to a single learning trial), and learning was inferred from measures of the amount of error remaining in the network at the end of each epoch.

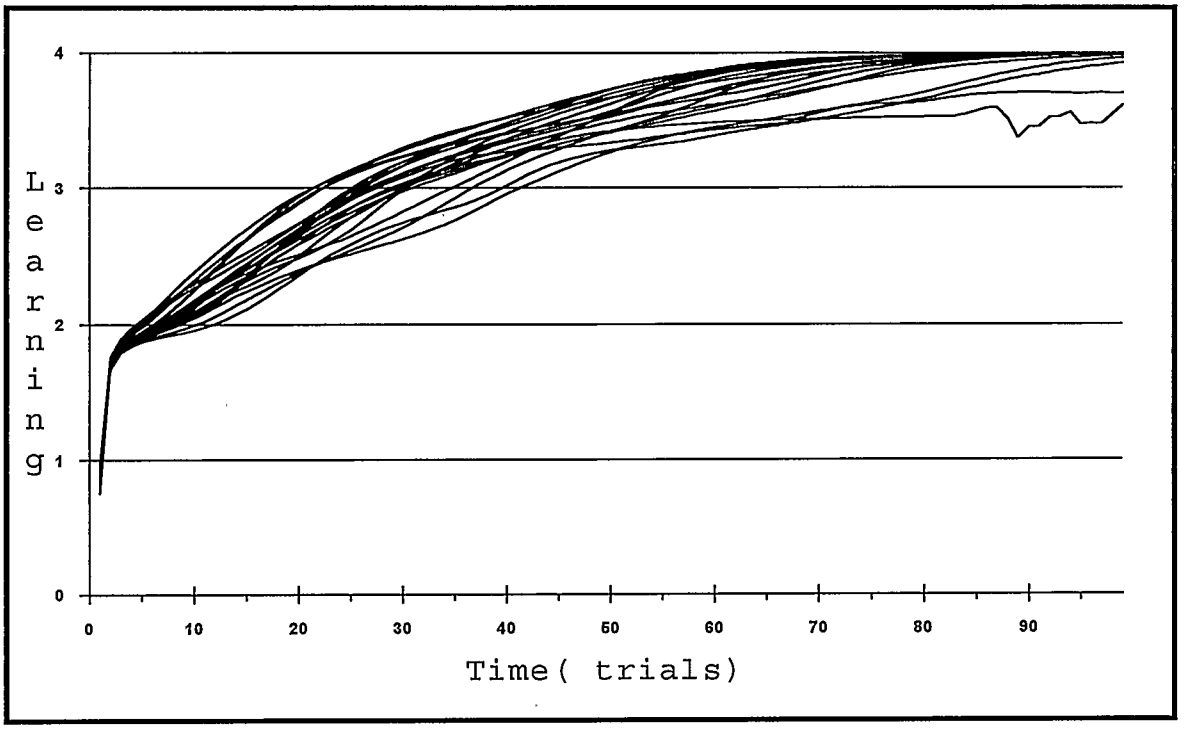

Figure 6.2 Temperature as a linear function of epochs.

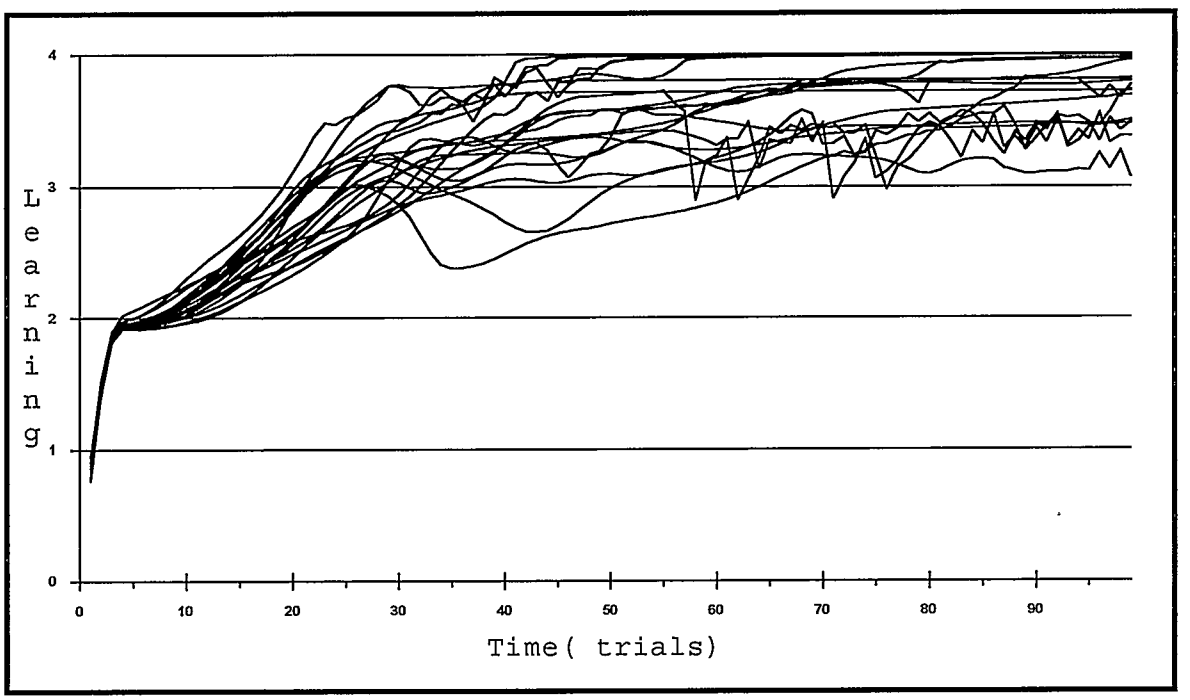

Figure 6.3 Temperature as a linear function of error.

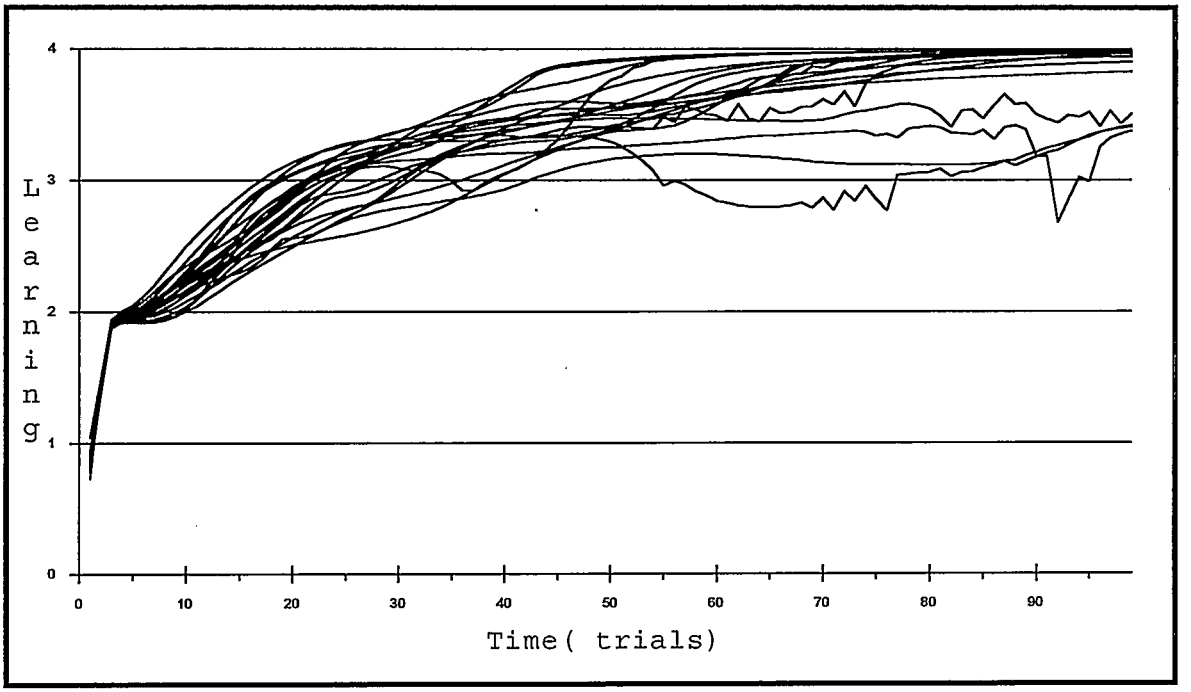

Figure 6.4 Temperature as a quadratic function of epochs.

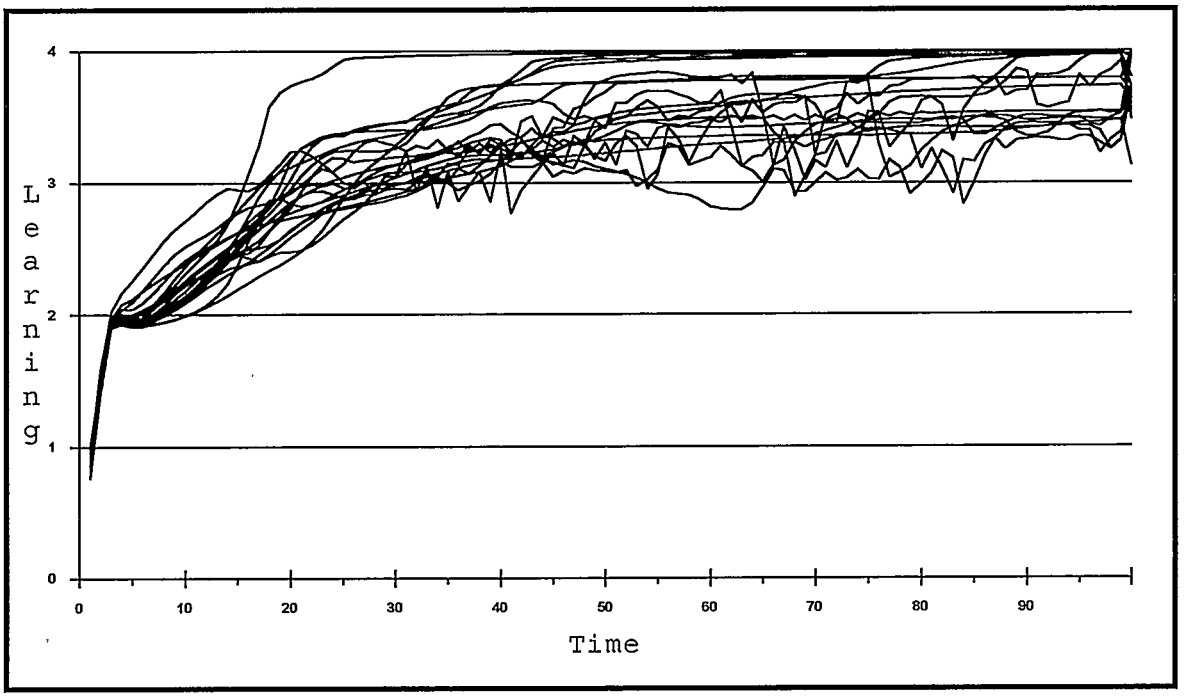

Figure 6.5 Temperature as a quadratic function of error.

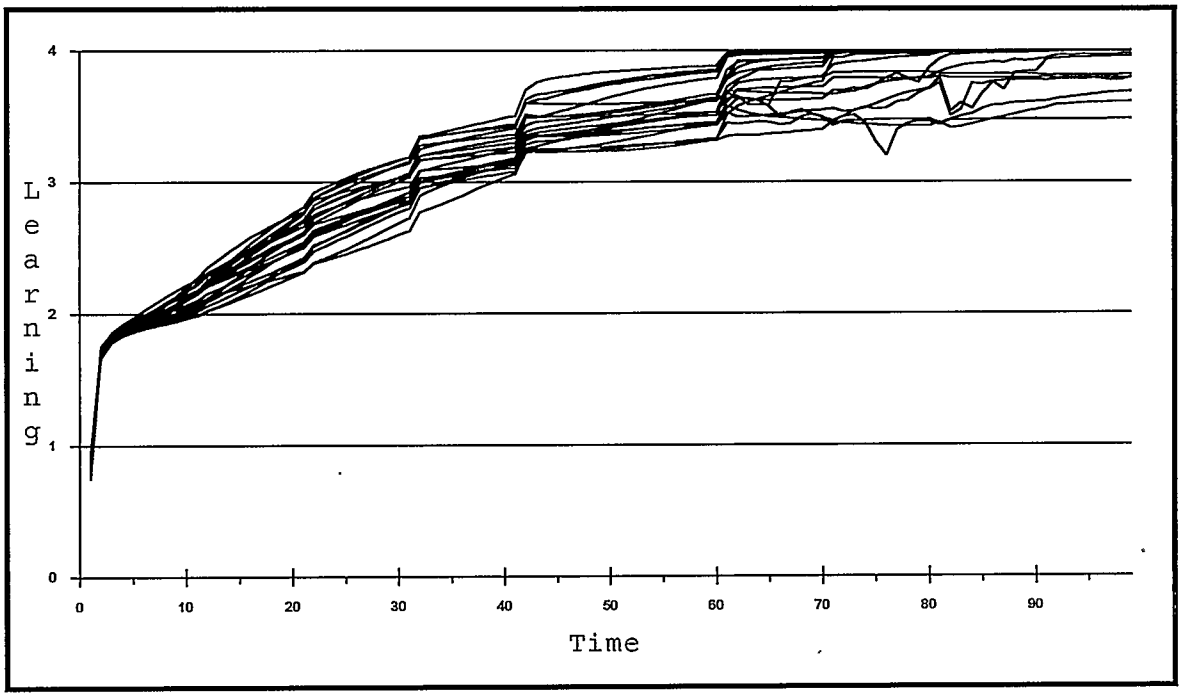

Figure 6.6 Temperature as a sigmoidal function of epochs.

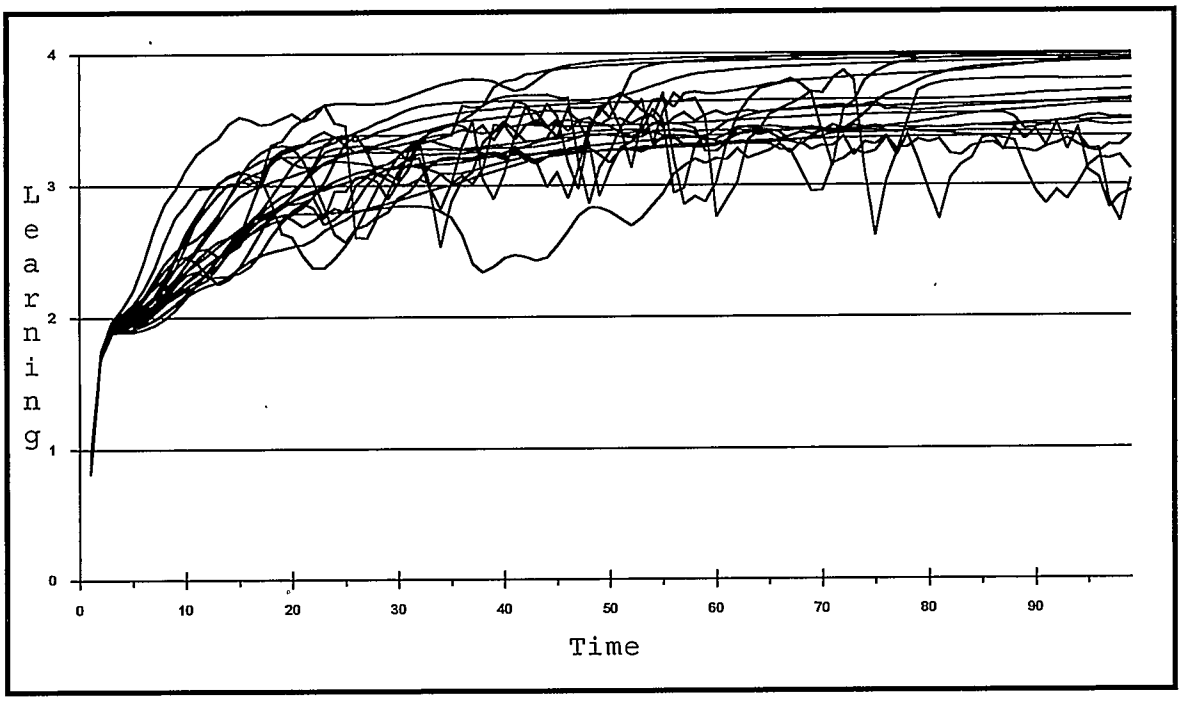

Figure 6.7 Temperature as a sigmoidal function of error.

For seven of the subjects in the human learning experiment there was a significant correlation between the series of response latencies and reaction times to the attention probes. The learning performance of these subjects formed the basis for comparisons with the performance of the simulations. The latency measures for these subjects produced distinct learning curves (see Figure 5.3). Figure 6.8 combines all seven learning curves in one graph to facilitate comparisons with the data collected from the simulations.

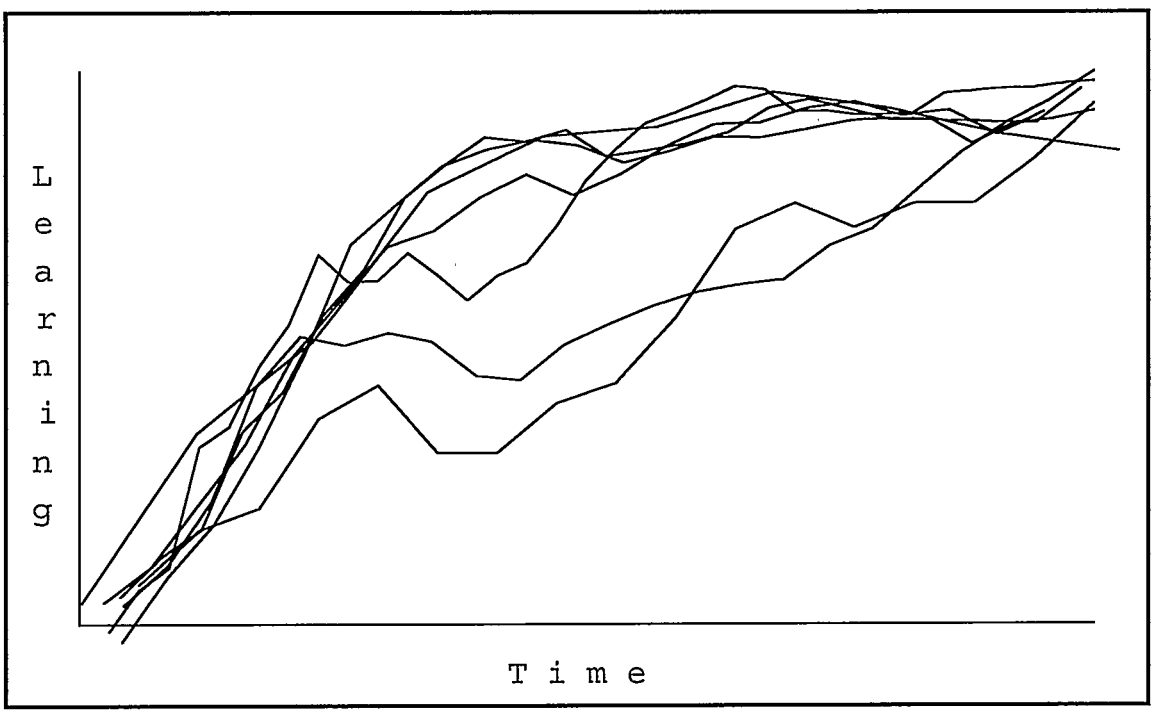

Figure 6.8 Learning curves based on inverse of average latency of response versus trial number.

The latencies were actually an inverse measure of learning, so the curves from Chapter 5 have been inverted

here to show a more "typical" learning curve. The "Time" axis represents the trials considered for each subject. Since this number varied considerably from subject to subject, no units are included for that axis. Instead, for all subjects, the curves were scaled so that the entire length of the axis represents all of the trials considered. The "Learning" axis represents the amount of learning shown by each subject and, again, no units are presented since the actual values varied considerably from subject to subject. Each curve was clipped and scaled to represent only the range of measures obtained for that subject over the trials considered. The origin represents the least amount of learning measured for the subject and the maximum value is near the top. Two of the seven subjects did not show appreciable learning on the first four and eight trials respectively. It appears that these subjects took longer to familiarize themselves with the learning task. Those initial trials are not represented here.

Discussion

As noted above, the objective of this simulated "experiment" was to find the mathematical function which would cause a changing temperature in a connectionist network to most accurately model the change in attention in the human learning experiment. Six alternatives were explored here and the results presented in the six sets of

graphs (Figures 6.2 through 6.7). The one which represents the best model of the human performance (represented by Figure 6.8) was selected through qualitative comparisons between the sets of graphs. There is no clear "winner" but a process of elimination based on comparisons of the nature of the simulations' performance and some artifacts in that performance give some indication as to which of the six is a better choice for modeling the effects of attention in human learning. Comparisons of the general performance of the simulations which varied temperature based on time versus those which based temperature on learning and a consideration of linear versus quadratic and sigmoidal transformations also provide arguments in favour of specific simulations.

Performance

Those factors which optimize computer performance are rarely the ones which optimize human performance. It is therefore not surprising that the simulation with the best performance is **not** the one which best models the human performance.

Quantitatively, the network with superior performance is the one in which temperature is linear with time (Figure 6.2). Early learning increases at a faster rate and the total amount of learning after 100 epochs is higher than with the other simulations.

Qualitatively the learning curves for most "subjects" in this condition are much smoother than those in other conditions. In connectionist network terms this smoothness indicates fewer diversions into local minima. Again in connectionist terms, this condition probably represents maximum network performance, but maximizing network performance was not the goal of these simulations. Instead, the goal was to model human performance. In the human learning experiment, human performance was much less uniformly smooth so a simulation which produces smooth learning curves is likely not the best model of human performance.

Artifacts

The performance of three of the simulations resulted in graphs with distinctive artifacts which did not appear in the graphs of the human performance. The presence of these artifacts detract from the usefulness of these simulations as models of the human performance.

The simulation in which temperature was sigmoidal with time (Figure 6.6) seemed to learn in a series of rapid spurts followed by a plateau during which little or no learning occurred. This unusual performance is probably due to an unexpected mathematical interaction. The back-propagation learning rule used in all simulations incorporates a sigmoidal transformation to determine level of activation of nodes on the forward pass of input through

the network. Using a sigmoidal transformation to also adjust temperature may have caused an unusual interaction in these complex mathematical formulas resulting in the step-like plateaus shown in the graph. In any case, it would seem that a back-propagation network which adjusts temperature in a manner which is sigmoidal with time is probably not appropriate for modeling human performance, though it may have interesting implications for other neural-network applications.

The graphs of the performance of the simulation in which temperature was varied as a linear function of error (Figure 6.3) show a distinctive "saddle" in the early part of the graph. This suggests that the rate of learning slowed down for a dozen or so epochs. Although similar saddles appear in other conditions, it is very pronounced in this condition.

In addition to the saddle artifact present in Figure 6.3, the simulation which varied temperature as a quadratic function of epoch produced performance which was almost as smooth as that shown in Figure 6.2.

A Better Choice

In contrast, the performance of the simulations in which temperature is either quadratic or sigmoidal with learning (Figures 6.5 and 6.7) have several characteristics in common with the human learning data. The average performance is good and it conform closely to the positively

decelerating learning curves observed in most human learning situations (see Figure 6.8). As with actual human performance, the performance of individual simulated "subjects" is somewhat erratic both within each subject and between subjects in the condition even to the point where several "subjects" might be considered outliers.

Time Versus Learning

In general, the three simulations which varied temperature as a function of learning (error) instead of time (epoch) produced performance which more closely resembles the human performance. All three simulations based on time (Figures 6.2, 6.4, and 6.6) produced learning curves which were generally much smoother. The performance of different simulated "subjects" was much more consistent within each of these three "conditions", even to the point of consistently reproducing the artifacts mentioned above. Human performance in general has considerable variability and the performance of the human subjects on this learning task is no exception. Because they reflect a similar amount of variability, one of the three models which vary temperature as a function of learning is likely to be a better model of human learning.

Sigmoidal as Linear and Quadratic

As discussed in the Method section above, a sigmoidal function may bear a close resemblance to either a linear or a quadratic function depending on the slope of the function

and any offset imposed to constrain the range of values. For this project, the simulation which varied temperature as a sigmoidal function of error used a balanced function with an intermediate slope. On the other hand, if that condition had used a sigmoidal function whose shape more closely approximated a linear function, the performance of that simulation would likely have approximated that of the simulation which used an actual linear function. Similarly, if the sigmoidal function had been clipped to resemble a quadratic function, the performance would have been similar to that of the simulation which used a quadratic function. One would expect, then, that differences in simulations using these three functions would be mostly a matter of degree and that a sigmoidal function could be made to approximate the performance of either of the other two functions.

In fact, the learning curves of the three simulations which varied temperature as a function of learning are similar. The main difference between them is the relative presence of the "saddle" artifact mentioned above. This artifact is very pronounced with the linear function, noticeable reduced with the quadratic function, and almost unnoticeable for many of the "subjects" in the condition which used the sigmoidal function. This would argue in favour of the sigmoidal function for a model of human learning.

Conclusions

The purpose of this series of simulations was to show that adding a temperature parameter to a back-propagation learning rule in a connectionist model of human learning and causing that parameter to decline as the network learns will improve network performance and will do so in a manner which is similar to the effect of attention in human learning.

The results of these simulations are not conclusive but, of the transformations investigated here, it would seem that a network which adjusts temperature sigmoidally with learning will more closely model human performance and one which decreases temperature linearly with time will maximize network performance.

CHAPTER SEVEN: CONCLUSIONS AND IMPLICATIONS

This chapter begins with a summary of the conclusions reached throughout this research project. The summary is followed by suggestions for further, follow-up research on the interaction between learning and attention using variations on the same methodology. The chapter ends with a discussion of broader implications for research within and across several disciplines.

Summary of Conclusions

This research project proceeded in phases consisting of a computer-based simulation, an experiment on human subjects, and then a series of additional computer-based simulations. Each phase of the project resulted in specific conclusions but each phase was also based, to at least some extent, on the conclusions of previous phases. To preserve the flow of the argument, the conclusions reached at each phase of the project are presented in the chapter describing that phase. The following is a brief summary of those conclusions.

The discussion of connectionist networks and related research presented in the first two chapters led to the conclusion that such networks are useful for modelling human cognitive processes. Further, this discussion suggested that models of human learning would likely require a multi-

layer network with a sophisticated learning mechanism such as the back-propagation learning rule.

The third chapter suggested that performance in a back-propagation network was improved by including a "temperature" parameter which decreased the variability of each node's output signal as the network learned. Chapter 4 concluded that this temperature parameter might be used to model attention in human learning, and suggested a methodology for measuring attention in a learning situation.

The results of the human learning experiment presented in the fifth chapter supported the hypothesis that attention declines in some continuous manner as learning increases and suggested several possible mathematical descriptions of that decline. Each of these possibilities was used as the basis for a back-propagation network simulation of the human learning experiment and Chapter 6 reports the results. The overall conclusion reached was that, of the possibilities investigated here, the most appropriate model of attention in human learning is a back-propagation network with a temperature term which declines sigmoidally as learning increases.

Project-Specific Implications

As with any research, some aspects of this project worked well and others did not. The things which worked well have implications for future research in a variety of

areas but the things which did not work as well have equally important implications for further research in this area and related areas. This section will discuss some of the limitations of the human learning experiment and the computer-based simulations of that task with a view to improving future research.

The Human Learning Experiment

One difficulty with this experiment concerned the fact that the procedure relied on reaction times as a measure of learning and of attention. Although the subjects were instructed to respond as quickly as possible, there is some evidence that a number of them may not have been focusing on these instructions. This may have been due in part to their lack of maturity (they were all grade ten students) and, in part, because they were focusing on learning the items to the exclusion of everything else. Future studies using this methodology may have more success by drawing subjects from a more mature population. In addition, the software which administers the learning task could be modified to provide feedback on the subject's speed of response. If this feedback were provided in the form of an ongoing arcade-style score, it is more likely that subjects would be motivated to respond as quickly as possible.

A second difficulty with the experimental design had to do with the subjects' persistence. As mentioned in Chapter 5, initially it was not clear whether latency would decline

by any appreciable amount until mastery was almost reached. On that basis, this study was designed to continue to test subjects long after the point of mastery. In fact, the data from this study suggests that latency begins to decline almost as soon as subjects begin to learn. Future studies which use this methodology could avoid some of the problems with persistence by establishing some criteria for stopping shortly after mastery is reached.

Related to the issue of persistence is the rate with which the subjects learned. Since the objective of this experiment was to reach mastery, the task was made relatively easy by having only twelve pairs of stimuli to be learned. For many of the subjects, this meant that mastery was approached after only a few trials and, consequently, most of the attention measures were taken over a relatively small range of learning measures. Since it would seem that attention begins to decline early in the task, the number of items could be increased to produce a slower rate of learning. This should delay the loss of persistence. It should also increase the chances of determining a more precise description of the relationship between learning and attention by providing attention measures over a broader range of learning scores.

The conclusions presented in the first section of this chapter were reached largely on the basis of qualitative comparisons between the performance of the human subjects

and the simulations. If a more homogenous population (perhaps using testing software which implements the above suggestions) could produce a more consistent performance, it might be profitable to investigate techniques for making quantitative statistical comparisons between the graphs of the human subjects' performance and the performance of the simulations.

The Simulations

Arguments presented in Chapter 6 suggest that a sigmoidal function can approximate linear and quadratic functions as well. The specific function used for the simulations which adjusted temperature sigmoidally had a balanced shape with an intermediate slope. Of the three simulations which based temperature on the amount of error, the one with the sigmoidal function seemed to be a slightly better model of the human performance than the ones with the linear function and the quadratic function. It is possible that a sigmoidal function of a different shape would produce an even better fit with the human performance. Further research using simulations with sigmoidal functions could investigate a range of slopes and offsets.

In general, the performance of the simulations which based temperature on error was more like the human performance than the simulations which based temperature on epoch number. The major exception to this was the presence of a saddle-like plateau in the early learning trials of

some simulated "subjects". Further investigation of this "saddle" artifact is necessary, especially in comparison with the slightly sigmoidal performance in the early learning trials of some of the human subjects.

Broad, Interdisciplinary Implications

As mentioned in the preface, this project is based on research from three disciplines: Education, Psychology, and Computer Science. The rest of this chapter will discuss the implications this research has for each of these areas and present some suggestions for further research.

Computer Science

Because of the sophisticated technology required to implement connectionist networks, this area of research owes much to computer science in general and the area of artificial intelligence in particular. However, computer science is not generally concerned with modeling human performance, and some aspects of the technology it produces are more appropriate for such models than others. One such technology is simulated annealing.

There are some similarities between the use of temperature in a back-propagation network and simulated annealing. In general, simulated annealing is used for combinatorial optimization in a wide range of domains (Laarhoven & Aarts, 1987; Vidal, 1993). One specific application is its use in neural networks called Boltzmann

machines (Hinton & Sejnowski, 19866; Aarts & Korst, 1989). Although it is this application of simulated annealing which inspired the use of the term temperature to describe the modification to the back-propagation learning rule used in the simulations for this project, the objective here was to modify the back-propagation learning rule to more accurately model human learning, not to model human learning using a Boltzmann machine.

Although Boltzmann machines have recently been used to model some aspects of human performance, there is at least one aspect of such a model which makes it less appropriate as a model of attention in human learning. Simulated annealing continuously decreases the temperature of the system in which it is implemented. No provision is made for increasing temperature if the combinatorial optimization is not going well. In some ways this could be compared with the simulations in which temperature was based on elapsed time but in the simulations in which temperature varies with learning, there would be numerous instances of temperature increases as the network escaped from local minima.

Human cognitive research borrows from computer science but work in this area often has something to give in return. This dissertation has focused mainly on the use of connectionist networks to model human cognitive processes and, consequently, more attention has been paid to the qualitative performance of the networks than the relative

speeds with which they learn. However, for many computing science researchers, the speed and accuracy of the network are just exactly what is of interest.

Manipulating the variability of the output function in the manner suggested here definitely improves the speed of learning in one specific back-propagation network. More research in this area would identify the extent to which a temperature term is useful for other such networks and how it influences other aspects of network performance.

Psychology

Empirical research is strongly influenced by the environment in which the research takes place. It is assumed that if this environment is properly controlled the performance of subjects can be meaningfully compared. The problem with cognitive research is that a significant amount of the "environment" influencing performance is inside the subject's head. It is difficult or impossible to control these influences.

Computer-based models of cognitive processes allow researchers to indirectly investigate these "in-the-head" influences to the extent that they are faithfully modelled. The results presented here are not strong enough to argue that the only way to faithfully model the influence of attention is with a sigmoidally declining temperature parameter, but they do suggest that such a parameter is one possible way of doing so in at least some learning

situations. Further research would identify just exactly what those situations might be.

Any time research within a specific paradigm (especially a relatively new one) is successful, it argues for the usefulness and credibility of the paradigm in general. In addition to what this research says about the specifics of modeling attention, it also provides general support for connectionist models and theories. This investigation was successful in that it definitely does suggest that a connectionist model of human learning should include a model of attention and that temperature may be an appropriate way to do this. To fully substantiate this suggestion, it will be necessary to conduct further research into the performance details of human subjects' learning and attention and into the implementation details of temperature in connectionist networks.

Education

As a model of cognitive processing, connectionism places a strong emphasis on learning. Intuitively, this suggests that such a model would have implications for education. As yet, few of these implications have been made explicit but connectionism has only recently received general acceptance even in the cognitive science community.

To the extent that connectionist models in general are a faithful representation of human cognitive processes, the characteristic responses of a connectionist network to

learning have implications for the way human learning environments should be structured. For example, the process of automatic generalization suggests a mechanism whereby generalizations will be acquired from repeated exposure to typical instances. This argues in favour of discovery learning. On the other hand, this mechanism also suggests that forming associations between new concepts and existing super-ordinate concepts will be facilitated if the super-ordinate concept is sensitized in advance. This suggests that at least some form of advance organizer will assist integration of new information. The fact that a connectionist model incorporates both of these mechanisms may suggest why both of these teaching strategies appear to work.

The behaviour of the specific connectionist networks in this research project also have implications for educational practice. The results provide support for declining demands on attentional resources as a cognitive skill is mastered. Converting controlled cognitive processes to automatic ones is an important part of learning and to optimize this learning, attentional resources should be fully utilized. This argues for the continuous introduction of new material and new perspectives and maybe even new skills even while the skill at hand is still being mastered.

To maximize learning in any one individual, that individual's progress should be closely monitored and their

curriculum should be constantly updated to optimize the use of their attentional resources. This would be almost impossible with the currently dominant, group-centred approach to education, but even with a more student-centred approach there are significant practical difficulties concerning assessment, record-keeping, and delivery of curriculum.

Computer technology is already being used to address many of these difficulties but most of the assessment is product oriented and delivery is rarely individualized. One way for computer-based instruction to individualize curriculum is to maintain a profile of the learner based on on-going assessment of student performance. A measure of the amount of attentional resources being devoted to the task at hand could be used to adjust the pace of delivery and even the content. To the extent that response latencies and even reaction time probes accurately measure the use of attentional resources, these measures should be added to the learner profile.

People's minds are infinitely more complex than any computer-based connectionist network but connectionist models of human cognition can help us understand more about human learning and this increased understanding can lead to more informed educational practice.

BIBLIOGRAPHY

Aarts, E. and Korst, J. (1989). *Simulated annealing and Boltzmann machines: A stochastic approach to combinatorial optimization and neural computing.* Chichester: John Wiley & Sons.

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior, 22,* 261-295.

Baddeley (1990). *Human Memory,* Needham Heights, MA:

Baddeley, A., Lewis, V., Eldridge, M. and Thomson, N. (1984). Attention and retrieval from long-term memory. *Journal of Experimental Psychology: General, 113*(4), 518-540.

Bechtel, W. (1985). Contemporary connectionism: are the new parallel distributed processing models cognitive or associationist? *Behaviorism, 13*(1), 53-61.

Bereiter, C. and Scardamalia, M. (1987). *Psychology of Written Composition.* Hillsdale, NJ: Erlbaum.

Bower, G. H. and Clapper, J. P. (1989). Experimental methods in cognitive science. In M. J. Posner (Ed.) *Foundations of Cognitive Science.* Cambridge, MA: The MIT Press.

Caudill, M. (1988). Neural networks primer: Part III. *AI Expert, 3*(6), 53-59.

Caudill, M. (1991). Neural network training tips and techniques. *AI Expert, 6*(1), 56-61.

Cohen, J. D., Dunbar, K., and McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review, 97*(3), 332-361.

Epstein, W. (1988). Has the time come to rehabilitate Gestalt theory? *Psychological Research, 50*(1), 2-6.

Estes, W. K. (1988). Towards a framework for combining connectionist and symbol-processing models. *Journal of Memory and Language, 27*, 196-212.

Fodor, J. and Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: a critical analysis. *Cognition, 28*, 3-71.

Gluck, M. A. & Bower, G. H. (1988). Evaluating an adaptive network model of human learning. *Journal of Memory and Language, 27*, 166-195.

Gopher, D. & Navon, D. (1980). How is performance limited: Testing the notion of central capacity. *Acta Psychologica, 46*, 161-180.

Hecht-Nielsen, R. (1988). Neurocomputing: picking the human brain. *IEEE Spectrum, 25*(3), 36-41.

Hinton, G. E. and Sejnowski, T. J. (1986). Learning and relearning in Boltzmann machines. In D. E. Rumelhart and J. L. McClelland (Eds), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Cambridge MA: MIT Press.

John, E. R., Tang, Y., Brill, A. B., Young, R., and Ono, K. (1986). Double-labeled matabolic maps of memory. *Science, 233*, 1167-1175.

Jones, W. P. and Hoskins, J. (1987). Back-propagation: a generalized delta learning rule. *Byte, 12*(10), 155-162.

Josin, G. (1987). Neural-network heuristics: three heuristic algorithms that learn from experience. *Byte, 12*(10), 183-192.

Kaplan, S., Weaver, M. and French, R. (1990). Active symbols and internal models: towards a cognitive connectionism. *AI and Society*, *4*, 51-71.

Keele, S. W. (1972). Attention demands of memory retrieval. *Journal of Educational Psychology*, *93*, 245-248.

Kehoe, E. J. (1989). Connectionist models of conditioning: A tutorial. *Journal of the Experimental Analysis of Behavior*, *52*(3), 427-440

Klimesch, W. (1987). A connectivity model for semantic processing. *Psychological Research*, *49*, 53-61.

Kosko, B. (1987). Constructing an associative memory. *Byte*, *12*(9), 137-144.

Laarhoven, P. J. M. van and Aarts, E. H. L. (1987). *Simulated annealing: Theory and applications*. Dordrecht: D. Reidel Publishing Company.

Leeuwen, C. van (1989). PDP and Gestalt: an integration. *Psychological Research*, *50*, 199-201.

Loftus, G. R. (1993). A picture is worth a thousand p values: On the irrelevance of hypothesis testing in the microcomputer age. *Behavior Research Methods, Instrumentation, & Computers*, *24*, (in press).

MacWhinney, B., Leinbach, J., Taraban, R. & McDonald, J. (1989). Language learning: Cues or rules? *Journal of Memory and Language*, *28*(3), 255-277.

Marr, D. (1982). *Vision*. San Francisco, CA: W. H. Freeman.

Maren, A. J., Jones, D. & Franklin, S. (1990). Configuring and optimizing the back-propagation network. In A. Maren, C. Harston, & R. Pap (Eds), *Handbook of Neural Computing Applications* (pp. 233-250). Sandiego, CA: Academic Press.

Massaro, D. W. (1988). Some criticisms of connectionist models of human performance. *Journal of Memory and Language, 27*, 213-234.

McClelland, J. L. (1988). Connectionist models and psychological evidence. *Journal of Memory and Language, 27*, 107-123.

McClelland, J. L., Rumelhart, D. E. and Hinton, G. E. (1986). The Appeal of PDP. In D. E. Rumelhart and J. L. McClelland (Eds), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Cambridge MA: MIT Press.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review, 63*, 81-97.

Miller, L. (1988). Behaviorism and the new science of cognition. *The Psychological Record, 38*, 3-18.

Minsky, M. (1977). Frames-system theory. In P. N. Johnson-Laird and P. C. Watson (Eds), *Thinking*. Cambridge MA: Cambridge University Press.

Minsky, M. and Papert, S. (1969). *Perceptrons: An Introduction to Computational Geometry*. Cambridge, MA: MIT Press.

Norman, D. A. and Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology, 7*, 44-64.

Norris, D. (1990). How to build a connectionist idiot (savant). *Cognition, 35*, 277-291.p6 2

Oaksford, M., Charter, N. and Stenning, K. (1990). Connectionism, classical cognitive science and experimental psychology. *AI & Society, 4*, 73-90.

Obermeier, K. K. & Barron, J. J. (1989). Time to Get Fired Up. *Byte*, 14(8), 217-224.

Papert, S. (1988). One AI or many? *Daedelus, Winter*, 1-14.

Penny, C. G. (1988). A beneficial effect of part-list cuing with unrelated words. *Bulletin of the Psychonomic Society, 26*(4), 297-300.

Pylyshyn, Z. W. (1989). Computing in cognitive science. In M. J. Posner (Ed.) *Foundations of Cognitive Science*. Cambridge, MA: The MIT Press.

Rosenblatt, F. (1962). *Principles of Neurodynamics*. Washington, DC: Spartan Books.

Rumelhart, D. E. (1989). The architecture of mind: A connectionist approach. In M. J. Posner (Ed.) *Foundations of Cognitive Science*. Cambridge, MA: The MIT Press.

Rumelhart, D. E., Hinton, G. E. and McClelland, J. L. (1986). A General Framework for PDP. In D. E. Rumelhart and J. L. McClelland (Eds), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Cambridge MA: MIT Press.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(9), 533-536.

Rumelhart, D. E. and Norman, D. A. (1987). Representation of Knowledge. In *Issues in Cognitive Modeling*, Hillsdale, NJ: Lawrence Erlbaum.

Schneider, W. (1988). Sensitivity analysis in connectionist modeling. *Behavior Research Methods, Instruments, & Computers, 20*(2), 282-288.

Schneider, W. and Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review, 88*(1), 1-55.

Sejnowski, T. J. and Churchland, P. S. (1989), Brain and Cognition. In M. J. Posner (Ed.) *Foundations of Cognitive Science*. Cambridge, MA: The MIT Press.

Shank, R. C. and Abelson, R. P. (1977). Scripts, plans, and knowledge. In P. N. Johnson-Laird and P. C. Watson (Eds), *Thinking*. Cambridge MA: Cambridge University Press.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioural and Brain Science, 11*, 1-74.

Sperling, G. and Dosher, B. A. (1986). Strategy and optimization in human information processing. In K. R. Boff, L. Kaufman and J. P. Thomas (Eds) *Handbook of Perception and Human Performance: Volume I*, New York, NY: John Wiley and Sons.

Vidal, R. V. V. (Ed.) (1993). *Applied simulated annealing*. Berlin: Springer-Verlag.

Walker, S. F. (1990). A brief history of connectionism and its psychological implications. *AI and Society, 4*, 17-38.

Zeidenberg, M. (1987). Modeling the brain. *Byte, 12*(12), 237-246