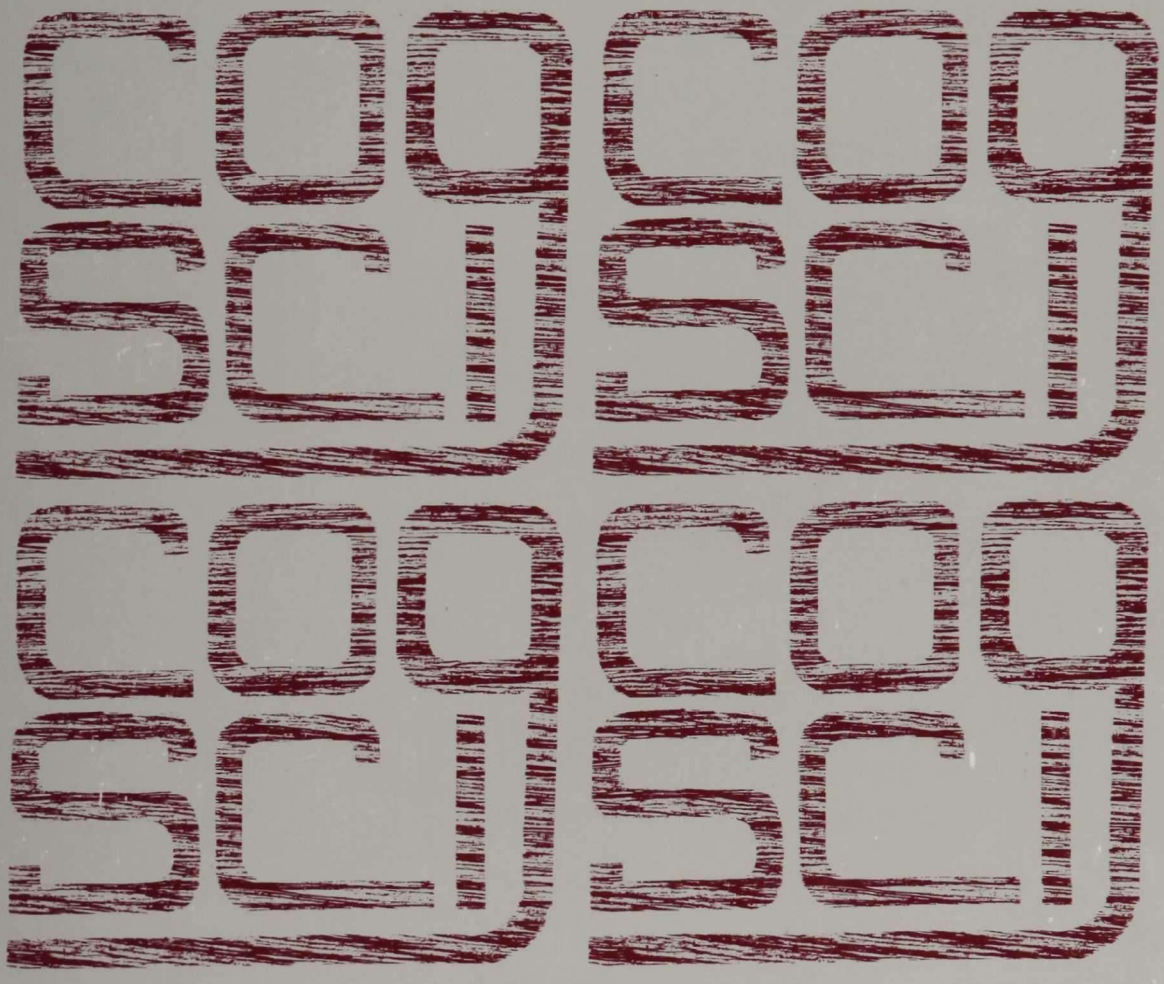
THE EIGHTH ANNUAL
CONFERENCE OF

THE COGNITIVE SCIENCE SOCIETY



1986

AMHERST, MASSACHUSETTS

LEA

Program of the

Eighth Annual Conference of the Cognitive Science Society

15–17 August 1986
Amherst, Massachusetts

LEA LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS
1986 Hillsdale, New Jersey London

Copyright 1986 by the Cognitive Science Society.

All rights reserved. No part of this book may be reproduced in any form, by photostat, microform, retrieval system, or any other means, without the prior written permission of the publisher.

Distributed by

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

ISBN 0-89859-864-8

Printed in the United States of America

THE EIGHTH ANNUAL CONFERENCE
OF THE COGNITIVE SCIENCE SOCIETY

Charles Clifton, Jr.
Chair of Conference

STEERING COMMITTEE

James Michael Royer
Local Arrangements Chair

Paul Cohen
Program Chair

Andrew Barto

Barbara Partee

Neil Stillings

Marie Bremser, Conference Secretary

REVIEWERS

Daniel Anderson	Anne Gardner	Wendy Lehnert	George Reynolds
James Anderson	Jay Garfield	Jack Lockhead	Chris Riesbeck
James Averill	Dedre Gentner	Thomas Malone	Edwina Rissland
Carole Beal	Arthur Graesser	Barbara Malt	Thomas Roeper
Jaime Carbonell	Stephen Grossberg	Edward Matthei	David Rosenbaum
Gregory Carlson	Allen Hanson	Gary Matthews	David Rumelhart
Neil Carlson	John Haugeland	James McClelland	Nestor Schmajuk
Vere Chappell	B. Hayes-Roth	David McDonald	T.Sejnowski
James Chumbley	John Henderson	Carolyn Mervis	Mallory Selfridge
John Clement	Keith Holyoak	Jose Mestre	Makiko Shinjo
Garrison Cottrell	Yumi Iwaski	John Moore	Edward Smith
Jill DeVilliers	Michael Jordan	Jerome Myers	Elliot Soloway
John Donahoe	Avarind Joshi	Terrence Parsons	Richard Sutton
Martha Farah	Raymond Kesner	David Pesetsky	Michael Tanenhaus
Don Fisher	Walter Kintsch	A. Pollatsek	Richard Vardaris
Fernanda Ferreira	Janet Kolodner	Michael Prietula	Raj Wall
Susan Fiske	Angelika Kratzer	James Pustejovsky	Arnold Well
Kay Fite	Judy Kroll	Robert Rafal	Beverly Woolf
Lyn Frazier	S. Kulikowski	Keith Rayner	

Officers of the Cognitive Science Society

John Anderson	Carnegie-Mellon University	1983-1988
Daniel G. Bobrow (Past Chair)	Xerox PARC	1979-1987
Gordon H. Bower	Stanford University	1980-1986
Joan Bresnan	Stanford University, Xerox PARC	1983-1988
Ann Brown	University of Illinois	1984-1989
Allan Collins	Bolt Beranek and Newman	1979-1987
Patrick J. Hayes	Schlumberger PA Research	1979-1985
Geoffrey E. Hinton	Carnegie-Mellon University	1986-1991
Ulric Neisser	Emory University	1985-1990
Donald A. Norman	UC San Diego	1979-1990
Zenon Pylyshyn (Chair 85-86)	Univ. of Western Ontario	1979-1985
David Rumelhart	UC San Diego	1986-1991
Edward E. Smith (Chair-Elect)	Bolt Beranek and Newman	1980-1986

THIS CONFERENCE IS SUPPORTED BY GRANTS FROM THE ALFRED P. SLOAN FOUNDATION AND THE PSYCHOLOGICAL SCIENCES DIVISION OF THE OFFICE OF NAVAL RESEARCH, AND BY THE GRADUATE SCHOOL OF THE UNIVERSITY OF MASSACHUSETTS

PROCEEDINGS OF THE EIGHTH ANNUAL CONFERENCE
OF THE
COGNITIVE SCIENCE SOCIETY
Amherst, Massachusetts
August 15-17, 1986

PRESENTED PAPERS

FRIDAY, AUGUST 15
Learning Distributed Representations of Concepts..... 1
Geoffrey E. Hinton
Representing Causal Schemata in Connectionist Systems..... 13
Richard M. Golden
Skill Learning and Repetition Priming in Symmetry Detection: Parallel Studies of Human Subjects and Connectionist Models..... 23
Neal J. Cohen, Irene T. Abrams, Walter S. Harley, Lisa Tabor, and Terrence J. Sejnowski
Adaptive Resonance Theory: Stable Self-organization of Neural Recognition Codes in Response to Arbitrary Lists of Input Patterns..... 45
Gail A. Carpenter and Stephen Grossberg
Representing Magnitude by Memory Resonance..... 63
Marten J. den Uyl
The Spacing Effect on Nettalk, a Massively-Parallel Network. 72
Charles R. Rosenberg and Terrence J. Sejnowski
Processing Verb Phrase Anaphors..... 90
Michael K. Tanenhaus and Greg N. Carlson
The Effect of the Discourse Center on the Local Coherence of a Discourse..... 96
Susan B. Hudson, Michael K. Tanenhaus, and Gary S. Dell
The Time Course of Pronoun Comprehension..... 102
Rosemary J. Stevenson
The Comprehension of Conceptual Anaphora in Discourse..... 110
Morton Ann Gernsbacher
Conditioning and Categorization: Some Common Effects of Informational Variables in Animal and Human Learning..... 126
Mark A. Gluck and Gordon H. Bower
Simultaneous Configural Classical Conditioning..... 141
Jeffrey C. Schlimmer and Richard H. Granger, Jr.
A Layered Network Model for Learning-to-Learn and Configuration in Classical Conditioning..... 154
E. James Kehoe
Simulation of the Classically Conditioned Nictitating Membrane Response by a Neuron-Like Adaptive Element: a Real-time Variant of the Sutton-Barto Model..... 176
Diane E. J. Blazis, John E. Desmond, John W. Moore, and Neil E. Berthier

SATURDAY, AUGUST 16
Introspection and Reasoning about the Beliefs of Other Agents..... 187
Anthony S. Maida

Causal Reasoning about Quantities..... 196
Kenneth D. Forbus and Dedre Gentner
Neither Pictures Nor Propositions: The Intentionality of Mental Imagery..... 208
Daniel Reisberg and Deborah Chambers
Methods for Evaluating the Validity of Hypothesized Analogies..... 223
John Clement
Using Frequency- and Orientation-Tuned Channels to Determine Surface Slant..... 235
Paul Kube
Adaptation, Brightness Perception and Their Correlation with Photoreceptor Responses..... 245
K.N. Leibovic
Midbrain Mechanisms for Orienting Visual Attention..... 260
Robert D. Rafal and Albrecht W. Inhoff
Repetition-Sensitive Components of Neural Activation: Evidence from Intracranial Recordings from the Human Temporal Lobe..... 272
Michael E. Smith, Eric Halgren, and Gary Heit
Default Defeaters in Explanation-Based Reasoning..... 283
Gilbert Harman, Richard Cullingford, Marie Bienkowski, Ken Salem, and Ian Pratt
Sphexishness, Epistemic Bounds, and a Priori Psychology..... 293
M. P. Smith
Counterfactual Conditionals and the Conjunction Fallacy..... 308
John M. Miyamoto and Emily Dibble
Should We Use Probability in Uncertain Inference Systems?... 320
Max Henrion
Specificity of Expertise in Clinical Reasoning..... 331
Guy-Marie Joseph and Vimla L. Patel
Reflexibility in Problem Solving: The Social Context of Expertise..... 346
James A. Levin, Margaret Riel, Moshe Cohen, Michael Goeller, Marcia Boruta, and Naomi Miyake
Intelligent Tutoring Systems for Scientific Inquiry Skills.. 353
Valerie Shute and Jeffrey Bonar
Focus and Learning in Program Design..... 371
Robert S. Rist
Towards Completely Integrated Parsing and Inferencing..... 381
Christopher K. Riesbeck and Charles E. Martin
Speak of the Devil: Representing Deictic and Speech Act Knowledge in an Integrated Lexical Memory..... 388
Michael Gasser and Michael G. Dyer
Repertory Actors' Memory for Their Parts..... 399
William L. Oliver and K. Anders Ericsson
A Computational Model which Addresses Errors of Over-Generalization and their Subsequent Disappearance in Early Child Language Acquisition..... 407
Jane C. Hill
Remarks on the Psychological Reality of the Subset Principle: Its Relation to Universal Grammar as a Model of the Initial State..... 420
Barbara Lust

SUNDAY, AUGUST 17
Invited Address: A Theory of Discourse Structure..... 432
Barbara J. Grosz
The Use of Remindings in Planning..... 442
Kristian Hammond
Knowing What to Ask Next and Why: Asking Pertinent Questions Using Cases and Hypotheticals..... 452
Kevin D. Ashley
Towards a Memory Architecture that Supports Reminding..... 467
Janet L. Kolodner and Richard E. Cullingford
Application of Cognitive Science Methods to Psychotherapeutic Problem Solving: A Case Study and Some Theory..... 478
Rolf Pfeifer and Marianne Leuzinger-Bohleber
Encoding Planning Knowledge for Recognition, Construction, and Learning..... 488
Charles Dolan and Michael Dyer
The Law as a Learning System..... 500
Edwina L. Rissland and Robert T. Collins
Solving Puzzles with a Connectionist Network Using a Hill-Climbing Heuristic..... 514
Alan H. Kawamoto
BoltzCONS: Reconciling Connectionism with the Recursive Nature of Stacks and Trees..... 522
David S. Touretzky
Attractor Dynamics and Parallelism in a Connectionist Sequential Machine..... 531
Michael I. Jordan
The Representation of Objects for Grasping..... 547
Thea Iberall
A Connectionist Learning Model for 3-Dimensional Mental Rotation, Zoom, and Pan..... 562
Bartlett W. Mel
Seeing-More-Than-Is-There: A Probe of Retinoid Networks..... 572
Arnold Trehub and Alexander Pollatsek

POSTER PRESENTATIONS

FRIDAY, AUGUST 17
An Architecture for Mathematical Cognition..... 585
Kurt Ammon
Cluster: An Approach to Modeling Context..... 595
Yigal Arens
Effects of Focal Brain Damage on Categorization of Visual and Haptic Features..... 601
Bruce Bolster, Jane Honeyman, Valerie Gillis, and Martha Wilson
Belief Maintenance with Uncertainty..... 607
A. Julian Craddock and Roger A. Browse
Some Factors Influencing the Comprehension of Pronouns in Text..... 613
Rosalind A. Crawley

A Psychological Investigation Into the Deictic Center..... 621
Joyce H. Daniels
Towards a Comparative Psychology of Cognitive Content: Exploring Tree Preference Asymmetries In Humans, Pigeons and Monkeys..... 627
Michael R. Dawson, Dwight S. Mazmanian, and William A. Roberts
Is Distributed Connectionism Compatible with the Physical Symbol System Hypothesis?..... 639
Mark Derthick and David C. Plaut
Topological and Dynamical Aspects of a Neural Network Model for Generation of Pursuit Motor Programs..... 645
R. Eckmiller
Three Constructive Algorithms for Network Learning..... 652
Stephen I. Gallant
A Neural Simulation of Classical Conditioning in Aplysia.... 661
Howard N. Henry
Putting Affect into Text..... 669
Eduard H. Hovy
Illusory Conjunctions of Objects and Forms: Integration Errors in a Very Short-term Store..... 676
Helene Intraub
The Literate Listener: Effects of Spelling on Syllable Judgments..... 684
Jola Jakimik and Rebecca Treiman
Modifying Explanations to Understand Stories..... 691
Alex Kass
Computer Strategy of Decision Making Under Time Pressure.... 697
Daniel Lacombe, Claude Sarrazin, and Claude Alain
Organizing Memory for Explanation..... 710
David B. Leake and Christopher C. Owens
A Neural Network Theory of Frontal Lobe Function..... 716
Daniel S. Levine
Views From a Kill..... 728
James H. Martin
Operating Principles and Personal Theories: An Outline of their Roles in Early Lexical Development..... 734
Carolyn B. Mervis
Studying the Problem Solving Behavior of Experts and Novices in Physics Via Computer-Based Problem-Analysis Environments..... 741
Jose Mestre and William Gerace
Cognitive Control of Autonomous Mobile Robots: Nested Hierarchical Intelligent Module..... 747
A. Meystel
Self-supervised Learning: A Scheme for Discovery of "Natural" Categories by Single Units..... 765
Paul Munro
Anatomising Lexical Decision in Phrasal Contexts: When Does Truck Not Prime Car?..... 773
Padraig G. O'Seaghdha
Strategy Selection and Degree of Expertise in Medical Reasoning..... 780
Vimla L. Patel, Jose F. Arocha, and Guy J. Groen

A Real-time Attentional-associative Network for Classical Conditioning of the Rabbit's NMR..... 794
Nestor A. Schmajuk and John W. Moore
Using Mental Schemata: An Experimental Analysis of Computer Skill Acquisition..... 808
Marc M. Sebrechts and Charlotte T. Furstenberg
A Model of Attention Focussing During Problem Solving..... 817
Jude W. Shavlik and Gerald F. DeJong
Two Problems with Backpropagation and other Steepest-descent Learning Procedures for Networks..... 823
Richard S. Sutton
Interest when Nothing Happens: A Note on Narrative Retardation..... 832
Ed S. H. Tan
A Model for Parsing, Learning and Recognizing Objects in a Complex Environment..... 843
Arnold Trehub
Mental Representation of Spatial Information: A Production System Model for Priming and Verification..... 852
Karl F. Wender, Monica Wagener, and Bernd Wittmann
Inverting a Connectionist Network Mapping by Back-Propagation of Error..... 859
Ronald J. Williams
Complement Selection and the Lexicon in Japanese..... 866
Toshiyuki Yamamoto
From Structure to Process..... 876
M. Zock

INDEX OF AUTHORS

Irene T. Abrams 23
Claude Alain 697
Kurt Ammon 585
Yigal Arens 595
Jose F. Arocha 780
Kevin D. Ashley 452
Neil E. Berthier 176
Marie Bienkowski 283
Diane E. J. Blazis 176
Bruce Bolster 601
Jeffrey Bonar 353
Marcia Boruta 346
Gordon H. Bower 126
Roger A. Browse 607
Greg N. Carlson 90
Gail A. Carpenter 45
Deborah Chambers 208
John Clement 223
Moshe Cohen 346
Neal J. Cohen 23
Robert T. Collins 500
A. Julian Craddock 607
Rosalind A. Crawley 611
Richard Cullingford 283
Richard E. Cullingford 467
Joyce H. Daniels 621
Michael R. Dawson 627
Gerald F. DeJong 817
Gary S. Dell 96
Marten J. den Uyl 63
Mark Derthick 639
John E. Desmond 176
Emily Dibble 308
Charles Dolan 488
Michael Dyer 488
Michael G. Dyer 388
R. Eckmiller 645
K. Anders Ericsson 399
Kenneth D. Forbus 196
C. T. Furstenberg 808
Stephen I. Gallant 652
Michael Gasser 388
Dedre Genter 196
William Gerace 741
Morton Ann Gernsbacher 110
Valerie Gillis 601
Mark A. Gluck 126
Michael Goeller 346
Richard M. Golden 13
Richard H., Jr Granger 141
Guy J. Groen 780
Stephen Grossberg 45
Barbara Grosz 432
Eric Halgren 272
Kristian Hammond 442
Walter S. Harley 23
Gilbert Harman 283
Gary Helt 272
Max Henrion 320
Howard N. Henry 661
Jane C. Hill 407
Geoffrey E. Hinton 1
Jane Honeyman 601
Eduard H. Hovy 669
Susan B. Hudson 96
Thea Iberall 547
Albrecht W. Inhoff 260
Helene Intraub 676
Jola Jakimik 684
Michael I. Jordan 531
Guy-Marie Joseph 331
Alex Kass 691
Alan H. Kawamoto 514
E. James Kehoe 154
Janet L. Kolodner 467
Paul Kube 235
Daniel Lacombe 697
David B. Leake 710
K. N. Leibovic 245
Marianne Leuzinger-Boh 478
James A. Levin 346
Daniel S. Levine 716
Barbara Lust 420
Anthony S. Maida 187
Charles E. Martin 381
James H. Martin 728
Dwight S. Mazmanian 627
Bartlett W. Mel 562
Carolyn B. Mervis 734
Jose Mestre 741
A. Meystel 747
Naomi Miyake 346
John M. Miyamoto 308
John W. Moore 176
John W. Moore 794
Paul Munro 765
Padraig G. O'Seaghdha 773
William L. Oliver 399
Christopher C. Owens 710
Vimla L. Patel 331
Vimla L. Patel 780
Rolf Pfeifer 478
David C. Plaut 639
Alexander Pollatsek 572
Ian Pratt 283
Robert D. Rafal 260
Daniel Reisberg 208
Margaret Riel 346
Christopher Riesbeck 381
Edwina L. Rissland 500
Robert S. Rist 371
William A. Roberts 627
Charles R. Rosenberg 72
Ken Salem 283
Claude Sarrazin 697
Jeffrey C. Schlimmer 141
Nestor A. Schmajuk 794
Marc M. Sebrechts 808
Terrence J. Sejnowski 23
Terrence J. Sejnowski 72
Jude W. Shavlik 817
Valerie Shute 353
M. P. Smith 193
Michael E. Smith 272
Rosemary J. Stevenson 102
Richard S. Sutton 823
Lisa Tabor 23
Ed S. H. Tan 832
Michael K. Tanenhaus 90
Michael K. Tanenhaus 96
David S. Touretzky 522
Arnold Trehub 572
Arnold Trehub 843
Rebecca Treiman 684
Monica Wagener 852
Karl F. Wender 852
Ronald J. Williams 859
Martha Wilson 601
Bernd Wittman 852
Toshiyuki Yamamoto 866
M. Zock 876

LEARNING DISTRIBUTED REPRESENTATIONS OF CONCEPTS

Geoffrey E. Hinton
Computer Science Department
Carnegie-Mellon University

ABSTRACT

Concepts can be represented by distributed patterns of activity in networks of neuron-like units. One advantage of this kind of representation is that it leads to automatic generalization. When the weights in the network are changed to incorporate new knowledge about one concept, the changes affect the knowledge associated with other concepts that are represented by similar activity patterns. There have been numerous demonstrations of sensible generalization which have depended on the experimenter choosing appropriately similar patterns for different concepts. This paper shows how the network can be made to choose the patterns itself when shown a set of propositions that use the concepts. It chooses patterns which make explicit the underlying features that are only implicit in the propositions it is shown.

1: Two simple theories of neural representation

There have been many different proposals for how conceptual information may be represented in neural networks. These range from extreme localist theories in which each concept is represented by a single neural unit (Barlow, 1972) to extreme distributed theories in which a concept corresponds to a pattern of activity over a large part of the cortex. These two extremes are the natural implementations of two different theories of semantics. In the structuralist approach, concepts are defined by their relationships to other concepts rather than by some internal essence. The natural expression of this approach in a neural net is to make each concept be a single unit with no internal structure and to use the connections between units to encode the relationships between concepts. In the componential approach each concept is simply a set of features and so a neural net can be made to implement a set of concepts by assigning a unit to each feature and setting the strengths of the connections between units so that each concept corresponds to a stable pattern of activity distributed over the whole network (Hopfield, 1982; Kohonen, 1977; Willshaw, Buneman, & Longuet-Higgins, 1969). The network can then perform concept completion (i.e. retrieve the whole concept from a sufficient subset of its features). The problem with componential theories is that they have little to say about how concepts are used for structured reasoning. They are primarily concerned with the similarities between concepts or with pairwise associations. They provide no obvious way of representing articulated structures composed of a number of concepts playing different roles within the structure.

2: Role-specific units

One way of using neural nets to implement articulated structures of the kind shown in the semantic net formalism in figure 1a is to assign a group of neural units to each possible role and to make the pattern of activity of the units in that group represent the concept that fills the role (Hinton, 1981). Each unit then represents the conjunction of a role with a feature of the concept filling that role (e.g. a unit might be active iff the agent is male). A proposition can then be represented by a stable combination of role fillers as shown in figure 1b. This is a fundamentally different method of representation from either of the two more obvious methods described above. It has the interesting property that the very same concept will have quite different representations when it is playing different roles.

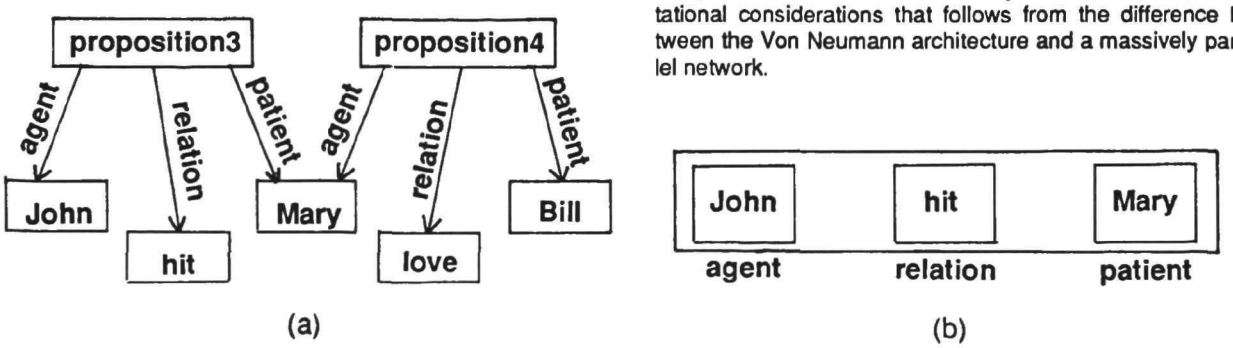
The use of multiple different representations of the same concept appears to be a serious flaw for two reasons: First, it appears to be uneconomical. Second, it is not clear how we know that John in the representation of "John hit Mary " has anything to do with John in the representation of "Mary divorced John".

The economic considerations are complex. The "obvious" way to represent that John is the agent of a certain proposition is to combine a canonical representation of John with a canonical representation of agent. In lisp, for example, a symbolic expression like (agent john) would be an obvious representation and a whole proposition might be represented by the expression ((agent john) (relation hit) (patient mary)). Alternatively, the roles might be implicitly represented by the position of an element in a list and so the whole proposition could be represented by (hit john mary). Either way, the very same symbol is used for John whatever role he plays in the proposition. In logic and in computer programming, the standard way of representing conjunctions is by composing symbolic expressions out of individual symbols. Given a conventional general purpose computer memory, it is easy to store arbitrary compositions of symbols.

If, however, we want to be able to retrieve a proposition from a partial description of its contents, the advantage of always using the very same representation for John is less clear. If the partial description includes the information that John is the agent, we would like this to pick out just those propositions which have John as agent. This is more specific than the propositions which have an agent and also have John in some role. It is the conjunction of John with agent that forms the retrieval cue, and so in a neural implementation it would make sense to have a specific representation for this conjunction. This conjunctive representation can then cause the effects required for completing the whole pattern of activity that represents the proposition. So, even if there is a representation in which John and agent are represented separately, it may be necessary to form a conjunctive representation for retrieval.

A similar argument can be made in other domains. In representing the graphemic structure of the word "chip", for example, it would be possible to use a representation such as ((4 p) (2 h) (3 i) (1 c)) but for a task like filling in the blank in "c-ip" it is inefficient to access all words which contain c and i and p. The identities and roles of the letters need to be conjoined to form more specific access cues. This argues that we need quite different neural representations for (1 c) and (4 c). For the purposes of access to the whole word, these representations need have nothing in common[1]. Indeed, parallel network

[1]In a conventional lisp program, it is easy to separate the representation from the procedures used for retrieval and so it is easy to use a representation that is not in a form that is helpful for retrieval. In a neural net it is probably more important to choose representations that can directly cause the required effects without the intervention of a complex interpreter. This is one of the many differences in representational considerations that follows from the difference between the Von Neumann architecture and a massively parallel network.



(a) (b)

Figure 1: (a) Part of a semantic net. (b) Three groups of units which have two *alternative* stable states that represent the two propositions in the semantic net.

models of reading use separate representations of letters in different positions within the word (McClelland & Rumelhart, 1981).

The second problem with role-specific representations is how to recognize the identity of the various different role-specific representations of the same concept. An efficient way to do this is to have a single, canonical representation for each concept and to have a mechanism for translating between role-specific representations and the canonical one. Hinton (1981) shows how this idea can be implemented in a neural net. It will not be discussed further in this paper.

2.1: Choosing role-specific representations

From now on, we assume that a concept playing a role within a larger structure is represented by a pattern of activity in a group of role-specific units, and we focus on the issue of how this pattern should be chosen. A simple solution is to use patterns selected at random, perhaps with the additional constraint that no two patterns are too similar. The use of random patterns is quite common in research in this area (Hopfield, 1982; Willshaw, 1981). It makes analysis easier and it is a sensible default in the absence of any other ideas about representation. However, it entirely eliminates one of the most interesting aspects of distributed representations: The ability to capture the similarity between concepts by the similarity of their representations, and the consequent ability to generalize new information in sensible ways. We illustrate this point in the simple domain of family relationships.

Figure 2 shows two family trees. All the information in these trees can be represented in simple propositions of the form (**person1 relationship person2**). These propositions can be stored as the stable states of activity of a neural network which contains a group of units for the role **person1**, a group for the role **relationship** and a group for the role **person2**. The net may also require further groups of units in order to achieve the correct interactions between the three role-specific groups. Figure 3 shows a network in which one further group has been introduced for this purpose. Units in this extra group detect combinations of features in the role-specific groups and can be used for causing appropriate completions of partial patterns. Suppose, for example, that one of the extra units becomes active whenever **person1** is old and **relationship** requires that both people be the same age (e.g. the relationship has-husband in the very conventional domain we use). The extra unit can then activate the unit

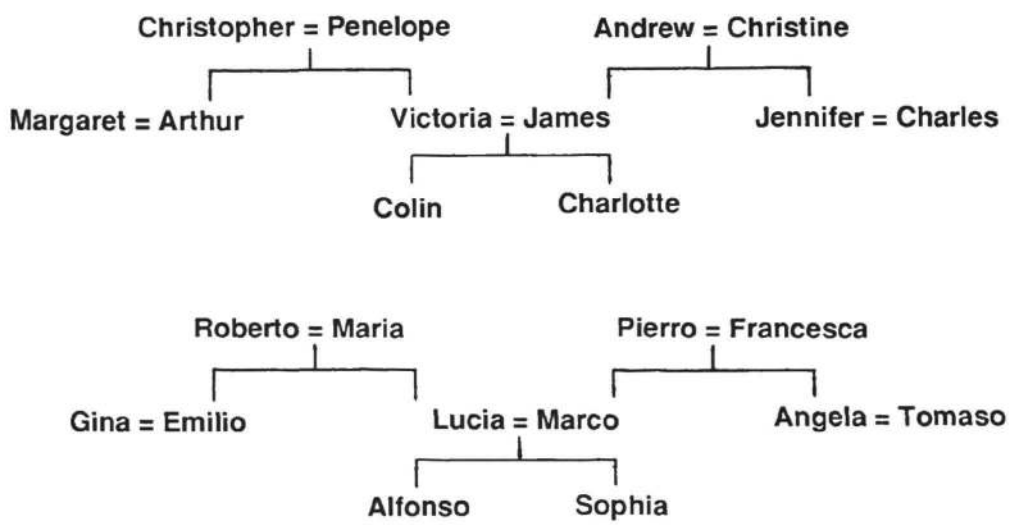


Figure 2: Two isomorphic family trees. The symbol "=" means "married to".

that represents the feature old within the **person2** group. An extra unit that works in this way will be said to encode a micro-inference. It uses some of the features of some of the role-fillers to infer some of the features of other role-fillers and it is typically useful in encoding many different propositions rather than just a single one. By dedicating a unit to a micro-inference that is applicable in many different propositions, the network makes better use of the information carrying capacity of its units than if it dedicated a single extra unit to each proposition. This is an example of the technique of coarse coding described in Hinton, McClelland & Rumelhart (1986). In describing how a micro-inference could be implemented, we assumed that there was a single unit within the **person1** group that was active whenever the pattern of activity in that group encoded an old person. This would not be true using random patterns, but it would be true using a componential representation.

Micro-inferences store propositions by encoding the underlying regularities of a domain. This form of storage has the advantage that it allows sensible generalization. If the network has learned the micro-inference given above it will have a natural tendency to make sensible guesses. If, for example, it is told enough about a new person, Jane, to know that Jane is old and it is then asked to complete the proposition (Jane has-husband ?) it will expect the filler of the **person2** role to be old. To achieve this kind of generalization of domain-specific regularities, it is necessary to pick a representation for Jane in the **person1** role that has just the right active units so that the existing micro-inferences can cause the right effects in the other role-specific groups. A randomly chosen pattern will not do.

The real criterion for a good set of role-specific representations is that it makes it easy to express the regularities of the domain. It is sensible to dedicate a unit to a feature like old because useful micro-inferences can be expressed in terms of this feature. There is another way of stating this point which enables us to avoid awkward questions about whether the network really understands what old means. Instead of saying that activity in a unit means that the person is old, we can simply specify the set of people for which the unit is active. Each unit then corresponds to a way of partitioning all the people into two subsets, and good representations are ones for which these partitions are helpful in expressing the regularities. The search for good representations is then a search in the space of possible sets of partitions[2].

[2]If the units can have intermediate activity levels or can behave stochastically, they do not correspond to clean cut partitions because there will be borderline cases. They are more like fuzzy sets, but the formal apparatus of fuzzy set theory (which is what defines the meaning of "fuzzy") is of no help here so we refrain from using the term "fuzzy" In much of what follows we talk as if units define clearcut sets with no marginal cases. This is just a useful idealisation.

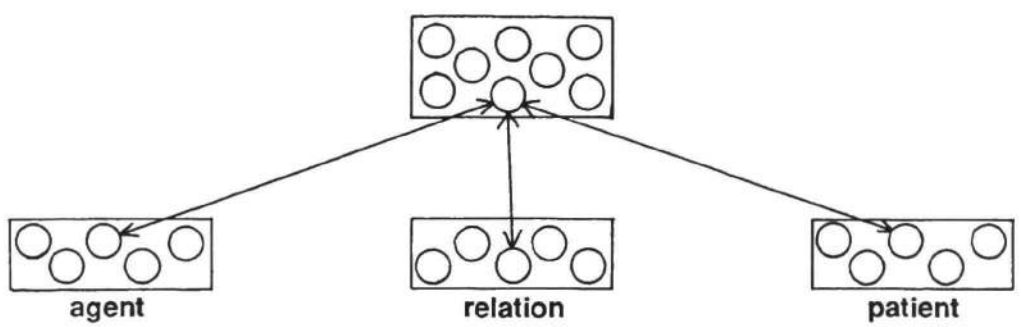


Figure 3: An extra group of units can be used to implement higher-order constraints between the role-specific patterns.

2.2: Giving the network the freedom to choose representations

The network shown in figure 3 has the disadvantage that it is impossible to present a proposition to the network without already having decided on the patterns of activity that represent the people and relationships. We would like the network to use its experience of a set of propositions to construct its own internal representations of concepts, and so we must have a way of presenting the propositions that is neutral with respect to the various possible internal representations. Figure 4 shows how this can be done. The network translates a neutral input representation in which each person or relationship is represented by a single active unit into its own internal representation before making any associations. In the input representation, all pairs of concepts are equally similar[3].

[3]The words of a natural language seem to work in a very similar way. They stand for concepts whilst indicating very little about how those concepts should be represented internally. Monomorphemic words with similar meanings do not generally have similar forms. So a pattern of activity based on the form of the word is not a good way of capturing the similarities between meanings. There must be a process that maps word forms into word meanings. This process must be far more complex than the simulations we present here because many word forms, like "bank", are ambiguous and so the process of going from the input representation to a representation of the word meaning cannot be performed separately for each word. The meaning of whole phrases or sentences must be used for disambiguation (Waltz & Pollack, 1985).

3: A network that learns distributed representations

In our attempts to show that neural networks can learn sensible distributed representations we have tried several different learning procedures. The most successful of these, so far, is the "back-propagation" procedure described in Rumelhart, Hinton & Williams (1986), and the simulation we present uses back-propagation. This learning procedure, which is briefly outlined in section 4, assumes that the units have real-valued outputs between 0 and 1 which are deterministic functions of their total inputs, where the total input, x_j, to unit j is given by

$$x_j = \sum_i y_i w_{ji} \qquad (1)$$

A unit has a real-valued output, y_j, which is a non-linear function of its total input.

$$y_j = \frac{1}{1+e^{-x_j}} \qquad (2)$$

The units are arranged in layers with a layer of input units at the bottom, any number of intermediate layers, and a layer of output units at the top. Connections within a layer or from higher to lower layers are forbidden: All connections go from lower layers to higher ones.

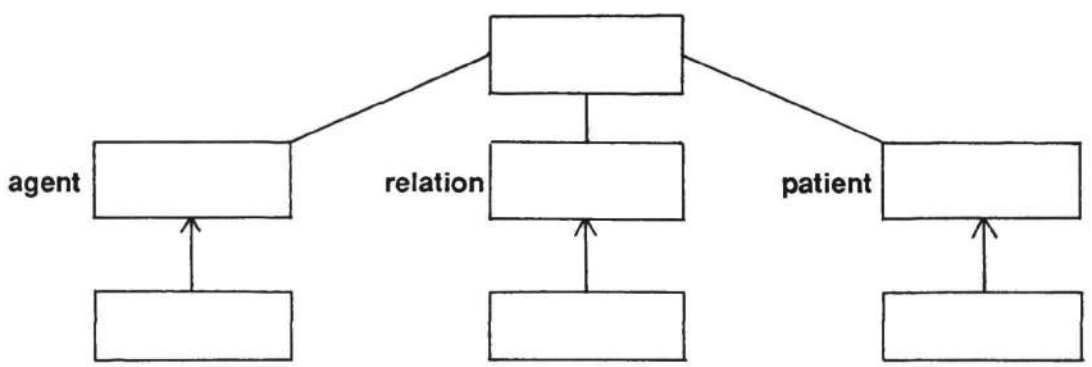


Figure 4: The state of each role-specific group can be fixed via a special input group. By varying the weights between the special input groups and the role-specific groups the network can develop its own role-specific representations instead of being forced to use representations that are pre-determined.

An input vector is presented to the network by setting the states of the input units. Then the states of the units in each layer are determined by applying Eq. 1 and 2 to the connections coming from lower layers. All units within a layer have their states set in parallel, but different layers have their states set sequentially, starting at the bottom and working upwards until the states of the output units are determined.

To use the back-propagation learning procedure we need to express the task of learning about family relationships in a form suitable for a layered network[4]. There are many possible layered networks for this task and so our choice is somewhat arbitrary: We are merely trying to show that there is at least one way of doing it, and we are not claiming that this is the best or only way. The network we used is shown in figure 5. It has a group of input units for the filler of the **person1** role, and another group for the filler of the **relationship** role. The output units represent the filler of the **person2** role, so the network can only be used to complete propositions when given the first two terms[5]. The states of the units in the input groups are clamped from outside and the network then determines the states of the output units and thus completes the proposition.

For some relationships, like uncle, there may be several possible fillers for the **person2** role that are compatible with a given filler of the **person1** role. In a stochastic network it would be reasonable to allow the network to choose one of the possibilities at random. In the deterministic network we decided to insist on an output which explicitly represented the whole set of possible fillers. This is easy to do because the neutral representation that we used for the output has a single active unit for each person and so there is an obvious representation for a set of people.

Using the relationships *{father, mother, husband, wife, son, daughter, uncle, aunt, brother, sister, nephew, neice }* there are 104 instances of relationships in the two family trees shown in figure 2. We trained the network on 100 of these instances. The training involved 1500 sweeps through the 100 examples with the weights being updated after each sweep. The details of the training procedure are given in section 4. After this substantial experience of the domain, the weights were very stable and the network performed correctly on all the training examples: When given a **person1** and a **relationship** as input it always produced activity levels greater than 0.8 for the output units corresponding to correct answers and activity levels of less than 0.2 for all the other output units. A typical example of the activity levels in all layers of the network is shown in figure 5.

The fact that the network can learn the examples it is shown is not particularly surprising. The interesting questions are: Does it create sensible internal representations for the various people and relationships that make it easy to express regularities of the domain that are only implicit in the examples it is given? Does it generalize correctly to the remaining examples? Does it make use of the isomorphism between the two family trees to allow it to encode them more efficiently and to generalize relationships in one family tree by analogy to relationships in the other?

3.1: The representations

Figure 6 shows the weights on the connections from the 24 units that are used to give a neutral input representation of **person1** to the 6 units that are used for the network's internal, distributed representation of **person1**. These

[4]Rumelhart et. al. describe another version of the procedure which does not require a layered net. It works for arbitrary recurrent networks, but requires more complex units that remember their history of activity levels. We have not applied this version to the family relationships task.

[5]We would have preferred it to perform completion when given any two terms. This could have been done by using a bigger network in which there were three input groups and three output groups, but learning would have been slower in the larger network and so we opted for the simpler case.

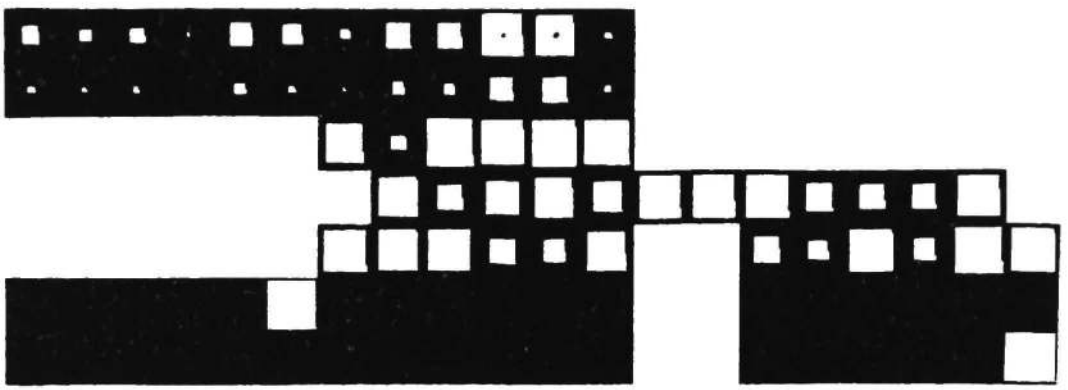


Figure 5: The activity levels in a five layer network after it has learned. The bottom layer has 24 input units on the left for representing person 1 and 12 units on the right for representing the relationship. The white squares inside these two groups show the activity levels of the units. There is one active unit in the first group (representing Colin) and one in the second group (representing has-aunt). Each of the two groups of input units is totally connected to its own group of 6 units in the second layer. These two groups of 6 must learn to encode the input terms as distributed patterns of activity. The second layer is totally connected to the central layer of 12 units, and this layer is connected to the penultimate layer of 6 units. The activity in the penultimate layer must activate the correct output units, each of which stands for a particular person2. In this case, there are two correct answers (marked by black dots) because Colin has two aunts. Both the input and output units are laid out spatially with the English people in one row and the isomorphic Italians immediately below.

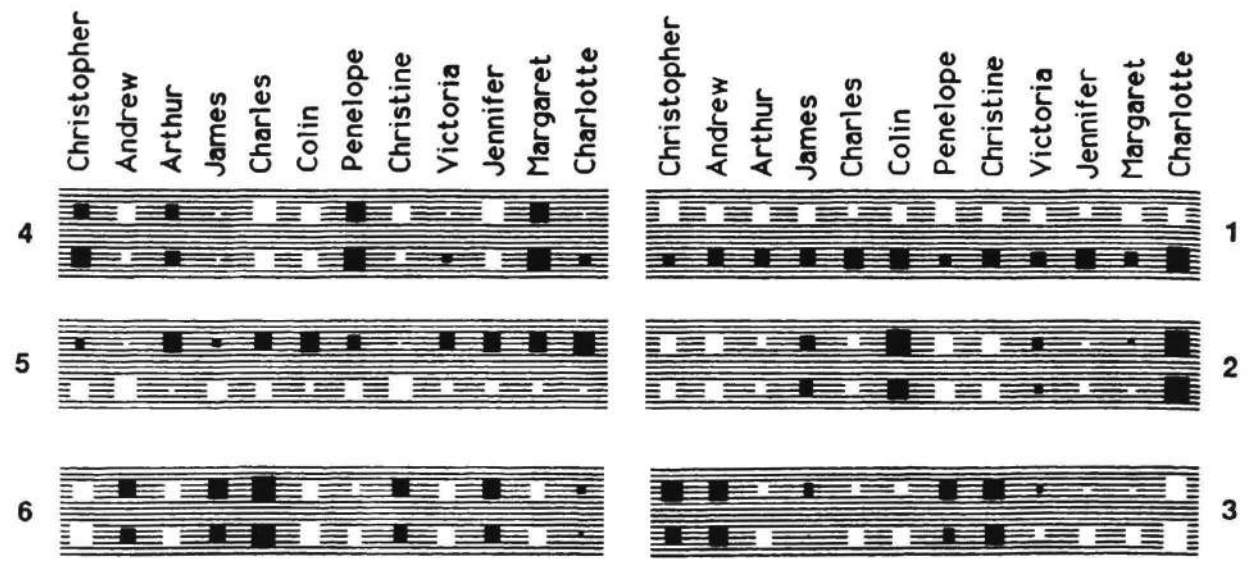


Figure 6: The weights from the 24 input units that represent people to the 6 units in the second layer that learn distributed representations of people. White rectangles stand for excitatory weights, black for inhibitory weights, and the area of the rectangle encodes the magnitude of the weight. The weights from the 12 English people are in the top row of each unit. Beneath each of these weights is the weight from the isomorphic Italian.

weights define the "receptive field" of each of the 6 units in the space of people. It is clear that at least one unit (unit number 1) is primarily concerned with the distinction between English and Italian and that most of the other units ignore this distinction. This means that the representation of an English person is very similar to the representation of their Italian equivalent. The network is making use of the isomorphism between the two family trees to allow it to share structure and it will therefore tend to generalize sensibly from one tree to the other.

Unit 2 encodes which generation a person belongs to. Notice that the middle generation is encoded by an intermediate activity level. The network is never explictly told that generation is a useful three-valued feature. It discovers this for itself by searching for features that make it easy to express the regularities of the domain. Unit 6 encodes which branch of the family a person belongs to. Again, this is useful for expressing the regularities but is not at all explicit in the examples[6].

It is initially surprising that none of the 6 units encodes sex. This is probably because of the particular set of relationship terms that were used. Each of the 12 relationship terms completely determines the sex of **person2** so the sex of **person1** is redundant. Figure 7 shows that one of the 6 units which encodes the twelve possible relationships is entirely devoted to predicting the sex of **person2**. If we had included relationships like spouse there would have been more pressure to encode the sex of **person1** because this would have been useful in constraining the possible fillers of the **person2** role.

3.2: Micro-inferences and scientific laws

There is an interesting analogy between the way in which the network represents propositions and the way in which scientists represent the structure of the natural world. In order to ex-

[6]In many tasks, features that are useful for expressing regularities between concepts are also observable properties of the individual concepts. For example, the feature male is useful for expressing regularities in the relationships between people and it is also related to sets of observable properties like hairyness and size. We carefully chose the input representation to make the problem difficult by removing all local cues that might have suggested the appropriate features.

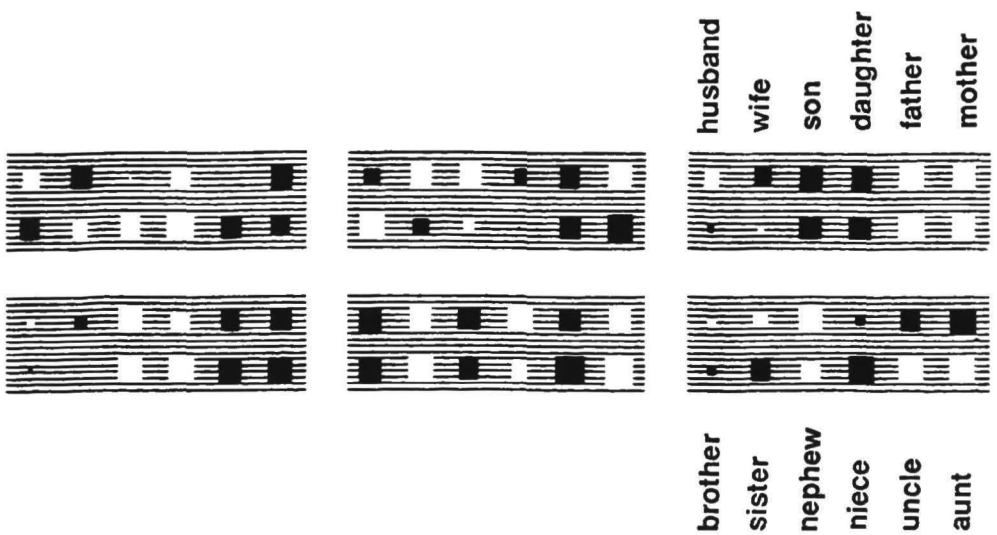


Figure 7: The weights from the 12 input units that represent relationships to the 6 units in the second layer that learn distributed representations of the relationships.

press the regularities in the data, a scientist must describe them using appropriate terms. For example, substances with widely different appearances must be grouped together into categories like "acid", "salt", or "base". If the appropriate terms are given in advance, the task is much easier than if the terms themselves must be discovered by searching for sets of terms that allow laws to be expressed. The gradient descent procedure used by the network also has its analog in scientific research. Initial definitions of the descriptive terms can be used to formulate laws and the apparent exceptions can often be used to refine the definitions.

Naturally, there are also many important differences between the way scientists proceed and the way learning occurs in the network. Scientists would not normally be satisfied if their theory consisted of a very large number of statistical "laws" and they needed a computationally intensive procedure to decide what the laws predicted.

3.3: Generalization

The network was trained on 100 of the 104 instances of relationships in the two family trees. It was then tested on the remaining four instances. The whole training and testing procedure was repeated twice, starting from different random weights. In one case the network got all four test cases correct and in the other case it got 3 out of 4, where "correct" means that the output unit corresponding to the right answer had an activity level above 0.5, and all the other output units were below 0.5. In the test cases, the separation between the activity levels of the correct units and the activity levels of the remainder were not as sharp as in the training cases. Figure 8 shows the activity levels of all 24 output units for each of the 4 test cases after training.

Any learning procedure which relied on finding direct correlations between the input and output vectors would generalize very badly on the family tree task. Consider the correlations between the filler of the **person1** role and the filler of the **person2** role. The filler of **person1** that is used in each of the generalization tests is negatively correlated with the correct output vector because it never occured with this output vector during training, and it did occur with other output vectors. The structure that must be discovered in order to generalize correctly is not present in the pairwise correlations between input units and output units.

The good generalisation exhibited by the network shows that the structure which it has extracted from the training examples agrees with the structure which we attribute to the domain. We would like to be able to say that the training set implicitly contains the information about how

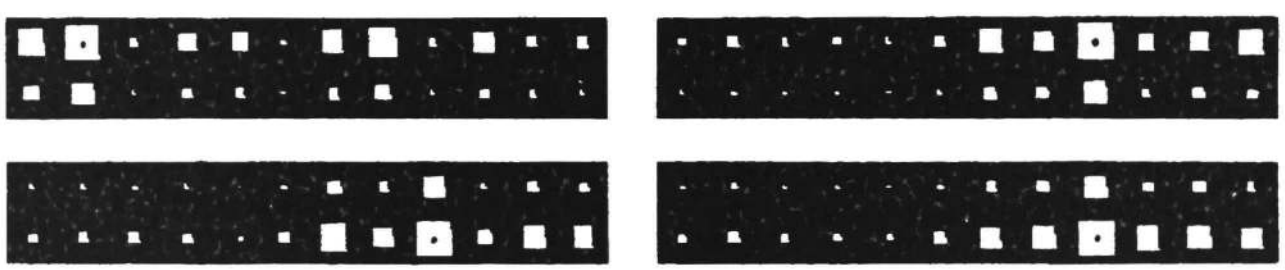


Figure 8: The activity levels of the output group in the four test cases that were not shown during training. The dots are on the correct answers. Notice that in every case the network has a slight tendency to activate the isomorphic person in the other family tree.

to generalize and that the network has correctly extracted this implicit information. But this requires a prescriptive domain-independent theory of how a set of examples should be used for making generalisations. Such a theory would constitute the "computational level" of understanding for learning research (Marr, 1982), and would be a major advance which could guide research at the algorithmic and implementation levels. Unfortunately, we know of no such theory and so we are restricted to showing that the learning procedure produces sensible generalizations in particular domains.

4: The back-propagation learning procedure

The aim of the learning procedure is to find a set of weights which ensure that for each input vector the output vector produced by the network is the same as (or sufficiently close to) the desired output vector. If there is a fixed, finite set of input-output cases, the total error in the performance of the network with a particular set of weights can be computed by comparing the actual and desired output vectors for every case. The error, E, is defined as

$$E = \frac{1}{2}\sum_c \sum_j y_{jc} - d_{jc})^2 \qquad (3)$$

where c is an index over cases (input-output pairs), j is an index over output units, y is the actual state of an output unit, and d is its desired state. To minimize E by gradient descent it is necessary to compute the partial derivative of E with respect to each weight in the network. This is simply the sum of the partial derivatives for each of the input-output cases. For a given case, the partial derivatives of the error with respect to each weight are computed in two passes. We have already described the forward pass in which the units in each layer have their states determined by the input they receive from units in lower layers using Eq. 1 and 2. The backward pass which propagates derivatives from the top layer back to the bottom one is more complicated.

The backward pass starts by computing $\partial E/\partial y$ for each of the output units. Differentiating Eq. 3 for a particular case, c, and suppressing the index c gives

$$\frac{\partial E}{\partial y_j} = y_j - d_j \qquad (4)$$

We can then apply the chain rule to compute dE/dx_j

$$\frac{\partial E}{\partial x_j} = \frac{\partial E}{\partial y_j}\frac{dy_j}{dx_j}$$

Differentiating Eq. 2 to get the value of dy_j/dx_j gives

$$\frac{\partial E}{\partial x_j} = \frac{\partial E}{\partial y_j} y_j(1-y_j) \qquad (5)$$

This means that we know how a change in the total input, x, to an output unit will affect the error. But this total input is just a linear function of the states of the lower level units and the weights on the connections, so it is easy to compute how the error will be affected by changing these states and weights. For a weight, w_{ji}, from i to j the derivative is

$$\frac{\partial E}{\partial w_{ji}} = \frac{\partial E}{\partial x_j}\frac{\partial x_j}{\partial w_{ji}}$$

$$= \frac{\partial E}{\partial x_j} y_i \qquad (6)$$

and for the output of the ith unit the contribution to $\partial E/\partial y_i$ resulting from the effect of i on j is simply

$$\frac{\partial E}{\partial x_j}\frac{\partial x_j}{\partial y_i} = \frac{\partial E}{\partial x_j} w_{ji}$$

so taking into account all the connections emanating from unit i we have

$$\frac{\partial E}{\partial y_i} = \sum_j \frac{\partial E}{\partial x_j} w_{ji} \tag{7}$$

We have now seen how to compute $\partial E/\partial y$ for any unit in the penultimate layer when given $\partial E/\partial y$ for all units in the last layer. We can therefore repeat this procedure to compute $\partial E/\partial y$ for successively earlier layers, computing $\partial E/\partial w$ for the weights as we go. The amount of computation required for the backward pass is of the same order as the forward pass (it is linear in the number of connections).

One way of using $\partial E/\partial w$ is to change the weights after every input-output case. This has the advantage that no separate memory is required for the derivatives. An alternative scheme, which we used in the research reported here, is to accumulate $\partial E/\partial w$ over all the input-output cases before changing the weights. The simplest version of gradient descent is then to change each weight by an amount proportional to the accumulated $\partial E/\partial w$.

$$\Delta w = -\varepsilon \frac{\partial E}{\partial w} \tag{8}$$

This method does not converge as rapidly as methods which make use of the second derivatives, but it is much simpler and can easily be inplemented by local computations in parallel hardware. It can be significantly improved, without sacrificing the simplicity and locality, by using an acceleration method in which the current gradient is used to modify the velocity of the point in weight space instead of its position.

$$\Delta w(t) = -\varepsilon \frac{\partial E}{\partial w(t)} + \alpha \Delta w(t-1) \tag{9}$$

where t is incremented by 1 for each sweep through the whole set of input-output cases, and α is an exponential decay factor between 0 and 1 that determines the relative contribution of the current gradient and earlier gradients on the weight change. Eq. 9 can be viewed as describing the behavior of ball-bearing rolling down the error-surface when the whole system is immersed in a liquid with viscosity determined by α.

The learning procedure is entirely deterministic, so if two units within a layer start off with the same connectivity and the same weights there is nothing to make them ever differ from each other. We therefore break symmetry by starting with small random weights.

The learning procedure often works better if it is not required to produce outputs as extreme as 1 or 0. To give an output of 1, a unit must receive an infinite total input and so the weights grow without bound. All the examples of back-propagation described in this paper use a more liberal error measure which treats all values above 0.8 as perfectly satisfactory if the output unit should be on and all values below 0.2 as perfectly satisfactory if the output unit should be off. Otherwise, the error is the squared difference from 0.8 or 0.2.

There are many aspects of the learning procedure which make it highly implausible as a model of learning in real neural networks. There are ways of removing the prohibition on recurrent connections (Rumelhart, Hinton & Williams, 1986) and it may be possible to overcome the need for an externally supplied desired output vector. But the back-propagation phase is central to the learning procedure and it is quite unlike anything known to occur in the brain. The connections are all used backwards, and the units use a different input-output function. We therefore view this learning procedure as an interesting way of demonstrating what can be achieved by gradient descent, without claiming that this is how gradient descent is actually implemented in the brain. Nevertheless, the success of the learning procedure suggests that it is

worth looking for other more plausible ways of doing gradient descent.

4.1: The learning parameters used for the family tree simulation

We tried several different values for the parameters ε and α in Eq. 9. We finally chose to use $\varepsilon = .005$ and $\alpha = .5$ for the first 20 sweeps through the 100 training examples and $\varepsilon = .01$ and $\alpha = .9$ for the remaining sweeps. The reasons for varying the parameters during learning and the methods used to choose reasonable parameters are discussed in more detail in Plaut, Nowlan & Hinton (1986). All the weights were initially chosen at random from a uniform distribution between –.3 and +.3.

To make it easier to interpret the weights, we introduced "weight-decay". Immediately after each weight change the magnitude of every weight was reduced by 0.2%. After prolonged learning the decay was balanced by $\partial E/\partial w$, so the final magnitude of each weight indicated its usefulness in reducing the error. Weight decay is equivalent to modifying the error function so that, in addition to requiring the error to be small, it requires the sum of the squares of the weights to be small. A side-effect of this modification is that it sometimes causes two units to develop very similar sets of weights with each weight being half as big as it would be if the job was done by a single unit. This is because $(.5w)^2+(.5w)^2 < w^2$.

To achieve negligible error without weight decay required 573 sweeps through the 100 training examples. The weights shown in figure 6 were obtained by allowing the learning to run for 1500 sweeps with weight-decay.

Acknowledgements

This research was supported by grants from the System Development Foundation.

Copyright © 1986 Geoffrey E. Hinton

References

Barlow, H. B. Single units and sensation: A neuron doctrine for perceptual psychology? *Perception*, 1972, *1*, , 371-394.

Hinton, G. E. Implementing semantic networks in parallel hardware. In G. E. Hinton & J. A. Anderson (Eds.), *Parallel Models of Associative Memory,* Hillsdale, NJ: Erlbaum, 1981.

Hinton, G. E., McClelland, J. L. & Rumelhart, D. E. Distributed representations. In D. E. Rumelhart, J. L. McClelland, & the PDP research group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition.,* Cambridge, MA: Bradford Books, 1986.

Hopfield. J. J. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences U.S.A.*, April 1982, *79*, 2554-2558.

Kohonen, T. *Associative memory: A system-theorectical approach.* Berlin: Springer, 1977.

Marr, D. *Vision.* San Francisco: Freeman, 1982.

McClelland, J. L., & Rumelhart, D. E. An interactive activation model of context effects in letter perception, Part I: An account of basic findings. *Psychological Review*, 1981, *88*, 375-407.

Plaut, D. C., Nowlan, S. J., & Hinton, G. E. *Experiments on back-propagation.* Technical Report CMU-CS-86-126, Carnegie-Mellon University, June 1986.

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

Waltz, D. L., & Pollack, J. B. Massively parallel parsing: A strongly interactive model of natural language interpretation. *Cognitive Science*, 1985, *9*, 51-74.

Willshaw, D. J., Buneman, O. P. & Longuet-Higgins, H. C. Nonholographic associative memory. *Nature*, 1969, *222*, 960-962.

Willshaw, D. Holography, associative memory, and inductive generalization. In G. E. Hinton & J. A. Anderson (Eds.), *Parallel models of associative memory,* Hillsdale, NJ: Erlbaum, 1981.

Representing causal schemata in connectionist systems

Richard M. Golden

Department of Psychology

Brown University

The connectionist approach to human memory is based upon the idea that knowledge can be stored implicitly in the form of real-valued interconnections among a set of simple "neuron-like" computing elements (Hinton & Anderson, 1981). The schema system approach (Rumelhart, 1980; Schank & Abelson, 1977) considers human memory to be organized in terms of many small packets of knowledge called schemata. If a knowledge packet is defined as some sequence of causally related events, then it is referred to as a "causal schema" or "script."

Although these two seemingly different approaches to the problem of modelling human memory might seem incompatible, they are actually intimately related (Rumelhart, Smolensky, McClelland, & Hinton, 1986; also see Touretzky & Hinton, 1985). In this paper, a connectionist model of how causal schemata are used in the recall of actions from simple stories is described. The paper is organized in the following manner. First, an explicit procedure for representing complex causal schemata as "neural activation patterns" is discussed in detail. Next, the fundamental neural mechanisms that are used to process and learn information are described and motivated from a probabilistic viewpoint. Finally, the resulting system is used to model some experimental data obtained by Bower, Black, and Turner (1979) in their studies of human memory for written text.

Representational assumptions

The fundamental entity in this model is called a "causal relationship." A Causal Relationship (CR) consists of an "initial situation," an "action," and a "final situation." If the "final situation" of one causal relationship is identical to the "initial situation" of another causal relationship, then the pair of causal relationships are said to be "causally linked." A collection of causal relationships that have been linked together in this manner is referred to as a causal schema. For example, let the notation (S3,A3,S5) indicate a causal relationship formed from an initial situation S3, an action A3, and a final situation S5. Figure 1 shows that CR (S1,A1,S3) is causally linked to CR (S3,A3,S5) which, in turn, is linked to CR (S5,A2,S6).

A basic assumption of the model discussed here is that a causal relationship corresponds to some unique pattern of neural activity in the brain. In addition, similar causal relationships are assumed to possess similar neural codings. More specifically, a causal relationship is represented as a 160-dimensional state vector (i.e., a list of 160 real numbers) where the ith element of the vector specifies the firing rate of the ith neuron in a neural network. Consider the causal relationship at the top of Figure 2. The initial situation field of this CR is interpreted as: "The actor is at a restaurant, the actor is hungry, and the actor is at the table." The action field of the CR is interpreted as: "The actor orders the meal." The final situation field is interpreted as: "The actor is at the table, the actor is hungry, the food is on the table."

To encode the initial situation field as a 64-dimensional subvector, the three 64-dimensional binary orthogonal subvectors corresponding to the states: "At_restaurant," "At_table," and "Hungry" are added together to form a composite 64-dimensional subvector. If an element of this composite subvector is non-negative, then the value of that element is set equal to +1, otherwise the value of that element is set equal to -1. The resulting modified composite subvector represents the initial situation field of the 160-dimensional causal relationship state vector. More formally, the 41 states in the "state dictionary" form a psychological *basis set* that

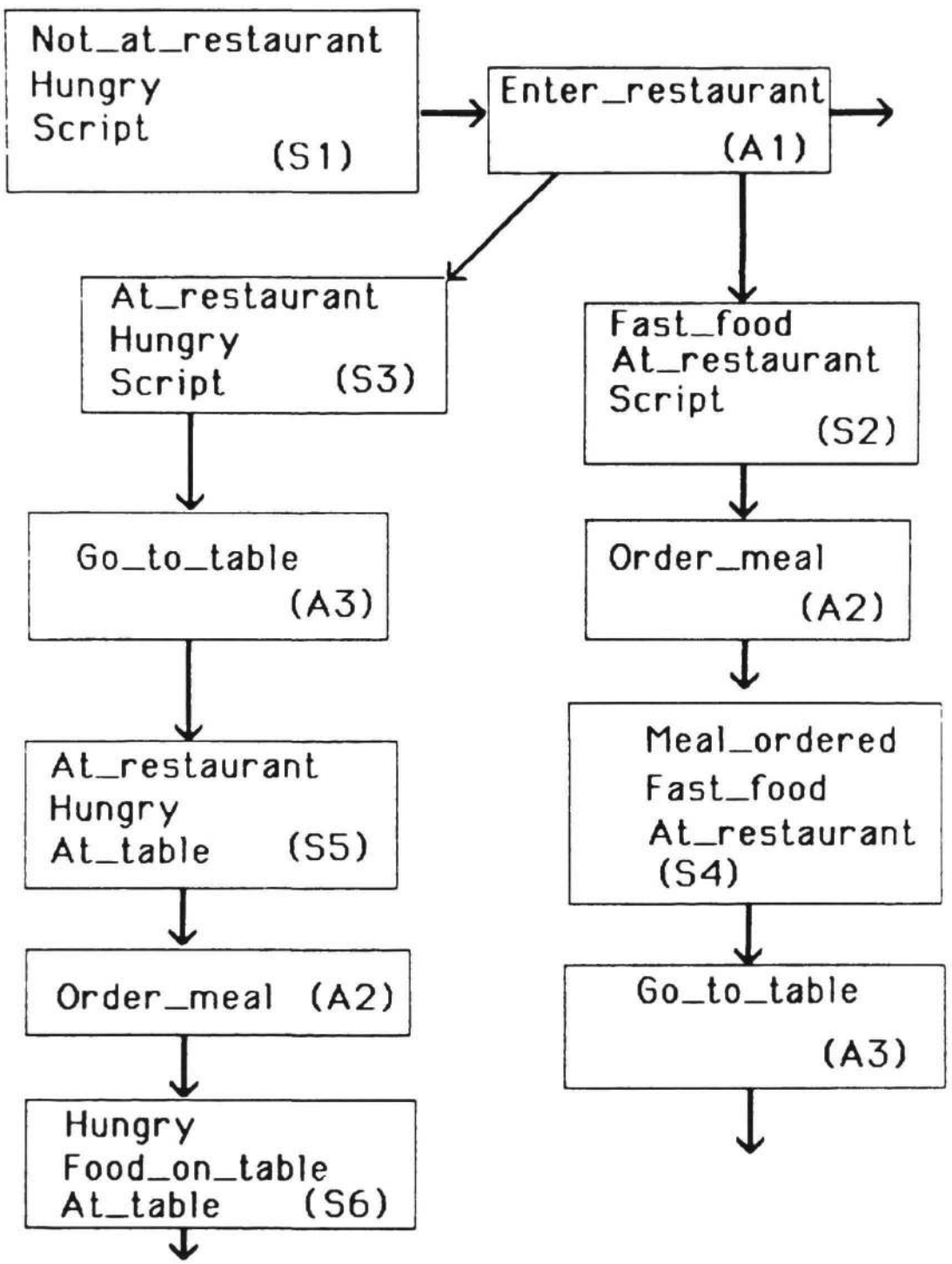


Figure 1. A portion of a causal schema. Note that such schemata may be represented as unordered collections of causal relationships.

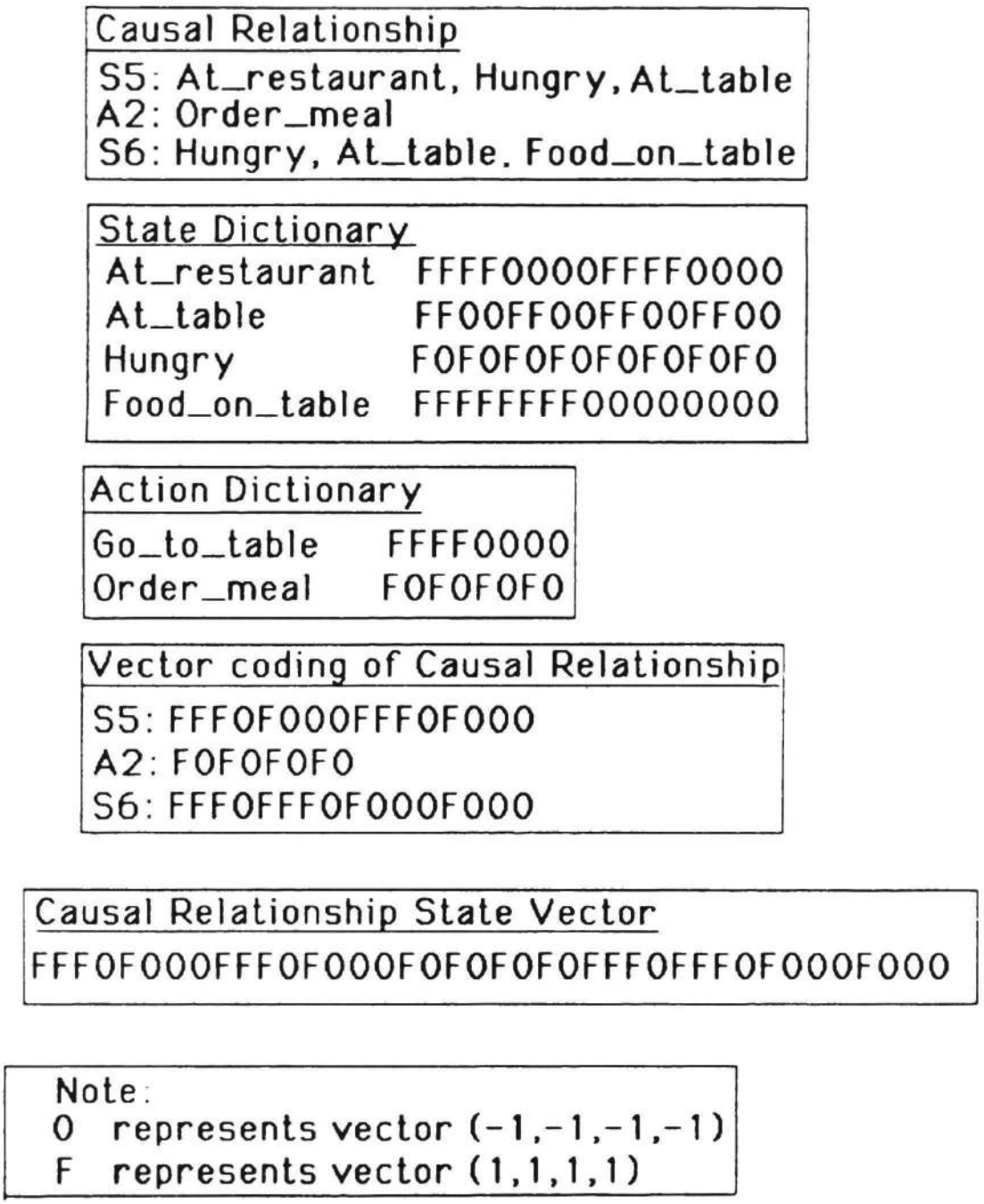


Figure 2. Representing a causal relationship as a state vector. The initial situation field of the CR is formed by adding the 64-dimensional subvectors in the state dictionary labelled "At_restaurant," "Hungry," and "At_table" together and assigning a +1 to the non-negative elements of the resulting vector and a -1 to the negative vector elements. The action field of the CR is simply looked up in the action dictionary. The final situation field of the CR is constructed in the same manner as the initial situation field. Note that the symbol F refers to a sequence of four positive ones, while the symbol 0 refers to a sequence of four negative ones.

can represent over 50,000 situations in a 64-dimensional state vector space.

The action field of the causal relationship is represented by a 32-dimensional binary orthogonal subvector whose value is obtained directly from the "action dictionary." The encoding procedure for the final situation field of the causal relationship vector is identical to the procedure used to encode the initial situation vector field.

Making "most probable" decisions with a neural model

The fundamental problem of content-addressable memory may be formulated as follows. Given some unusual or "improbable" vector $\mathbf{X}_0$, construct a more probable interpretation. More formally, we can search for a maximum of some probability density function P(**X**) in the vicinity of $\mathbf{X}_0$. This density function indicates the relative frequency of occurrence of a stimulus vector **X** within the environment. Within this framework, we can view a broad class of neural network models as specific gradient ascent algorithms that maximize P(**X**), while some popular neural network learning algorithms are viewed as procedures that estimate the general form of P(**X**). In particular, the learning process is viewed as a procedure that constructs a P(**X**) such that P(**X**) obtains a local maximum for each class of vectors learned by the system.

Memory Recall is Maximizing a Probability Density Function

The Brain-State-in-a-Box (BSB) neural model (Anderson, Silverstein, Ritz, & Jones, 1977) is an abstract nervous system model that was designed to study various psychological phenomena. Information in this system is represented by an N-dimensional state vector that specifies a particular pattern of firing rates over a group of N neurons. The *i*th neuron in the system is modelled as a simple linear integrator possessing a maximum and minimum firing rate. The system operates by amplifying an incoming activation pattern (state vector) using positive feedback until many of the neurons in the system have obtained their maximum or minimum firing rates. The assumption that neurons possess maximum and minimum firing rates implies that the neural activation pattern over the set of N neurons is constrained to lie within an N-dimensional hypercube. More formally the BSB model is defined as follows:

$$x_i(k+1) = S[x_i(k) + \sum_j a_{ij}x_j(k)] \qquad (1)$$

where $x_i(k)$ is the *i*th element of the state vector **X** at discrete time interval k, and a_{ij} is a synaptic weight representing the synaptic efficacy between the *i*th and *j*th neurons in the system. The linearized sigmoidal function S[a] is defined as follows. S[a] = +1 for $a > +1$, S[a] = -1 for $a < -1$, and S[a] = a for $-1 \leq a \leq +1$.

Now let $\mathbf{X}_0$ be the initial value of the system state vector. Let E(**X**) be defined as:

$$E(\mathbf{X}) = (-1/2)\mathbf{X}^T\mathbf{A}\mathbf{X} \qquad (2)$$

where the *ij*th element of the matrix **A** is the synaptic weight a_{ij}. Golden (1986) has demonstrated that the BSB model is an algorithm that transforms $\mathbf{X}_0$ into a new vector **X**, located in the vicinity of $\mathbf{X}_0$, such that $E(\mathbf{X}) \leq E(\mathbf{X}_0)$ under fairly general conditions.

Now define P(**X**) as the probability of **X** and let the form of P(**X**) be given as follows:

$$P(\mathbf{X}) = ke^{-E(\mathbf{X})} \qquad (3)$$

where k is a constant chosen such that $\int P(\mathbf{X}) = 1$ and E(**X**) is defined in (2). The gradient of P(**X**) with respect to **X** is calculated as follows:

$$\mathbf{GRAD}[P(\mathbf{X})] = -\,\mathbf{GRAD}[E(\mathbf{X})]P(\mathbf{X}). \qquad (4)$$

Equation (4) states that an algorithm that moves along the path of steepest descent with respect to $E(\mathbf{X})$, is also moving along the path of steepest ascent with respect to $P(\mathbf{X})$. Moreover, when the state vector is "improbable" (i.e., $P(\mathbf{X})$ is small), the step size will be small. But when the state vector is "probable" (i.e., $P(\mathbf{X})$ is large), the step size will be large. Finally note that since $P(\mathbf{X})$ is a monotonically decreasing function of $E(\mathbf{X})$, a neural network model that minimizes $E(\mathbf{X})$ is also maximizing $P(\mathbf{X})$. In psychological terms, the BSB model is constructing a "more probable" interpretation of the initial state vector $\mathbf{X}_0$.

Learning is Estimating the Form of the Probability Density Function

In the BSB model, the synaptic weight between the *i*th and *j*th neurons in the system is specified by a real number, a_{ij}, that corresponds to the *ij*th element in the **A** matrix. The set of synaptic weights specify the parameters of the probability density function in (3) and therefore also specify the "knowledge base" of the system. To obtain a set of synaptic weights responsive to a particular set of training stimuli, the autoassociative Widrow-Hoff learning rule (Anderson, 1983) is used. More specifically, at each learning trial, a state vector is randomly selected from the set of training stimuli. Next, this training stimulus is used to update the current set of synaptic weights according to the synaptic weight updating rule:

$$a_{ij}(k+1) = a_{ij}(k) + \gamma\,(x_i - \sum_m a_{im}(k)x_m)x_j \qquad (5)$$

where $a_{ij}(k)$ is the value of the synaptic weight between the *i*th and *j*th neurons in the system at learning trial k, x_i is the *i*th element of the training stimulus vector, and γ is a positive learning constant. The on-diagonal elements, a_{ii}, are not updated.

Let **A** be a matrix of synaptic weights formed by the coefficients a_{ij}. Let **X** be the random vector associated with some unknown stationary probability distribution function in the environment. The autoassociative Widrow-Hoff learning rule can be shown to be searching for an **A** matrix that minimizes the expected value of the Euclidean distance between **AX** and **X** where the expectation is taken with respect to **X** (Widrow, 1971). Let **C** be a value of the random vector **X**. Golden (in preparation) has demonstrated that if **C** is a hypercube vertex, **A** is symmetric, and **AC** is in the same quadrant of the hypercube as **C**, then **C** is a strict local maximum of the density function. In conjunction with the observation that (2) is an energy or Liapunov function, this implies that a region about **C** exists such that any state initiated in that region must approach **C** as time increases.

Psychologically, these arguments simply indicate that the autoassociative Widrow-Hoff learning rule connects the neurons in the system such that the neural network *implicitly* assigns high probabilities to stimuli that have been taught to the system. These neural interconnections are then used by the BSB neural network to reconstruct "more probable" interpretations of less probable or novel state vector stimuli.

The Causal Schema neural network model

The Causal Schema (CS) neural model is a special type of production system specifically designed to model causal schemata. The model makes specific qualitative predictions regarding the pattern of errors made by people in recalling short, simple stories from memory. To illustrate the operation and behavior of the model, an experiment performed by Bower, Black, and Turner (1979) is described and then simulated using the CS neural model. Additional tests of the model are discussed by Golden (in preparation).

Bower et al. (1979) had college students learn a series of very short stories that were organized about routine event sequences or "scripts." Some of the stories were generated from the same script and were therefore very similar to one another (e.g., "visiting a doctor" and "visiting a dentist"), while other stories studied by the subjects were quite distinctive. After an intervening task, the subjects were given the titles of the stories as cues and requested to recall the actions that were mentioned in the story. Bower et al. found that "stated" script actions (i.e., actions explicitly mentioned within a story) were recalled more frequently than "unstated" script actions (i.e., actions implicitly mentioned), and "unstated" script actions were recalled more frequently than "other" types of script actions. In addition, as the similarity between two stories learned by a subject was increased, the number of "unstated" script actions recalled by the subjects increased and the number of "stated" script actions recalled by the subjects decreased (Table 1).

Constructing a Long-term Memory

The first step to modelling the Bower et al. (1979) experiment is the development of a long-term memory for the CS neural model. Such a memory was constructed in the following manner. Four distinct causal schemata associated with the event sequences "going to a restaurant," "going to a fast food restaurant," "going to a lecture," and "going to a doctor" were constructed. Next, two variations upon each of these four basic causal schemata were constructed. The resulting set of twelve schemata are then completely specified by a total of 107 causal relationships. A matrix of synaptic weights was then constructed from this stimulus set of causal relationships by training the system using the autoassociative Widrow-Hoff learning rule for 1000 learning trials. The resulting set of synaptic weights was defined as the model's long-term memory. In the simulations described here, five such matrices were generated using different random number seeds in an attempt to model the long-term memory structures of five college students.

Golden (in preparation) describes some computer simulations illustrating how this type of long-term memory system can be used to control behavior. In particular, an incomplete CR representing an initial situation (e.g., (S1,0,0)) is presented to the BSB model which reconstructs the action field of the CR. The effect of this action upon the environment results in a new situation (e.g., (S2,0,0)) that, in turn, can be used by the BSB model to reconstruct the second action in some action sequence.

Modelling the Learning of Short Stories

After the 1000 learning trials using the "long-term memory" stimulus set of 107 CRs were completed, the simulated "subjects" were trained with 24 "story sets." In particular, each subject was taught a single story set for an additional 100 learning trials and then tested. Each story set consisted of two similar stories derived from the same causal schema and one very distinctive story derived from another causal schema. A "story" simply consisted of a collection of five causal relationship state vectors that were implicitly linked together using the causal schema state vector encoding procedure that was described earlier. Note that the system's knowledge of stories is stored over the *same* set of synaptic weights as the system's long-term memory.

Modelling the Recall Process

Figure 3 illustrates the main flow of control when the CS neural model is requested to recall a story from memory. The BSB model is provided with an initial situation and action field (corresponding to the title of the story), and reconstructs the final situation field. The final situation field is then used to form the initial situation field of a new state vector. The action and final situation fields of this new state vector are filled with zeroes. The new state vector is

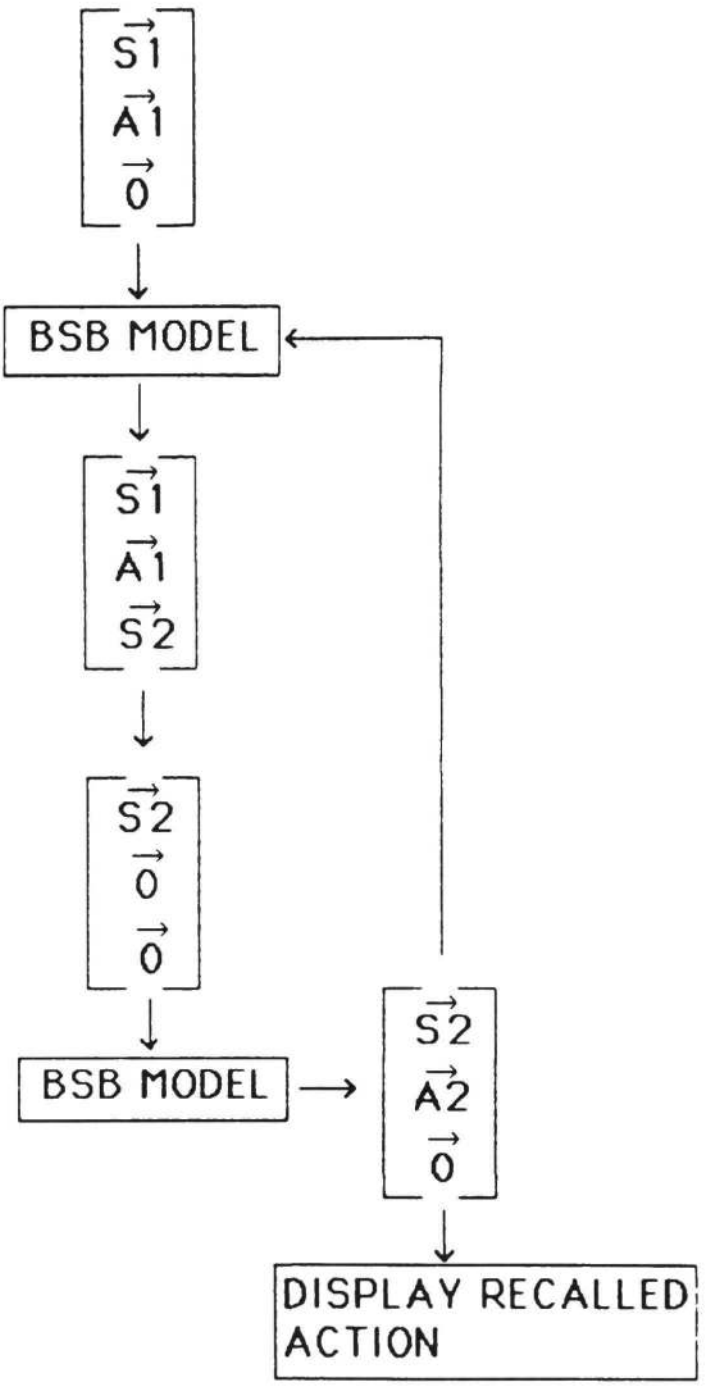


Figure 3. Flow of control during story recall. The CR representing the story title, (S1,A1,0), is transformed by the BSB model into (S1,A1,S2) thus reconstructing a final situation field for the partially specified (S1,A1,0). The final situation of (S1,A1,S2) is then used to form (S2,0,0) which is presented to the BSB model. The BSB model recalls an action, A2, from memory and this action is recorded by the experimenter. CR (S2,A2,0) may now be used as a memory cue to recall the next action in the story and the above cycle is repeated.

then submitted to the BSB model which reconstructs a new action. This new action is recorded as the first action recalled by the model, and the new action and new initial situation are used to initiate the cycle once again to recall the second action from memory.

Computer Simulation Results

Table 2 provides the results of the computer simulations which may be compared with the results obtained in the Bower et al. (1979) study. Like the human data in Table 1, "stated" actions are recalled more frequently than "unstated" actions which are recalled more frequently than "other" actions. In addition, as the number of related stories that are learned by the computer subjects increases, the number of stated actions recalled decreases and the number of unstated actions recalled increases. The interaction of "number of related stories" and "action type" was highly significant ($p < 0.01$) in the computer simulations treating either story sets or computer subjects as random factors.

Summary

A connectionist model of causal schemata in human memory has been described that makes specific qualitative predictions about experiments involving memory for written text. As an example, the performance of the model was compared with human subjects' performance in a specific psychological experiment. For this particular experiment, the CS model successfully captured the general qualitative characteristics of human recall memory for simple stories. In addition, a procedure for representing complex causal schemata as collections of neural activation patterns (state vectors) and a probabilistic interpretation of memory recall and learning in the BSB model were discussed.

Acknowledgements

This research was supported in part by a grant from the National Science Foundation to J. A. Anderson, administered by the Memory and Cognitive Processes section (Grant BNS-82-14728). I am grateful to David Cooper for suggesting that some analyses of connectionist models might be viewed within a probabilistic framework. I would also like to thank the members of the Brown University neural modelling group (particularly Jim Anderson and Mike Rossen) for their comments and advice.

References

Ackley, D. A., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science, 9,* 147-169.

Anderson, J. A. (1983). Cognitive and psychological computation with neural models. *IEEE transactions on systems, man, and cybernetics, 5,* 799-815.

Anderson, J. A., Golden, R. M., & Murphy, G. L. (1986). *Concepts in distributed systems.* Paper presented at S.P.I.E. Advanced Institute Series Hybrid and Optical Computers, Leesburg, Virginia.

Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review, 84,* 413-451.

Bower, G. H., Black, J. B., & Turner, T. J. (1979). Scripts in memory for text. *Cognitive Psychology, 11,* 177-220.

Golden, R. M. (1986). The "Brain-State-in-a-Box" neural model is a gradient descent algorithm. *Journal of Mathematical Psychology, 30*, 73-80.

Golden, R. M. (in preparation). *Modelling causal schemata in human memory: A connectionist approach.* Unpublished doctoral dissertation, Brown University, Providence, RI.

Hinton, G. E., & Anderson, J. A. (1981). *Parallel models of associative memory.* Hillsdale, NJ: Erlbaum.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences USA, 79*, 2554-2558.

Hopfield, J. J. (1984). Neurons with graded response have collective properties like those of two-state neurons. *Proceedings of the National Academy of Sciences USA, 81*, 3088-3092.

Rumelhart, D. E. (1980). Schemata: The building blocks of cognition. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension.* Hillsdale, NJ: Erlbaum.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). Parallel distributed processing models of schemata and sequential thought processes. In J. L. McClelland and D. E. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. v. 1: Foundations.* Cambridge, MA: Bransford Books/MIT Press.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding.* Hillsdale, NJ: Erlbaum.

Smolensky, P. (1984). The mathematical role of self-consistency in parallel computation. *Proceedings of the Sixth Annual Conference of the Cognitive Science Society.* Boulder, Colorado.

Touretzky, D. S., & Hinton, G. E. (1985). Symbols among the neurons: Details of a connectionist inference architecture. *Proceedings of the Ninth International Joint Conference on Artifical Intelligence, v. 1.*, Los Angeles, CA., pp. 238-243.

Widrow, B. (1971). Adaptive filters. In R. E. Kalman and N. DeClaris (Eds.), *Aspects of network and system theory.* New York: Holt, Rinehart, & Winston.

Table 1

Average number of actions recalled by human subjects
(adapted from Bower et al., 1979)

		Number of stated actions	Number of unstated actions	Number of other actions
Number of related stories	1	3.03	0.80	0.39
	2	2.27	1.26	0.35

Table 2

Average number of actions recalled by the CS neural model
(Computer simulation)

		Number of stated actions	Number of unstated actions	Number of other actions
Number of related stories	1	2.47	0.10	0.02
	2	1.62	0.40	0.00

SKILL LEARNING AND REPETITION PRIMING IN SYMMETRY DETECTION: PARALLEL STUDIES OF HUMAN SUBJECTS AND CONNECTIONIST MODELS[1]

Neal J. Cohen

Irene T. Abrams

Walter S. Harley

Lisa Tabor

Department of Psychology
The Johns Hopkins University

Terrence J. Sejnowski

Department of Biophysics
The Johns Hopkins University

ABSTRACT

The present paper is a preliminary report of our work exploring skill learning and repetition priming in parallel studies of mirror symmetry detection in humans and network models. The memory mechanisms supporting the acquisition of skill and repetition priming in humans have been the subject of much speculation. On one account, drawing on the distinction between procedural and declarative learning, these learning phenomena grow out of experience-based tuning and reorganization of processing modules engaged by performance in a given domain, in a manner that is intimately tied to the operation of those modules. Such learning appears similar to that suggested by the incremental learning algorithms currently being explored in massively-parallel connectionist models (e.g., the Boltzmann machine). In the present work, both learning phenomena were observed in the behavioral data from human subjects and the simulation data from the network models. The network models showed priming effects from the start of de novo learning despite being designed to handle generalization to new materials - the essence of skill learning - and without additional mechanisms designed to provide a temporary advantage for recently presented material. Priming occurred for the human subjects despite the use of novel materials for which pre-existing representations cannot already be present in memory. These findings support the notion that skill learning and repetition priming are linked to basic incremental learning mechanisms that serve to configure and reorganize processing modules engaged by experience.

---

[1] Supported by a Biomedical Research Support Grant from the Division of Research Resources, NIH and a grant from the Sloan Foundation to NJC, and by grants from the National Science Foundation, System Development Foundation, General Electric Corporation, Exxon Education Foundation, Allied Corporation Foundation, Westinghouse, and Smith, Kline & French Laboratories to TJS. We thank Caroline McKeldin for research assistance and Valerie Mehl for manuscript preparation. Reprint requests should be sent to Neal J. Cohen, Department of Psychology, The Johns Hopkins University, Baltimore, MD 21218.

INTRODUCTION

The memory mechanisms supporting the ability to acquire new skills and the additional facilitation of performance due to repetition priming effects (i.e., superiority in task performance for repeated materials) have been the subject of much recent work in cognitive science and cognitive neuropsychology. Skill learning and repetition priming are both exhibited as a consequence of experience in amnesic patients despite their impaired recall and recognition of the specific materials used to train and test skilled performance, poor recollection of the learning experiences during which skilled performance was developed, and poor insight into the nature of the knowledge underlying the increasingly skilled performance (see Cohen, 1985; Squire & Butters, 1984 for reviews). Skill learning and repetition priming have been dissociated from aspects of explicit remembering (recall and recognition) in normal subjects as well (see Kolers, 1979; Schacter, 1985; Tulving, 1985 for reviews). We have argued that these learning phenomena reflect the operation of a procedural memory system that influences the organization of perceptual and action systems in a way that does not depend upon explicit representation of particular learning experiences or of rules about the world (Cohen, 1985; Cohen, Eichenbaum, DeAcedo & Corkin, 1985; Squire & Cohen, 1984). On this view, skilled performance in a given domain grows out of the tuning and reorganization of processing and action modules engaged by performance in that domain; the learning that results consequent to experience in that domain is intimately tied to the operation of its processing components.

The characteristics of procedural learning and its relationship to the declarative system have been explored most extensively by John Anderson, who has modeled the procedural system in a production-system framework in the context of a broader conceptualization of memory and cognition (e.g., Anderson, 1982, 1983). Our view of the procedural system, however, and the explanation of skill learning and repetition priming that it suggests, seems to bear strong similarity to that of the incremental learning algorithms currently being explored in massively-parallel network models.

One example of network architectures is the Boltzmann machine, which has been applied successfully to such problems as figure-ground separation in

visual perception (Sejnowski & Hinton, 1986). The processing units in this architecture are binary and connected symmetrically with coupling strengths or weights that can have positive or negative values. There exists a learning algorithm for the Boltzmann machine that allows the network to automatically find a set of weights for solving a problem given only examples of typical inputs and the required outputs (Hinton & Sejnowski, 1983; Ackley, Hinton & Sejnowski, 1985). The Boltzmann machine learning algorithm has recently been used for learning to detect the axis of mirror symmetry in checkerboard-based patterns (Sejnowski, Kienker & Hinton, 1986). This problem is a second-order predicate in the sense of Minsky and Papert (1969) and is beyond the capability of the perceptron learning algorithm (Rosenblatt, 1959) and Hopfield networks (Hopfield, 1982). The crucial difference that distinguishes Boltzmann machines from perceptrons and allows them to solve difficult problems is the presence of additional units between the input and output layers, called hidden units. These hidden units can be used as feature detectors for solving the problem; the learning algorithm discovers the optimal set of feature detectors by shaping the weights among units through incremental changes.

In nearly all previous work with connectionist models, the focus has been on the ability of a network to generalize from examples to new instances on which the network was not previously trained[2]. The work with human subjects suggests that skill learning, measured in terms of the improvement in performance for novel materials in the trained domain, is closely associated with repetition priming effects, the additional facilitation in performance specific to the materials actually presented during training. In order to examine whether these learning phenomena are linked in the network models as they are in humans, we have been exploring skill learning and repetition priming in both Boltzmann machine architectures and human subjects in studies of mirror symmetry detection for checkerboard-based patterns. It is important to note that this enterprise is not in any way intended to offer the Boltzmann machine implementation of symmetry-detection learning as a detailed model of how humans actually acquire skill in this domain. Rather it is intended to

---

[2] Although McClelland & Rumelhart (1986) and Carpenter and Grossberg et al. (1985) have considered repetition priming effects in models of word recognition.

explore whether simple, incremental learning algorithms can manifest both stimulus-generalizable skill learning and stimulus-specific repetition priming phenomena in the same networks.

Perception of mirror symmetries is superb in humans (e.g., Kohler, 1929; Garner, 1962; Bruce & Morgan, 1975; Barlow & Reeves, 1979), and, indeed, has been assumed to play an important role in the early stages of visual processing (e.g., Marr & Nishihara, 1978; Biederman, 1985). It is not surprising, then, that studies of mirror symmetry detection in humans have focused on its psychophysical characteristics rather than on its improvement with practice. The acquisition of visual symmetry recognition has been studied in pigeons (Delius & Nowak, 1982) and Boltzmann machines (Sejnowski, Kienker & Hinton, 1986), but this study is, to our knowledge, the first to examine the acquisition of mirror symmetry detection in humans. We assume that despite their already developed skill in symmetry detection, humans will show considerable practice effects, learning how best to apply symmetry-detection routines to this particular class of visual materials, e.g., learning about the critical visual features that are diagnostic of mirror symmetry for different axes in these particular checkerboard-based patterns.

An important question that has been raised in previous work on repetition priming in humans, and especially in work with human amnesic patients, is the extent to which priming depends upon a structured, pre-existing knowledge base in which particular facts can be temporarily activated by recent experience, rather than upon acquisition of new knowledge (e.g., see Fowler, Napps & Feldman, 1985; Schacter, 1985; Gordon, 1986). The present work addresses this issue in two ways: First, the checkerboard-based patterns used as stimuli in the present studies are novel materials, and are not a part of subjects' pre-existing knowledge structures. That is, any priming effects obtained here could not be accounted for by postulating the temporary activation of already stored logogen- or pictogen-like representations. Second, the computer simulation work explores the possibility of priming effects in networks learning de novo, where no pre-existing knowledge is provided about topography, symmetry, or checkerboard-based patterns. The present paper is a preliminary report of findings from our parallel studies of human subjects and computer simulations relevant to understanding the mechanisms of repetition priming and its relationship to skill learning.

METHOD

Subjects

Twenty-two students at The Johns Hopkins University volunteered to be subjects in the present studies. Of these, 12 served in the priming experiment and 10 served in the axis generalization experiment.

Materials

Stimuli were computer-generated, mirror-symmetric visual patterns constructed by designating as purple approximately 40% (range = 31-49%) of 81 cells within a 9x9 blue display. Each pattern was symmetrical about one and only one of four axes: horizontal, vertical, left diagonal, or right diagonal. The precise distribution of purple and blue cells within a pattern, given a particular axis of symmetry, was determined randomly by an IBM PC. The patterns were approximately 13 cm on each side and subtended approximately 8 deg of visual angle.

Apparatus

Patterns were presented on an IBM PC color display controlled by an IBM PC. Reaction times and response axes were recorded via a Summagraphics optical mouse.

General Procedure

Subjects were seated individually in front of an IBM PC color display controlled by an IBM PC, and held a Summagraphics optical mouse in their dominant hand. They were presented with a series of 400 patterns (divided evenly among the 4 axes) organized into blocks of 60-80. Within each block there was an equal number of patterns representing each axis, with no axis occurring more than 3 times in succession. Subjects initiated each trial by pressing the left button of the mouse. Each trial consisted of a visual pattern presented for 83 msec followed, after a 17 msec unfilled delay, by a visual mask. The mask consisted of a regular, alternating-color checkerboard whose cells were identical in size and color to the test stimuli. It remained on the screen until a response was recorded, or until 8 sec had elapsed, ending the trial. Subjects were instructed to indicate the axis of symmetry

by moving the mouse on the table-mounted pad in the direction analogous to the axis perceived on the display. They were instructed to respond as quickly and as accurately as possible, guessing whenever in doubt. After each response, feedback was given in the form of a re-presentation of the pattern with the addition of a line indicating the correct axis. The feedback pattern remained on the screen for 2 sec.

Prior to the start of the test series, subjects received a practice session consisting of 20 stimuli (divided equally among the axes presented in the subsequent experimental condition), each presented with a feedback line indicating the correct axis.

Exp. 1: Priming experiment

Subjects were trained on a combination of repeating (primed) and nonrepeating patterns. Of the 400 stimuli presented to each subject, there were 240 nonrepeating patterns (60 of each axis) common to all subjects, with each subject receiving a different, pseudo-randomized order. The remaining 160 stimuli were the primes, consisting of a set of patterns that repeated at one of three different rates: In the 1/10 condition, 4 patterns (1 of each axis) repeated in every 10+2 stimuli; in the 1/20 condition, 8 patterns (2 of each axis) repeated in every 20+6 stimuli; in the 1/40 condition, 16 patterns (4 of each axis) repeated in every 40+10 stimuli. Note that the overall percentage of prime stimuli in each series was maintained at 40% across the different conditions by doubling the number of prime patterns for each halving of the prime repetition rate. The prime patterns were yoked across the three priming-rate conditions such that a given prime pattern appeared in each condition. Stimuli were presented in 5 blocks of 80.

Exp. 2: Axis generalization experiment

This experiment consisted of a training phase, in which patterns were symmetric about one of only two axes, and a testing phase, in which patterns were symmetric about one of four axes, as in the priming experiment. The training phase consisted of 100 stimuli symmetric about either the left and right diagonal axes, for one group of subjects, or the horizontal-vertical axes, for another group. These stimuli, presented in 2 blocks of 50, were different for each subject. The testing phase consisted of 300 stimuli (75 of

each axis) common to all subjects, with each subject receiving a different, pseudo-random order. The test stimuli were presented in 5 blocks of 60.

Computer Simulations

Networks

Two networks were studied. Both had 81 input units, representing the 81 cells of the 9x9 array, and 4 output units, representing the 4 possible axes of mirror symmetry. They differed with respect to the number and connectivity of the hidden units. One network had 12 hidden units, each fully and symmetrically connected to all of the input units and output units (hereafter called the global-unit network). The other network had 72 hidden units, each connected to the 4 output units and to one of 9 3x3 sections of the 9x9 stimulus array (hereafter called the local-unit network). For both networks, the hidden units were not interconnected.

General procedure

The learning algorithm and implementation details are the same as those described in Sejnowski, Kienker and Hinton (1986). All networks started with all weights set to zero except for the axis generalization experiment, in which the training phase ran until performance reached asymptote and then switched to the testing phase until performance again reached asymptote. Approximately 40,000-100,000 patterns were presented in each simulation.

Materials

Patterns were randomly generated following the same constraints as those used in the human experiments. Patterns were composed of approximately 40% (range = 31-49%) of the 81 cells in a 9x9 array being "on" in such a way as to be symmetric about one and only one of the four axes.

Exp. 1: Priming experiment

The networks were trained on a combination of repeating (primed) and nonrepeating patterns. In each simulation, two priming rates were included, with one being a multiple of 2 or 3 that of the other. In each case (with the exception of the condition with the highest priming rate; see below), there were at least 20 different patterns that served as prime stimuli for each

priming rate, divided equally among the different axes. There were six different conditions, involving priming rates of 1/50 and 1/150, 1/100 and 1/200, 1/200 and 1/400, 1/400 and 1/800, 1/800 and 1/1600, and 1/833 and 1/2500.

Five simulation conditions were run with the global-unit network with several replications. For three of these, the overall percentage of prime stimuli was held constant at 30% (20 different patterns each at 1/100 and 1/200; 80 different patterns each at 1/400 and 1/800; and 160 different patterns each at 1/800 and 1/1600). For the other two simulations, the overall percentage of prime stimuli was permitted to vary, with 9 and 20 prime patterns respectively at 1/50 and 1/150 (31% prime stimuli), and 20 prime patterns each at 1/833 and 1/2500 (3.2% prime stimuli).

Five simulation conditions were run with the local-unit network with several replications. For two of these, the overall percentage of prime stimuli was held constant at 30% (40 different patterns each at 1/200 and 1/400; and 80 different patterns each at 1/400 and 1/800). For the other three simulations, the overall percentage of prime stimuli was permitted to vary, with 20 different patterns each at 1/100 and 1/200 (30% prime stimuli), at 1/200 and 1/400 (15% prime stimuli), and 1/400 and 1/800 (7.5% stimuli).

Exp. 2: Axis generalization experiment

In the training phase, the networks were presented with patterns symmetric about one of only two axes, either the left and right diagonal axes, for one set of simulations, or the horizontal-vertical axes, for another set of simulations. In the testing phase, the networks were presented with patterns symmetric about one of the four possible axes, as in the priming experiment.

RESULTS

Human Experiments

Exp. 1: Priming experiment

Subjects began training with performance well above chance levels and showed steady skill learning across the 5 blocks of the experiment. With practice, responses to the nonrepeated patterns became increasingly accurate

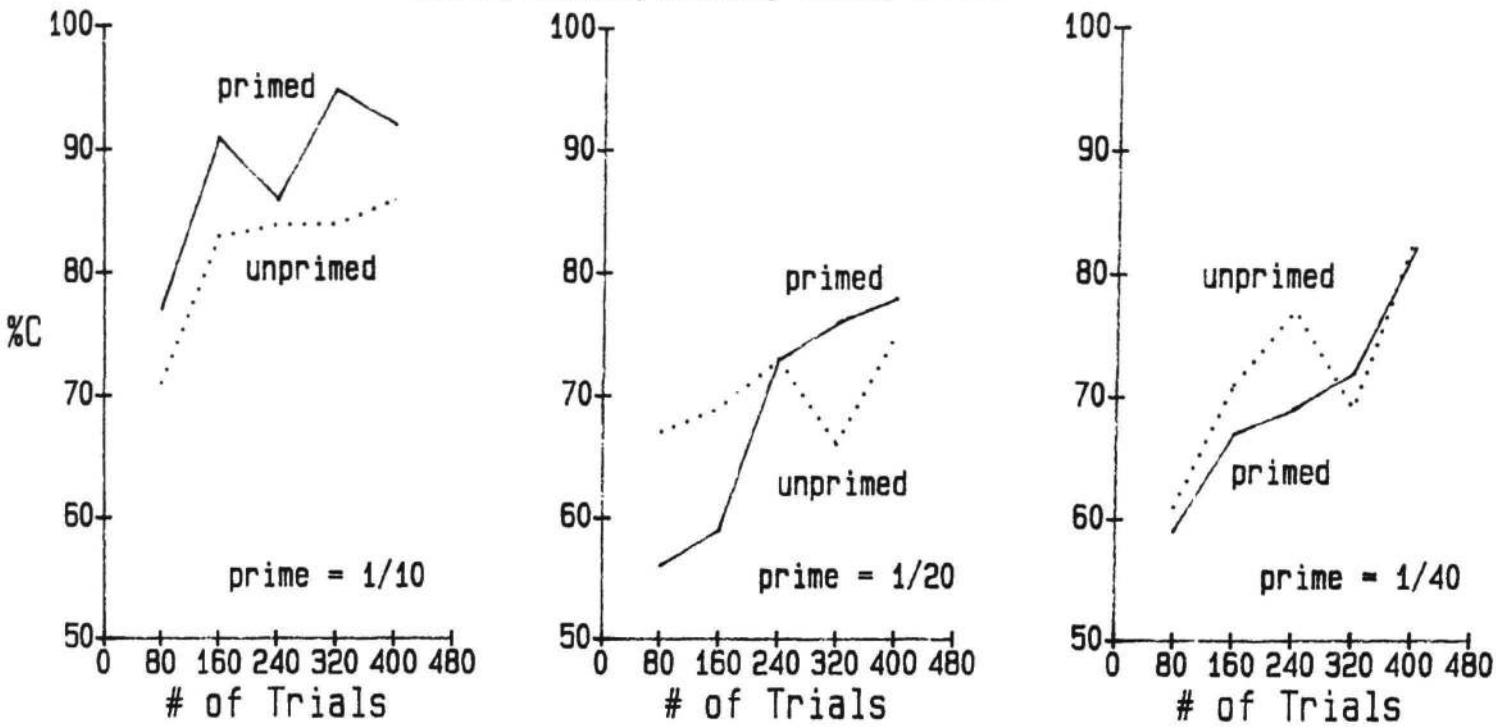


Figure 1. Percent correct across trials for primed and unprimed patterns. Across priming conditions (1/10, 1/20, and 1/40), the overall percentage of prime stimuli was held constant.

(see Figure 1) and had shorter latencies (see Figure 2). Subjects also showed clear repetition priming effects, measured both in terms of percent correct (Figure 1) and reaction time (Figure 2). Repetition priming was evident for all blocks for the 1/10 repetition rate and seemed to emerge by the last two blocks for the 1/20 repetition rate; no priming was observed for the 1/40 repetition rate.

One interesting result concerns performance for the different axes when analyzed separately. Performance was superior for the horizontal and vertical axes, for which subjects presumably have had a good deal more real-world experience prior to the present studies, than for the diagonal axes, which seem to play less of a role in our perceptual experience. The observed difference in performance was maintained across all 5 blocks of the experiment (see Figure 3).

Exp. 2: Axis generalization experiment

Results for the 2-axis training and 4-axis testing phases of this experiment are presented in Figure 4 (only the first 3 of the 5 testing blocks are shown). Data are presented separately for training on horizontal-vertical

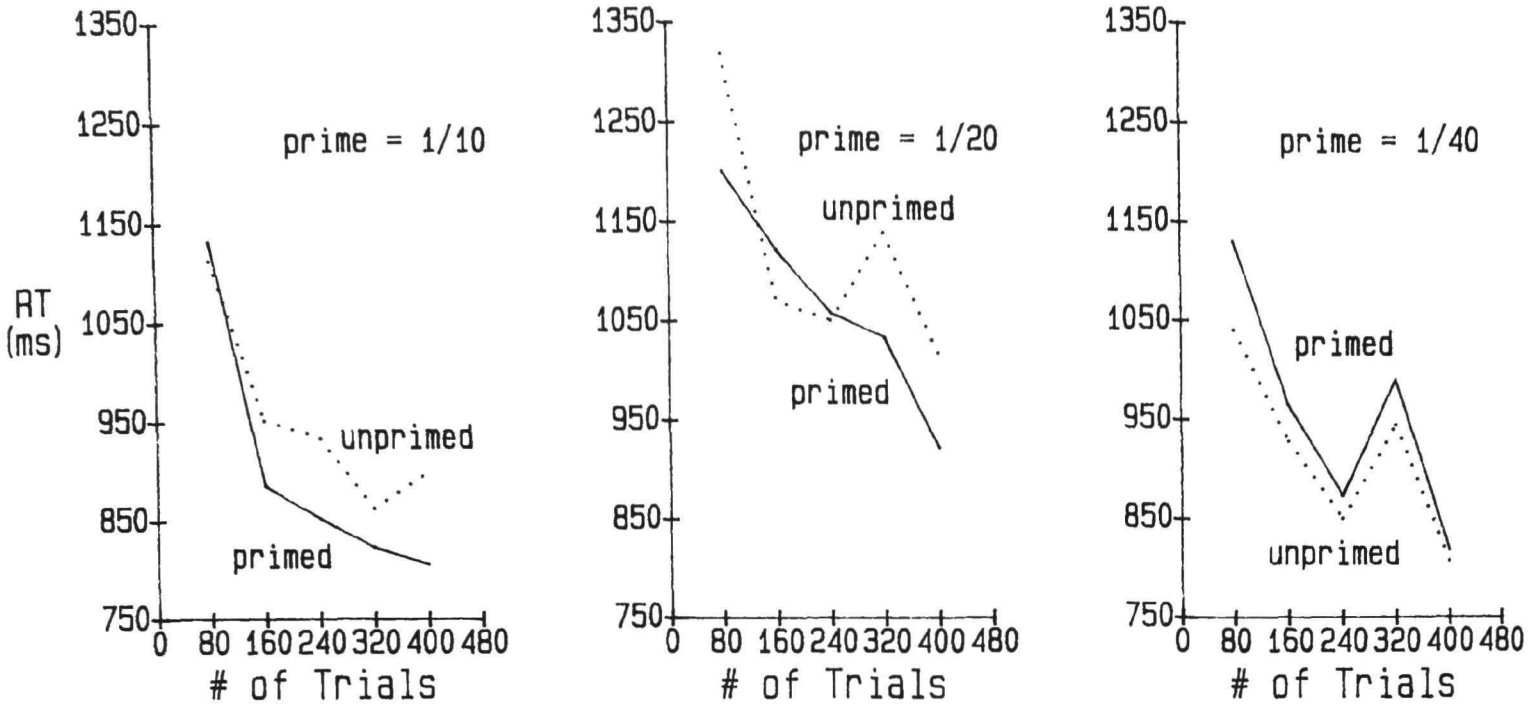


Figure 2. Reaction times (in msec) across trials for primed and unprimed patterns. Across priming conditions (1/10, 1/20, and 1/40), the overall percentage of prime stimuli was held constant.

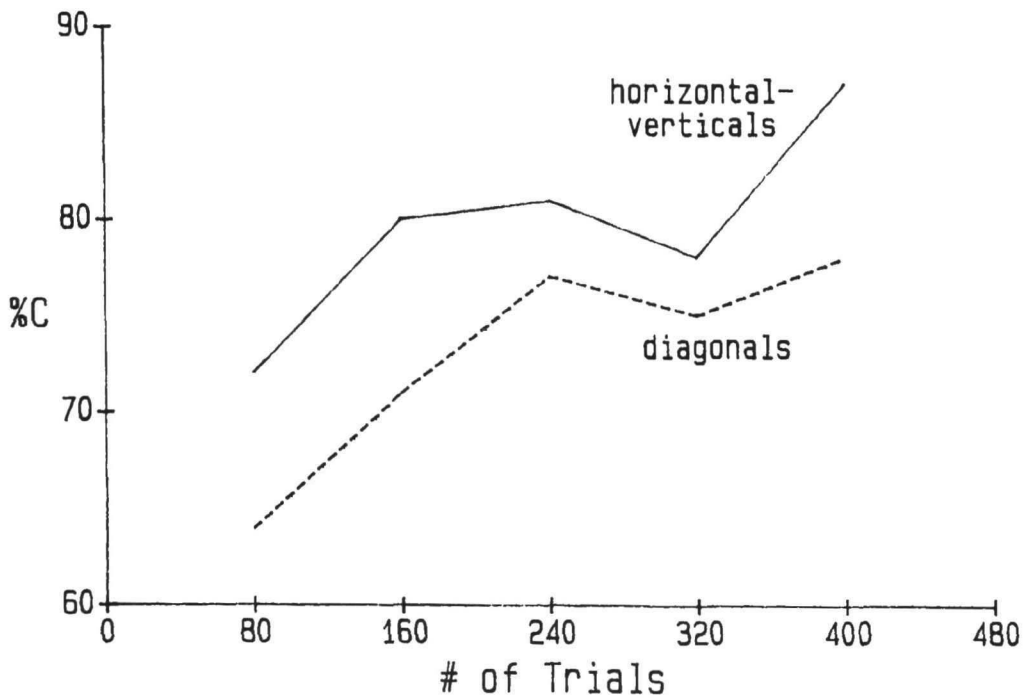


Figure 3. Percent correct across trials for horizontal-vertical or diagonal unprimed boards. Data are averaged across the different priming conditions.

axes versus diagonal axes (middle and bottom panels), and are also averaged together as trained versus untrained axes (top panel). The skill learning exhibited by subjects for new patterns did not generalize across axes. After two blocks of training with two axes, performance on the untrained axes at the start of the 4-axis testing phase was no better than - indeed, on average was poorer than - performance at the beginning of the training phase. Note, however, that these data do not take into account differences in number of response alternatives between the training and testing phases. To the extent that the observed performance includes some amount of guessing distributed

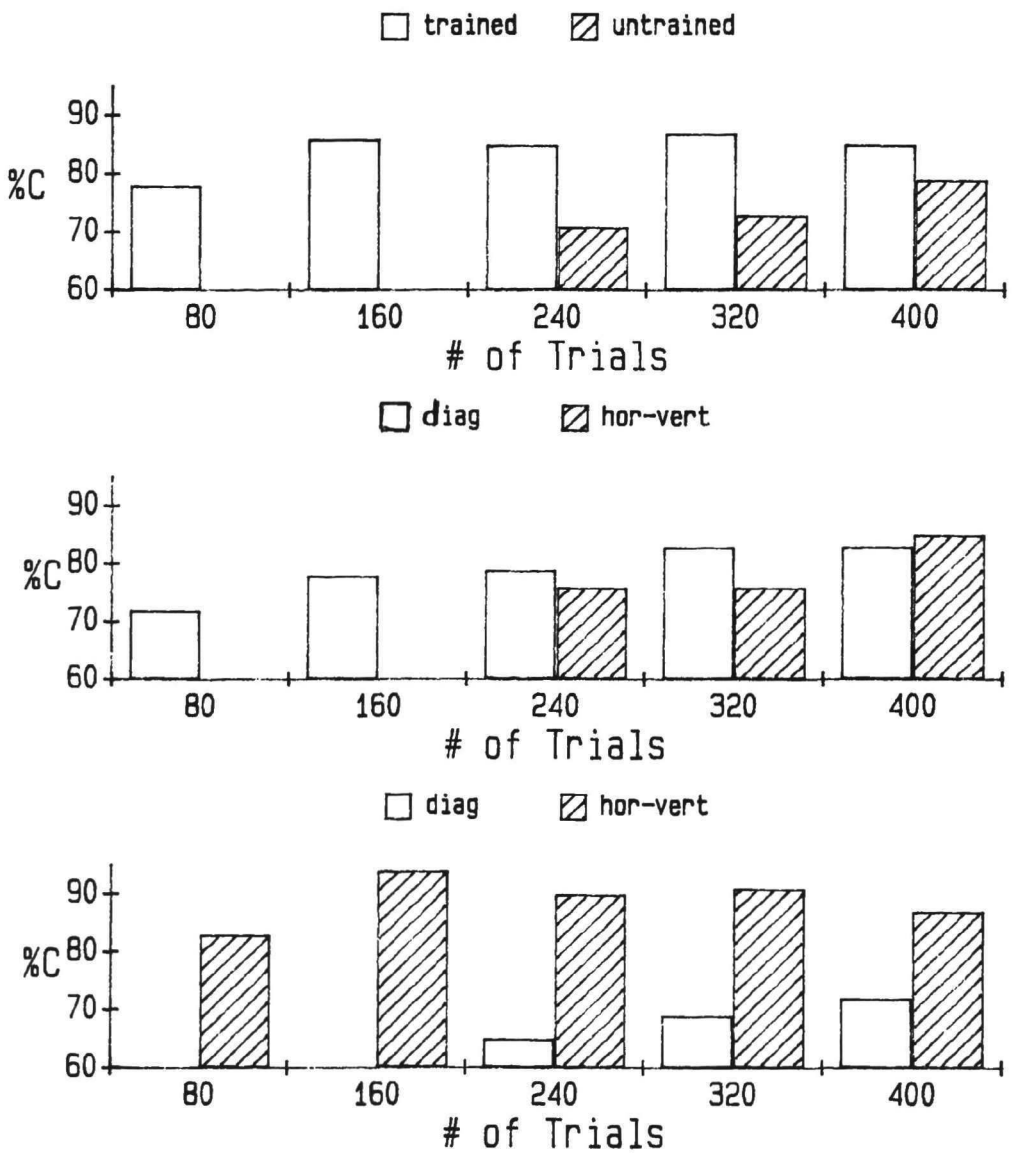


Figure 4. Percent correct across trials when trained on either diagonal (middle) or horizontal-vertical axes (bottom) and tested on all four axes. Overall effect of training is shown at top.

among the set of response alternatives, the contribution of guessing to performance would be different for training than for testing, artificially lowering the scores on the testing phase. Our inspection of the errors and observations of subjects taking the test suggests that rather little contribution is made to the observed performance by distributing guesses among the set of response alternatives, thereby lessening its presumed impact.

The clearest finding of this experiment was the difference between performance for the diagonal axes and performance for the horizontal-vertical axes, amplifying the effect seen in Exp. 1. First, looking only at the training results, subjects' initial performance was better for the horizontal-vertical axes (bottom panel) than for the diagonal axes (middle panel), and there was more improvement over the two training blocks for the horizontal-vertical axes than for the diagonal axes. Moreover, the effects of 2-axis training on 4-axis test performance was dramatically different for the two sets of axes: Training on the horizontal-vertical axes produced a huge advantage in performance for these axes over the diagonal axes when tested in the 4-axis condition (bottom panel), whereas training on the diagonal axes served only to boost performance on these axes to the level of the horizontal-vertical axes when tested in the 4-axis condition (middle panel).

Computer simulations

Exp. 1: Priming experiment

For each network, simulations showed both skill learning for nonrepeated items and repetition priming for repeated items, even over very long lags (see Figures 5-8). The finding of priming in the global-unit network for the 1/1600 condition (see Figure 6) deserves particular emphasis. The priming effect was apparent with fewer than 12 presentations of a given prime pattern spread out over 20,000 stimuli. In fact, in a subsequent study (Cohen, Abrams, Harley, Tabor, Gordon, & Sejnowski, 1986), we have demonstrated priming with as few as 6 presentations of a given prime stimulus spread out among 20,000 patterns. In addition, the priming effect was remarkably reproducible across replications. That is, for those sets of simulations that included overlap among the priming rates (e.g., for the local-unit network, the three simulations had priming rates of 1/100 and 1/200, 1/200 and 1/400, and 1/400 and 1/800; look across panels of Figure 7), the performance of

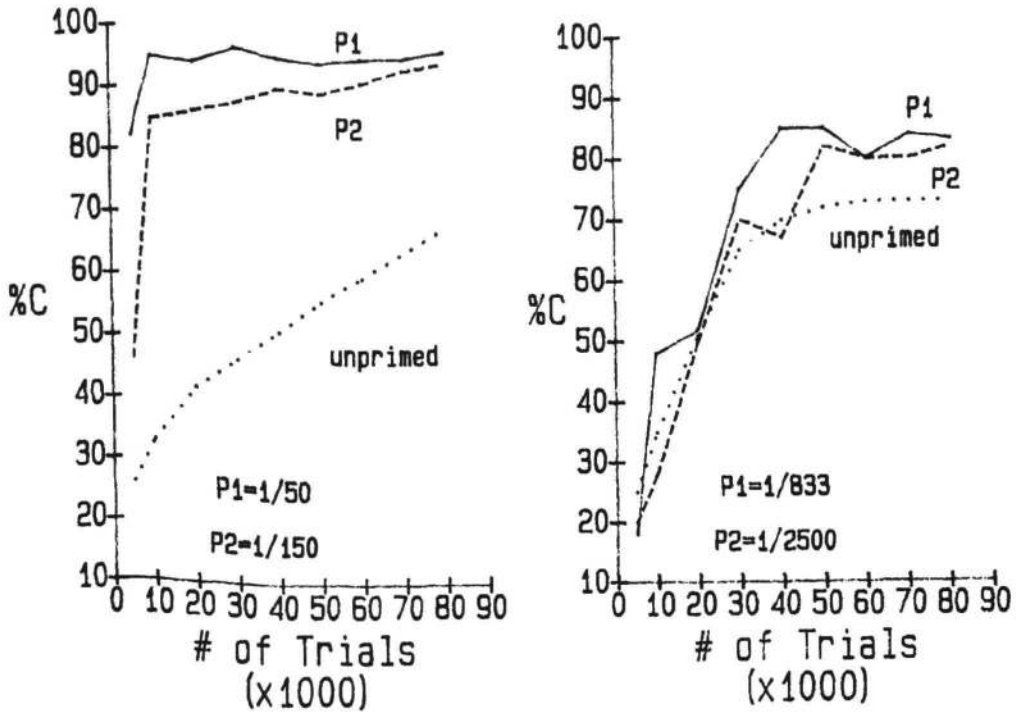


Figure 5. Percent correct for global-unit network for primed and unprimed stimuli. Each of two simulations tested performance on two different priming rates. The overall percentage of prime stimuli varied across simulations.

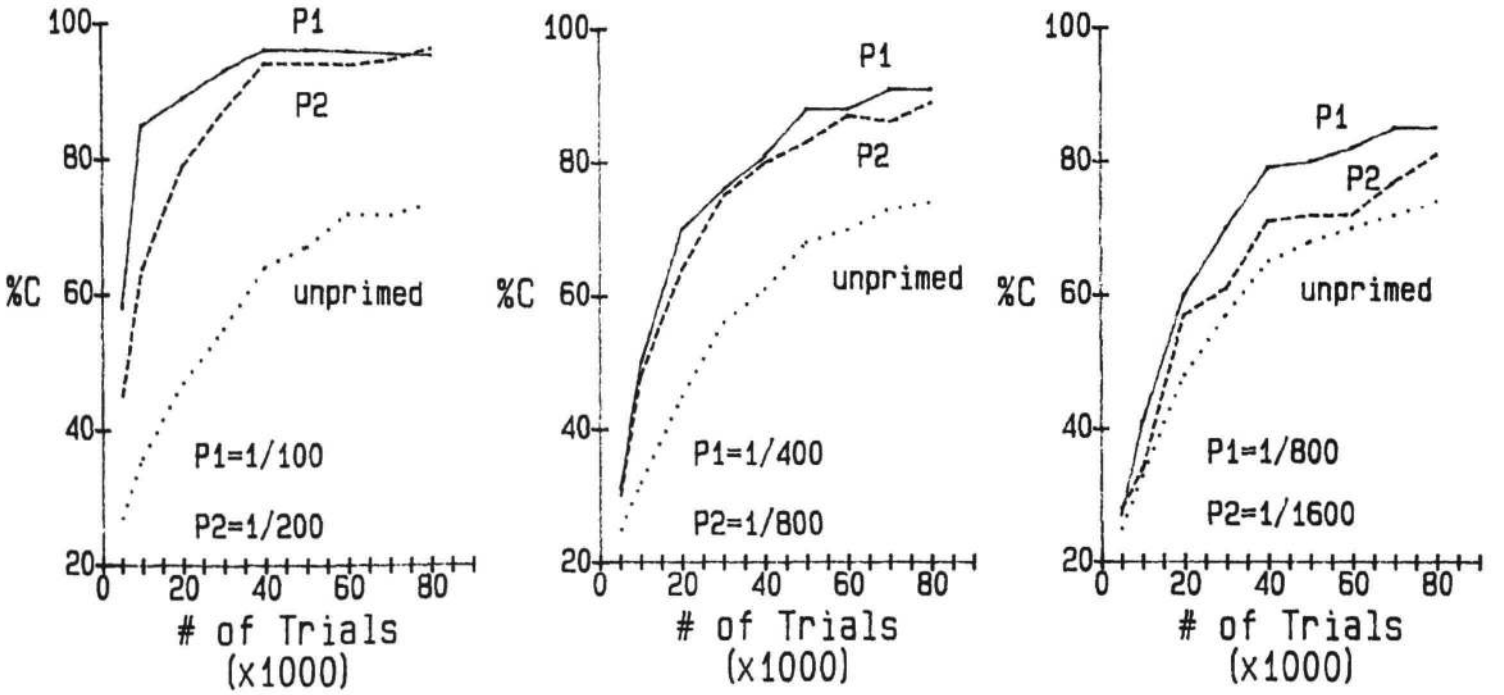


Figure 6. Percent correct for global-unit network for primed and unprimed stimuli. Each of three simulations tested performance on two different priming rates. The overall percentage of prime stimuli was maintained across simulations.

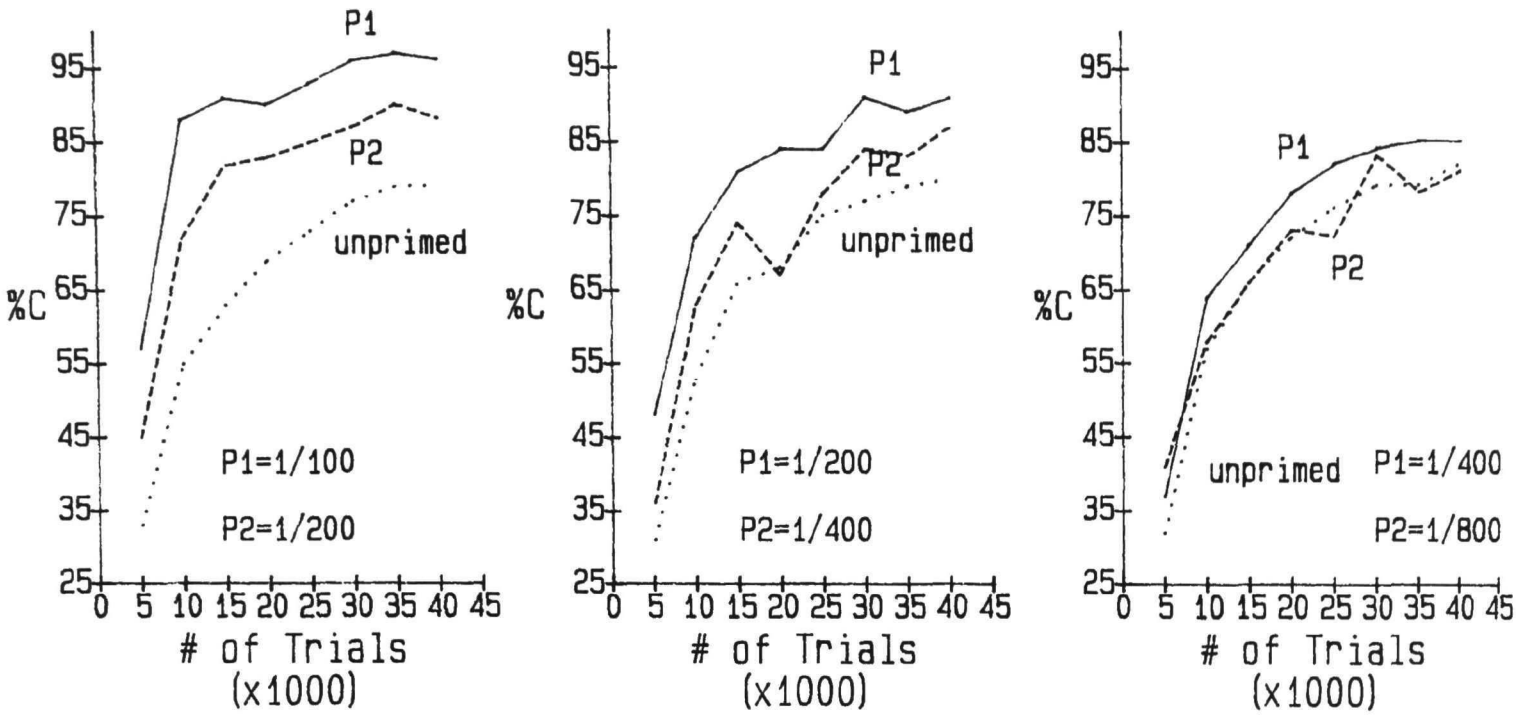


Figure 7. Percent correct for local-unit network for primed and unprimed stimuli. Each of three simulations tested performance on two different priming rates. The overall percentage of prime stimuli was permitted to vary conditions.

network for the same priming rate on different occasions was virtually identical. The performance of both networks for the different axes was nearly invariant, unlike the results for human subjects.

The two networks differed in a number of interesting ways. First, though both networks started learning de novo, the local-unit network learned much more rapidly, reaching a higher level of performance for nonrepeated stimuli in 40,000 trials (see Figures 7 & 8) than did the global-unit network in 80,000 trials (see Figures 5 & 6). Second, despite the superior skill learning of the local-unit network, the global-unit network was the more sensitive one to priming, showing a priming effect even at a priming rate of 1/1600 (see Fig. 6) and at 1/3200 (Cohen et al., 1986), whereas the local-unit network showed only marginal priming at a priming rate of 1/800 (see Figures 7 & 8). Third, the global-unit network but not the local-unit network showed a trade-off between skill learning and repetition priming as a function of priming rate when the overall percentage of prime stimuli was permitted to vary. Thus, performance for the prime stimuli was better in the 1/50 and 1/150 condition than in the 1/833 and 1/2500 condition, whereas performance for the nonrepeated stimuli was poorer in the 1/50 and 1/150 condition than in

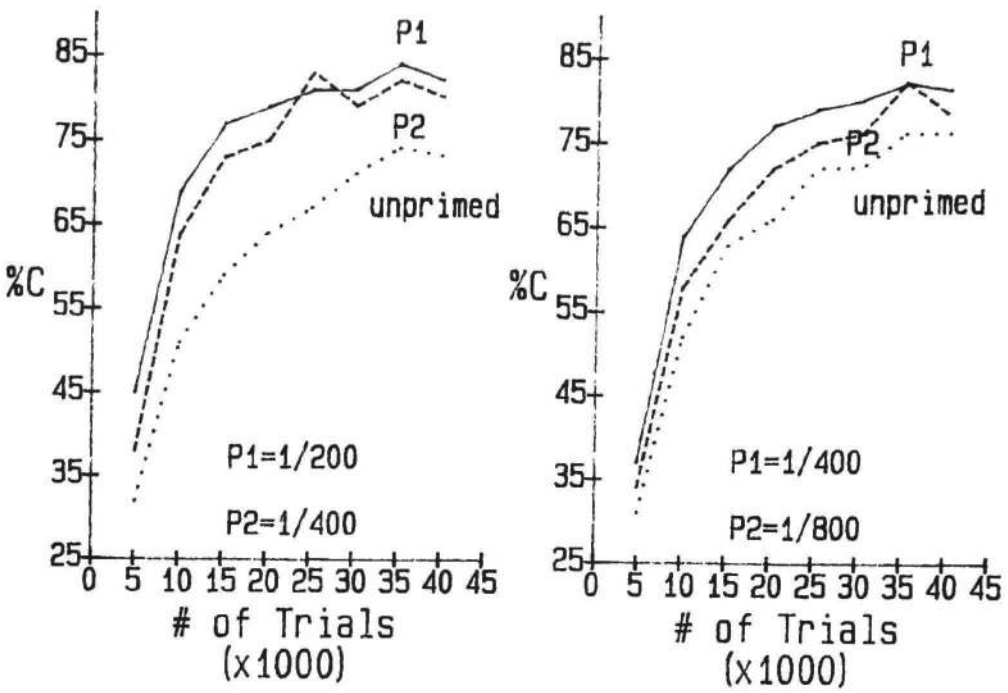


Figure 8. Percent correct for local-unit network for primed and unprimed stimuli. Each of two simulations tested performance on two different priming rates. The overall percentage of prime stimuli was maintained across simulations.

the 1/833 and 1/2500 condition (see Figure 5). The trade-off disappeared, however, when the overall percentage of prime stimuli was held constant across the different priming rate conditions (see Figure 6). For the local-unit network, there was no trade-off regardless of whether the overall percentage of prime stimuli varied or was constant (see Figures 7 & 8).

Inspection of the weights for the hidden units of the two networks was illuminating. The hidden units in the global-unit network behaved in the manner reported by Sejnowski et al. (1986). The weights were frequently antisymmetric about one or more axes, and often were also symmetric about one or more other axes. The amount of antisymmetry was striking. The spatial distribution of the weights to input units corresponding to different portions of the receptive field varied considerably among hidden units and represented a number of different types of geometric features as well as some isolated cells in the array. Many of the geometrical features, such as linear stripes or angles, were quite global, spanning the entire width or length of the array. Frequently, such features were represented in one hidden unit along with their complement in a different unit. Finally, the weights to some

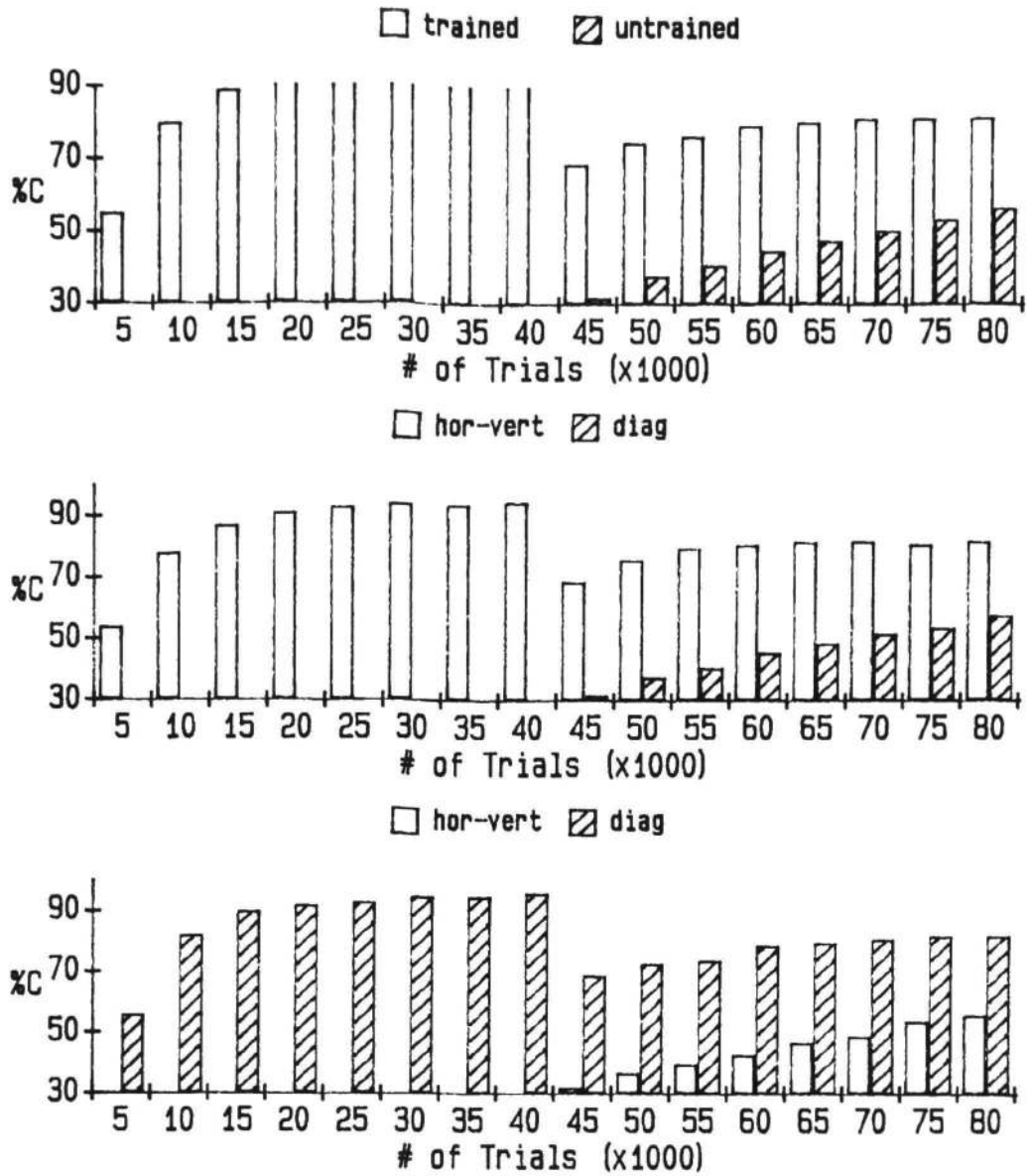


Figure 9. Percent correct for global-unit network when trained on either horizontal-vertical (middle) or diagonal axes (bottom) and tested on all four axes. Overall training effect is shown at top.

hidden units were very small, suggesting that these units played little or no role in the solution to the problem.

The hidden units in the local-unit network behaved in a similar fashion, although their receptive fields were quite local, being restricted to particular 3x3 sections of the 9x9 input array. Here, too, the number of units whose weights were set up to detect antisymmetry was striking. The presence of hidden units with weights representing geometrical features and their complement in a unit with a matching receptive field was noted here as well. An interesting aspect of the hidden units in this network was the

apparent reliance of the network on units with receptive fields in the center 3x3 section of the input array. This section is relevant to detection of any of the four axes. Accordingly, the weights for such units were typically sensitive to multiple axes and, in particular, were antisymmetric for one or more axes while being symmetric for one or more others.

Exp. 2: Axis generalization experiment

Both networks trained on a particular pair of axes showed a decrement in overall performance when switched to testing with all 4 axes (see Figures 9 &

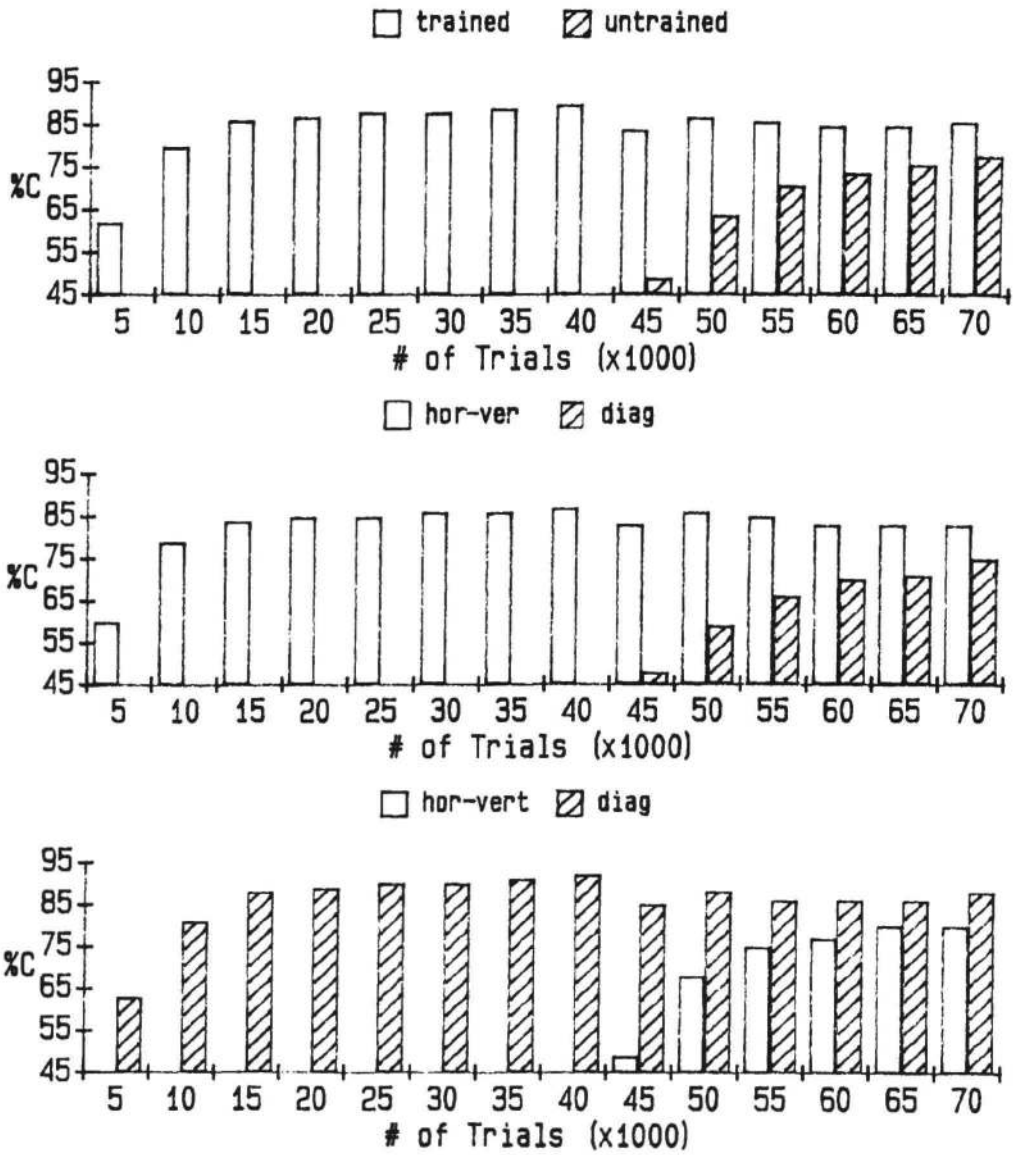


Figure 10. Percent correct for local-unit network when trained on either horizontal-vertical (middle) or diagonal axes (bottom) and tested on all four axes. Overall training effect is shown at top.

10). Not only did performance decline for the trained axes, especially in the global-unit network, but performance on the previously untrained axes was poorer at the start of the 4-axis testing than was performance for the other pair of axes at the start of the 2-axis training. Again, this was particularly evident for the global-unit network. Unlike the performance of human subjects, neither of the networks showed a preference for the horizontal-vertical axes over the diagonal axes, or vice versa. (Note that these data, like those presented for human subjects, do not take into account the difference between the training and testing phases in number of response alternatives. Preliminary data from simulations in which the networks are trained on 2 axes and then tested on only the other 2 axes suggests that the previously untrained axes neither benefit nor suffer from prior training on other axes.)

The weights for the hidden units in the two networks were similar to those discussed above. One finding of interest that comes from this particular task was the preponderance of hidden units in the local-unit network that were responsive to both of the two axes during the 2-axis training phase. Inspection of the hidden units with receptive fields in the center 3x3 section of the input array revealed an organization of weights that nearly always was symmetric for one axis and antisymmetric for the other. For the other units, in most of the cases in which the unit was responsive to only one axis, it was responsive to antisymmetry in that axis. Hidden units with receptive fields in portions of the input array not relevant to distinguishing between the axes being tested had small weights and were inactive. Upon switching to 4-axis testing, these inactive units rapidly took on weights appropriate to the new axes, making this network rather more responsive to the change in stimulus parameters than was the global-unit network. The global-unit network was forced to reconfigure its hidden-unit weights to conform to the new stimulus parameters.

DISCUSSION

The present work documents skill learning and repetition priming in symmetry detection for both human subjects and the Boltzmann machine. The behavioral data from humans and the simulation data from the network models share certain qualitative similarities, but are different in some significant

quantitative and qualitative respects. The difference quantitatively between human and model performance is striking: The number of trials required for the networks to learn the task is some two orders of magnitude greater than that required by humans. Moreover, the performance of the networks for unprimed stimuli seems to plateau at approximately 70-80% correct, depending upon the condition, whereas the performance of our human subjects continues to improve with practice, and will approach 100% correct with enough trials. Note, however, that humans come to this task with considerable topographic knowledge and familiarity with symmetry detection, and their initial performance on the task is well above chance. By contrast, the Boltzmann machine starts its learning de novo; it must learn simultaneously about both symmetry and the critical aspects of symmetry detection for these particular materials. Its performance on the present task starts at chance. In addition to this difference in what must be learned, the discrepancy between the relatively small number of units used in the network models compared with the presumably huge number of neurons in the human visual system sensitive to symmetry is likely to be an important factor. Finally, note that it is extraordinarily difficult to know how to scale the performance of the network models vis-a-vis human performance in the absence of specific claims about the psychological relevance of "trials" or "machine operations" for these models. Until such claims are made, conclusions based on quantitative comparisons can be offered only tentatively.

In terms of qualitative comparisons of the experimental data and simulation data, it is noteworthy that humans showed a strong superiority for detecting vertical symmetries. The superiority of performance in both experiments for horizontal-vertical axes over diagonal axes was due overwhelmingly to superior performance for the vertical axis. Across the 5 blocks of Exp. 1, performance for the vertical axis (84% correct) showed nearly as much advantage over performance for the horizontal axis (77% correct) as for the diagonal axes (74% correct). This was as true in the initial block (% correct: v = 75, h = 66, d = 63) as in the final block (% correct: v = 91, h = 84, d = 81). This result is consistent with previous psychophysical work (Barlow & Reeves, 1979), but at odds with the performance of the networks. The superiority for vertical symmetries is thought to be due at least in part to the sensitivity of cells near the vertical midline to

local features around the midline. Other features relevant to processing mirror symmetries must be processed more globally. Our modeling attempts included both a local-unit and global-unit network, a difference in connectivity of the hidden units that had important implications for the way in which the model was able to internally represent mirror symmetries. The global-unit network, with less pre-wired structure than the local-unit network, was more influenced by the structure imposed by the stimulus array: It remained responsive to primes over long lags and was more sensitive to the change from 2 to 4 axes. The limited receptive fields of processing elements in the local-unit network would seem to provide a closer model of the type of processing used by biological visual systems, although the connectivity among the local processing units actually established by learning in our networks may be quite different from that attained by biological systems. Evaluation of which of these networks provides a closer match to human performance awaits further testing.

The most important finding of the present studies is the emergence of priming from the network model without any additional mechanisms, such as short-term changes to the weights. These networks were designed to handle generalization, which is the essence of skill learning; the fact that they prove to be sensitive to individual items even over enormous lags is a striking finding, one that further links skill learning and repetition priming. This is the basic qualitative similarity between the simulation and human performance data.

Two conclusions can be drawn from the study of symmetry detection in humans and massively-parallel network models. First, priming for humans in this task cannot depend upon the activation of some pre-existing representation of the primed materials, since the stimuli used in these studies were novel. Second, priming is a latent property of incremental learning in the parallel network model and occurs from the start of de novo learning. Thus, both the behavioral data from humans and simulation data from the model suggest that repetition priming is not a separate mechanism and does not depend on an already structured system. Priming may instead be an integral feature of the basic learning process that configures and reorganizes processing modules. Parallel work with humans and network models, such as that reported in the present paper, can provide important insights about

possible mechanisms of learning and memory.

REFERENCES

Ackley, D.H., Hinton, G.E., & Sejnowski, T.J. (1985). A learning algorithm for Boltzmann machines. Cognitive Science, 9, 147-169.

Anderson, J.R. (1982). Acquisition of cognitive skill. Psychological Review, 89, 369-406.

Anderson, J.R. (1983). The Architecture of Cognition. Cambridge: Harvard University Press.

Barlow, H.B. & Reeves, B.C. (1979). The versatility and absolute efficiency of detecting mirror symmetry in random dot displays. Vision Research, 19, 783-793.

Bruce, V.G. & Morgan, M.J. (1975). Violations of symmetry and repetition in visual patterns. Perception, 4, 239-249.

Biederman, I. (1985). Human image understanding: Recent research and a theory. Computer Vision, Graphics, and Image Processing, 32, 29-73.

Carpenter, G.A. & Grossberg, S. (1986). Neural dynamics of category learning and recognition: Attention, memory consolidation, and amnesia. In J. Davis, R. Newburgh, & E. Wegman (Eds.), Brain Structure, Learning, and Memory. AAAS Symposium Series.

Cohen, N.J. (1984). Preserved learning capacity in amnesia: Evidence for multiple memory systems. In L.R. Squire & N. Butters (Eds.), Neuropsychology of Memory. New York: Guilford Press.

Cohen, N.J., Eichenbaum, H., DeAcedo, B.S., & Corkin, S. (1985). Different memory systems underlying acquisition of procedural and declarative knowledge. Annals of the New York Academy of Sciences, 444, 54-71.

Delius, J.D. & Nowak, B. (1982). Visual symmetry recognition by pigeons. Psychological Research, 44, 199-212.

Fowler, C.A., Napps, S.E., & Feldman, L. (1985). Relations among regular and irregular morphologically related words in the lexicon as revealed by repetition priming. Memory & Cognition, 13, 241-255.

Garner, W.R. (1962). Uncertainty and Structure as Psychological Concepts. New York: Wiley.

Gordon, B. (1986). Preserved learning of novel information in amnesia: Evidence for multiple memory systems. (submitted)

Hinton, G.E. & Sejnowski, T.J. (1983). Optimal perceptual inference. Proceedings of the IEEE Computer Society Conference on Computer Vision & Pattern Recognition, Washington, DC, 448-453.

Hopfield, J.J. (1982). Neural networks and physical systems with emergent collective computational abilities, Proceedings of the National Academy of Sciences (USA), 79, 2554-2558.

Kohler, W. (1929). Gestalt Psychology. NY: Harcourt Brace.

Kohlers, P.A. (1975). A pattern-analyzing basis of recognition. In L.S. Cermak & F.I.M. Craik (Eds.), Levels of Processing in Human Memory. Hillsdale, NJ: L. Erlbaum.

Marr, D. & Nishihara, H.K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. Perceedings of the Royal Society (London; series B), 200, 269-294.

Minsky, M. & Papert, S. (1969). Perceptrons. Cambridge: MIT Press.

Rosenblatt, F. (1959). Principles of Neurodynamics. NY: Spartan.

Rumelhart, D.E. & McClelland, J.L. (1986). Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Cambridge: MIT Press.

Schacter, D.L. (1985). Priming of old and new knowledge in amnesic patients and normal subjects. Annals of the New York Academy of Sciences, 444, 41-53.

Sejnowski, T.J. & Hinton, G.E. (1986). Separating figure from ground with a Boltzmann Machine. In M.A. Arbib & A.R. Hanson (Eds.), Vision, Brain & Cooperative Computation. Cambridge: MIT Press.

Sejnowski, T.J., Kienker, P.K., & Hinton, G.E. (in press). Learning symmetry groups with hidden units: Beyond the perception. Physica D.

Squire, L.R. & Butters, N. (1984). Neuropsychology of Memory. NY: Guilford Press.

Squire, L.R. & Cohen, N.J. (1984). Human memory and amnesia. In J.L. McGaugh, N.W. Weinberger, & G. Lynch (Eds.), Neurobiology of Learning and Memory. NY: Guilford Press.

Tulving, E. (1985). How many memory systems are there? American Psychologist.

ADAPTIVE RESONANCE THEORY: STABLE SELF-ORGANIZATION OF NEURAL RECOGNITION CODES IN RESPONSE TO ARBITRARY LISTS OF INPUT PATTERNS

Gail A. Carpenter† and Stephen Grossberg‡
Center for Adaptive Systems
Department of Mathematics
Boston University
Boston, Massachusetts 02215

1. SELF-ORGANIZATION OF NEURAL RECOGNITION CODES

A neural network, called an *adaptive resonance theory* (ART) architecture, for the learning of recognition categories is described herein. Real-time network dynamics for this model have been completely characterized through mathematical analysis and computer simulations. The architecture self-organizes and self-stabilizes its recognition codes in response to arbitrary orderings of arbitrarily many and arbitrarily complex binary input patterns. Top-down attentional and matching mechanisms are critical in self-stabilizing the code learning process. The architecture embodies a parallel search scheme which updates itself adaptively as the learning process unfolds. After learning self-stabilizes, the search process is automatically disengaged. Thereafter input patterns directly access their recognition codes without any search. Thus recognition time does not grow as a function of code complexity. A novel input pattern can directly access a category if it shares invariant properties with the set of familiar exemplars of that category. These invariant properties emerge in the form of learned critical feature patterns, or prototypes. The architecture possesses a context-sensitive self-scaling property which enables its emergent critical feature patterns to form. They detect and remember statistically predictive configurations of featural elements which are derived from the set of all input patterns that are ever experi-

† Supported in part by the Air Force Office of Scientific Research (AFOSR 85-0149 and AFOSR F49620-86-C-0037), the Army Research Office (ARO DAAG-29-85-K-0095), and the National Science Foundation (NSF DMS-84-13119).

‡ Supported in part by the Air Force Office of Scientific Research (AFOSR 85-0149 and AFOSR F49620-86-C-0037) and the Army Research Office (ARO DAAG-29-85-K-0095).

Acknowledgements: We wish to thank Cynthia Suchta and Carol Yanakakis for their valuable assistance in the preparation of the manuscript.

enced. Four types of attentional process—priming, gain control, vigilance, and intermodal competition—are mechanistically characterized. Top-down priming and gain control are needed for code matching and self-stabilization. Attentional vigilance determines how fine the learned categories will be. If vigilance increases due to an environmental disconfirmation, then the system automatically searches for and learns finer recognition categories. A new nonlinear matching law (the 2/3 Rule) and new nonlinear associative laws (the Weber Law Rule, the Associative Decay Rule, and the Template Learning Rule) are needed to achieve these properties. All the rules describe emergent properties of parallel network interactions. The architecture circumvents the noise, saturation, capacity, orthogonality, and linear predictability constraints that limit the codes which can be stably learned by alternative recognition models. In addition, ART circuits have elsewhere been used to analyse data about speech perception, word recognition and recall, visual perception, olfactory coding, evoked potentials, thalamocortical interactions, attentional modulation of critical period termination, and amnesias (Banquet and Grossberg, 1986; Carpenter and Grossberg, 1985a, 1985b, 1986a, 1986b, 1986c; Grossberg, 1976a, 1976b, 1978a, 1980, 1986a; Grossberg and Stone, 1986a, 1986b). In the following pages, we describe intuitively some key properties of the model.

2. STABILITY-PLASTICITY DILEMMA: MULTIPLE INTERACTING MEMORY SYSTEMS

An adequate self-organizing recognition system must be capable of plasticity in order to learn about significant new events, yet it must also remain stable in response to irrelevant or often repeated events. In order to prevent the relentless degradation of its learned codes by the "blooming, buzzing confusion" of irrelevant experience, an ART system is sensitive to *novelty*. It is capable of distinguishing between familiar and unfamiliar events, as well as between expected and unexpected events.

Multiple interacting memory systems are needed to monitor and adaptively react to the novelty of events. Within ART, interactions between two functionally complementary subsystems are needed to process familiar and unfamiliar events. Familiar events are processed within an attentional subsystem. This subsystem establishes ever more precise internal representations of and responses to familiar events. It also builds up the learned top-down expectations that help to stabilize the learned bottom-up codes of familiar events. By itself, however, the attentional subsystem is unable simultaneously to maintain stable representations of familiar categories and to create new categories for unfamiliar patterns. An isolated attentional subsystem is either rigid and incapable of creating new categories for unfamiliar patterns, or unstable and capable of ceaselessly recoding the categories of

familiar patterns.

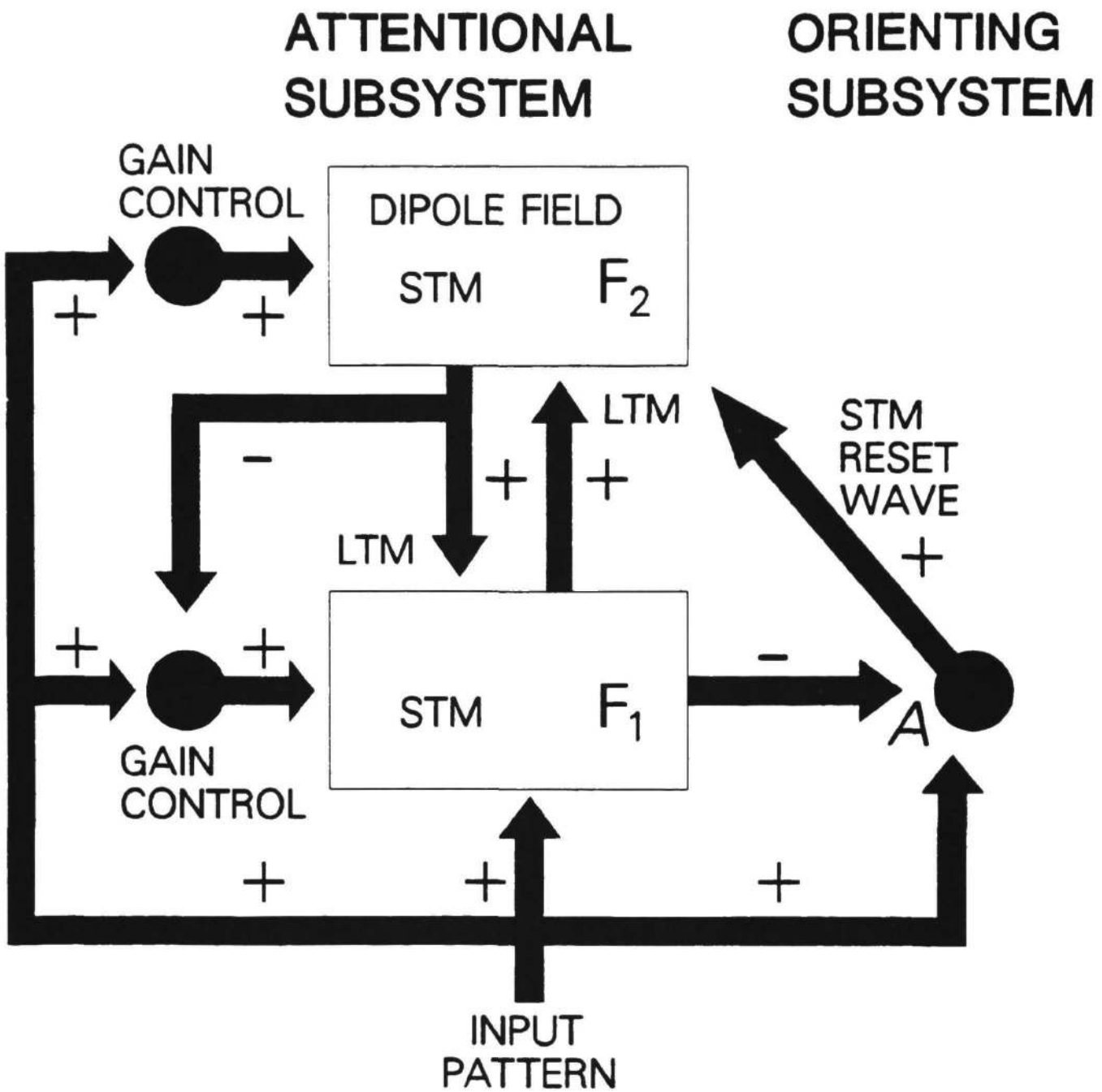
The second subsystem is an orienting subsystem that resets the attentional subsystem when an unfamiliar event occurs. The orienting subsystem is essential for expressing whether a novel pattern is familiar and well represented by an existing recognition code, or unfamiliar and in need of a new recognition code. Figure 1 schematizes the architecture that is analysed herein.

An ART system dynamically reorganizes its recognition codes to preserve its stability-plasticity balance as its internal representations become increasingly complex and differentiated through learning. By contrast, many classical adaptive pattern recognition systems become unstable when they are confronted by complex input environments. Unlike many alternative models the present model can deal with arbitrary combinations of binary input patterns. In particular, it places no orthogonality or linear predictability constaints upon its input patterns. The model computations remain sensitive no matter how many input patterns are processed. The model does not require that very small, and thus noise-degradable, increments in memory be made in order to avoid saturation of its cumulative memory. The model can store arbitrarily many recognition categories in response to input patterns that are defined on arbitrarily many input channels. Its memory matrices need not be square, so that no restrictions on memory capacity are imposed by the number of input channels. Finally, all the memory of the system can be devoted to stable recognition learning. It is not he case that the number of stable classifications is bounded by some fraction of the number of input channels or patterns. Thus a primary goal of the present article is to intuitively describe neural networks capable of self-stabilizing the self-organization of their recognition codes in response to an arbitrarily complex environment of input patterns.

Four properties are basic to the workings of the networks that we characterize herein.

A. Self-Scaling Computational Units: Critical Feature Patterns

Properly defining signal and noise in a self-organizing system raises a number of subtle issues. Pattern context must enter the definition so that input features which are treated as irrelevant noise when they are embedded in a given input pattern may be treated as informative signals when they are embedded in a different input pattern. The system's unique learning history must also enter the definition so that portions of an input pattern which are treated as noise when they perturb a system at one stage of its self-organization may be treated as signals when they perturb the same system at a different stage of its self-organization. The present systems automatically self-scale their computational units to embody context- and learning-dependent definitions of signal and noise.



1. Anatomy of the attentional-orienting system: Two successive stages, F_1 and F_2, of the attentional subsystem encode patterns of activation in short term memory (STM). Bottom-up and top-down pathways between F_1 and F_2 contain adaptive long term memory (LTM) traces which multiply the signals in these pathways. The remainder of the circuit modulates these STM and LTM processes. Modulation by gain control enables F_1 to distinguish between bottom-up input patterns and top-down priming, or template, patterns, as well as to match these bottom-up and top-down patterns. Gain control signals also enable F_2 to react supraliminally to signals from F_1 while an input pattern is on. The orienting subsystem generates a reset wave to F_2 when mismatches between bottom-up and top-down patterns occur at F_1. This reset wave selectively and enduringly inhibits active F_2 cells until the input is shut off.

One property of these self-scaling computational units is schematized in Figure 2. In Figure 2a, each of the two input patterns is composed of three features. The patterns agree at two of the three features, but disagree at the third feature. A mismatch of one out of three features may be designated as informative by the system. When this occurs, these mismatched features are treated as signals which can elicit learning of distinct recognition codes for the two patterns. Moreover, the mismatched features, being informative, are incorporated into these distinct recognition codes.

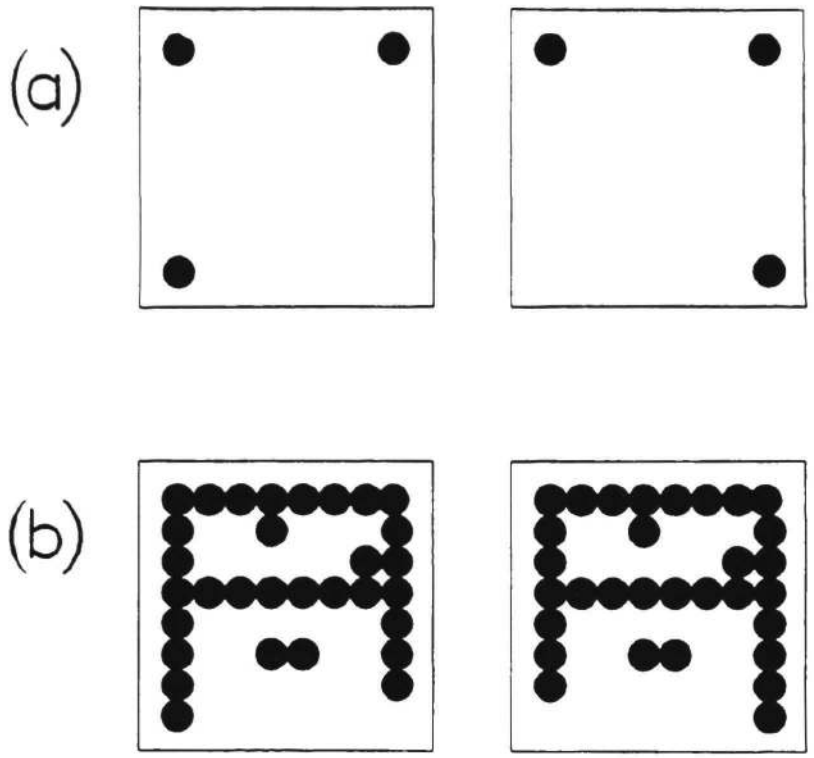
In Figure 2b, each of the two input patterns is composed of thirty-one features. The patterns are constructed by adding identical subpatterns to the two patterns in Figure 2a. Thus the input patterns in Figure 2b disagree at the same features as the input patterns in Figure 2a. In the patterns of Figure 2b, however, this mismatch is less important, other things being equal, than in the patterns of Figure 2a. Consequently, the system may treat the mismatched features as noise. A single recognition code may be learned to represent both of the input patterns in Figure 2b. The mismatched features would not be learned as part of this recognition code because they are treated as noise.

The assertion that *critical feature patterns* are the computational units of the code learning process summarizes this self-scaling property. The term *critical feature* indicates that not all features are treated as signals by the system. The learned units are *patterns* of critical features because the perceptual context in which the features are embedded influences which features will be processed as signals and which features will be processed as noise. Thus a feature may be a critical feature in one pattern (Figure 2a) and an irrelevant noise element in a different pattern (Figure 2b).

The need to overcome the limitations of featural processing with some type of contextually sensitive pattern processing has long been a central concern in the human pattern recognition literature. Experimental studies have led to the general conclusions that "The trace system which underlies the recognition of patterns can be characterized by a central tendency and a boundary" (Posner, 1973, p.54), and that "just listing features does not go far enough in specifying the knowledge represented in a concept. People also know something about the relations between the features of a concept, and about the variability that is permissible on any feature" (Smith and Medin, 1981, p.83). We illustrate herein how these properties may be achieved using self-scaling computational units such as critical feature patterns.

B. Self-Adjusting Memory Search

No pre-wired search algorithm, such as a search tree, can maintain its efficiency as a knowledge structure evolves due to learning in a unique input environment. A search order



2. Self-scaling property discovers critical features in a context-sensitive way: (a) Two input patterns of 3 features mismatch at 1 feature. When this mismatch is sufficient to generate distinct recognition codes for the two patterns, the mismatched features are encoded in LTM as part of the critical feature patterns of these recognition codes. (b) Identical subpatterns are added to the two input patterns in (a). Although the new input patterns mismatch at the same one feature, this mismatch may be treated as noise due to the additional complexity of the two new patterns. Both patterns may thus learn to activate the same recognition code. When this occurs, the mismatched feature is deleted from LTM in the critical feature pattern of the code.

that may be optimal in one knowledge domain may become extremely inefficient as that knowledge domain becomes more complex due to learning.

The ART system considered herein is capable of a parallel memory search that adaptively updates its search order to maintain efficiency as its recognition code becomes arbitrarily complex due to learning. This self-adjusting search mechanism is part of the network design whereby the learning process self-stabilizes by engaging the orienting subsystem (Section 5).

None of these mechanisms is akin to the rules of a serial computer program. Instead, the circuit architecture as a whole generates a self-adjusting search order and self-stabilization as emergent properties that arise through system interactions. Once the ART architecture is in place, a little randomness in the initial values of its memory traces, rather than a carefully wired search tree, enables the search to carry on until the recognition code

self-stabilizes.

C. Direct Access to Learned Codes

A hallmark of human recognition performance is the remarkable rapidity with which familiar objects can be recognized. The existence of many learned recognition codes for alternative experiences does not necessarily interfere with rapid recognition of an unambiguous familiar event. This type of rapid recognition is very difficult to understand using models wherein trees or other serial algorithms need to be searched for longer and longer periods as a learned recognition code becomes larger and larger.

In an ART model, as the learned code becomes globally self-consistent and predictively accurate, the search mechanism is automatically disengaged. Subsequently, no matter how large and complex the learned code may become, familiar input patterns *directly access*, or activate, their learned code, or category. Unfamiliar patterns can also directly access a learned category if they share invariant properties with the critical feature pattern of the category. In this sense, the critical feature pattern acts as a prototype for the entire category. As in human pattern recognition experiments, an input pattern that matches a learned critical feature pattern may be better recognized than any of the input patterns that gave rise to the critical feature pattern (Posner, 1973; Posner and Keele, 1968, 1970).

Unfamiliar input patterns which cannot stably access a learned category engage the self-adjusting search process in order to discover a network substrate for a new recognition category. After this new code is learned, the search process is automatically disengaged and direct access ensues.

D. Environment as a Teacher: Modulation of Attentional Vigilance

Although an ART system self-organizes its recognition code, the environment can also modulate the learning process and thereby carry out a teaching role. This teaching role allows a system with a fixed set of feature detectors to function successfully in an environment which imposes variable performance demands. Different environments may demand either coarse discriminations or fine discriminations to be made among the same set of objects. As Posner (1973, pp.53-54) has noted:

> "If subjects are taught a tight concept, they tend to be very careful about classifying any particular pattern as an instance of that concept. They tend to reject a relatively small distortion of the prototype as an instance, and they rarely classify a pattern as a member of the concept when it is not. On the other hand, subjects learning high-variability concepts often falsely classify patterns as members of the concept, but rarely reject a member of the concept incorrectly...The situation

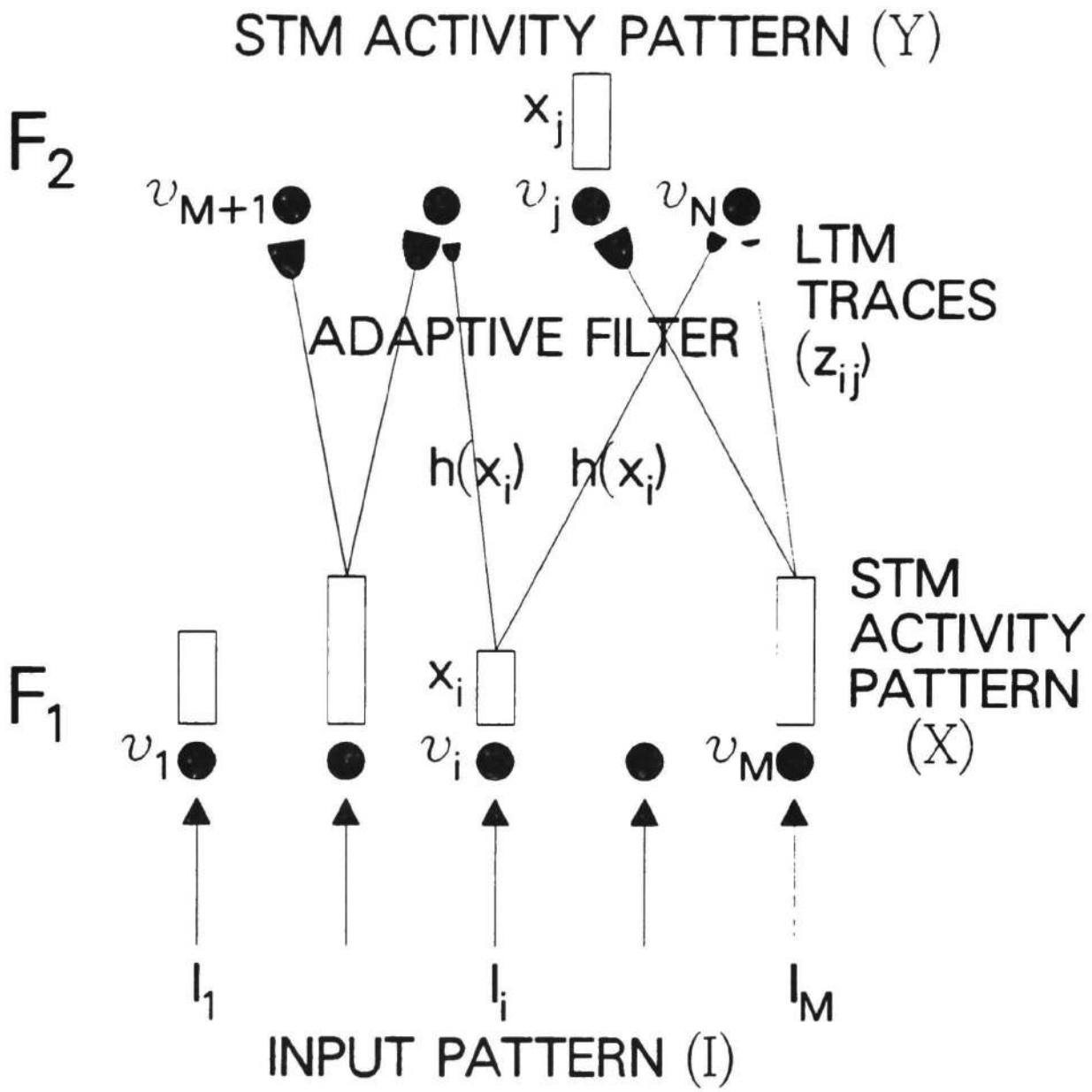
largely determines which type of learning will be superior."

In an ART system, if an erroneous recognition is followed by negative reinforcement, then the system becomes more *vigilant*. This change in vigilance may be interpreted as a change in the system's attentional state which increases its sensitivity to mismatches between bottom-up input patterns and active top-down critical feature patterns. A vigilance change alters the size of a single parameter in the network. The *interactions* within the network respond to this parameter change by learning recognition codes that make finer distinctions. In other words, if the network erroneously groups together some input patterns, then negative reinforcement can help the network to learn the desired distinction by making the system more vigilant. The system then behaves *as if* it has a better set of feature detectors.

The ability of a vigilance change to alter the course of pattern recognition illustrates a theme that is common to a variety of neural processes: a one-dimensional parameter change that modulates a simple nonspecific neural process can have complex specific effects upon high-dimensional neural information processing.

3. BOTTOM-UP ADAPTIVE FILTERING AND CONTRAST-ENHANCEMENT IN SHORT TERM MEMORY

The remainder of the article intuitively summarizes key model properties. We begin by considering the typical network reactions to a single input pattern I within a temporal stream of input patterns. Each input pattern may be the output pattern of a preprocessing stage. Different preprocessing is given, for example, to speech signals and to visual signals before the outcome of such modality-specific preprocessing ever reaches the attentional subsystem. The preprocessed input pattern I is received at the stage F_1 of an attentional subsystem. Pattern I is transformed into a pattern X of activation across the nodes, or abstract "feature detectors", of F_1 (Figure 3). The transformed pattern X represents a pattern in short term memory (STM). In F_1 each node whose activity is sufficiently large generates excitatory signals along pathways to target nodes at the next processing stage F_2. A pattern X of STM activities across F_1 hereby elicits a pattern S of output signals from F_1. When a signal from a node in F_1 is carried along a pathway to F_2, the signal is multiplied, or *gated*, by the pathway's long term memory (LTM) trace. The LTM gated signal (i.e., signal times LTM trace), not the signal alone, reaches the target node. Each target node sums up all of its LTM gated signals. In this way, pattern S generates a pattern T of LTM-gated and summed input signals to F_2 (Figure 4a). The transformation from S to T is called an *adaptive filter*.



3. Stages of bottom-up activation: The input pattern I generates a pattern of STM activation X across F_1. Sufficiently active F_1 nodes emit bottom-up signals to F_2. This signal pattern S is gated by long term memory (LTM) traces within the $F_1 \rightarrow F_2$ pathways. The LTM gated signals are summed before activating their target nodes in F_2. This LTM-gated and summed signal pattern T generates a pattern of activation Y across F_2. The nodes in F_1 are denoted by $v_1, v_2, \ldots, v_M$. The nodes in F_2 are denoted by $v_{M+1}, v_{M+2}, \ldots v_N$. The input to node v_i is denoted by I_i. The STM activity of node v_i is denoted by x_i. The LTM trace of the pathway from v_i to v_j is denoted by z_{ij}.

The input pattern T to F_2 is quickly transformed by interactions among the nodes of F_2. These interactions contrast-enhance the input pattern T. The resulting pattern of activation across F_2 is a new pattern Y. The contrast-enhanced pattern Y, rather than the input pattern T, is stored in STM by F_2.

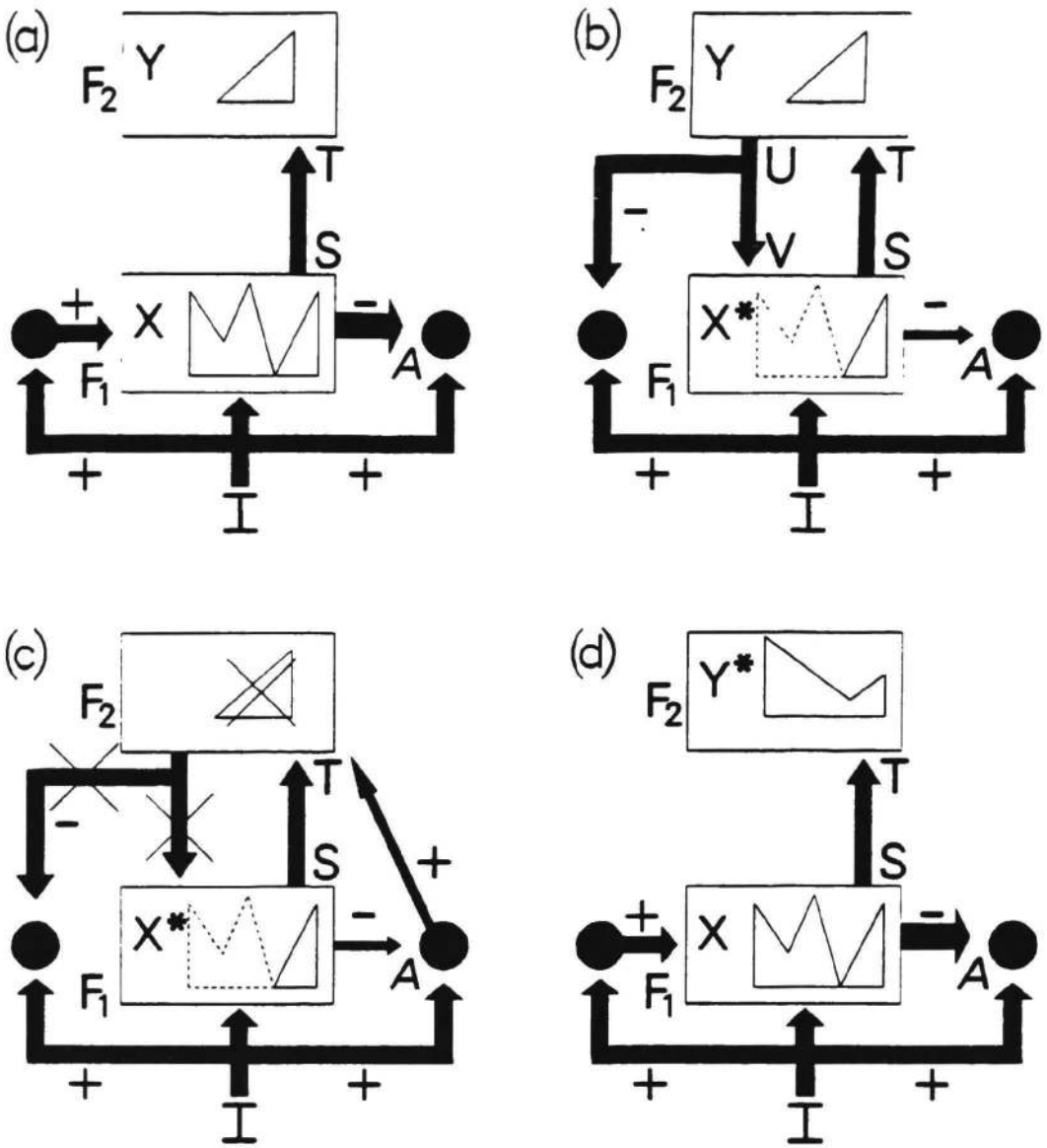
A special case of this contrast-enhancement process is one in which F_2 chooses the node which receives the largest input. The chosen node is the only one that can store activity in STM. In general, the contrast enhancing transformation from T to Y enables more than one node at a time to be active in STM. Such transformations are designed to simultaneously represent in STM several groupings, or chunks, of an input pattern (Cohen and Grossberg, 1986a, 1986b, 1986c; Grossberg, 1978a, 1986a). When F_2 is designed to make a choice in STM, it selects that global grouping of the input pattern which is preferred by the adaptive filter. This process automatically enables the network to partition all the input patterns which are received by F_1 into disjoint sets of recognition categories, each corresponding to a particular node (or "pointer," or "index") in F_2.

All the LTM traces in the adaptive filter, and thus all learned past experiences of the network, are used to determine the recognition code Y via the transformation $I \rightarrow X \rightarrow S \rightarrow T \rightarrow Y$. However, only those nodes of F_2 which maintain stored activity in the STM pattern Y can elicit new learning at contiguous LTM traces. Because the recognition code Y is a more contrast-enhanced pattern than T, many F_2 nodes which receive positive inputs ($I \rightarrow X \rightarrow S \rightarrow T$) may not store any STM activity ($T \rightarrow Y$). The LTM traces in pathways leading to these nodes thus influence the recognition event but are not altered by the recognition event. Some memories which influence the focus of attention are not themselves attended.

4. TOP-DOWN TEMPLATE MATCHING AND STABILIZATION OF CODE LEARNING

As soon as the bottom-up STM transformation $X \rightarrow Y$ takes place, the STM activities Y in F_2 elicit a top-down excitatory signal pattern U back to F_1 (Figure 4b). Only sufficiently large STM activities in Y elicit signals in U along the feedback pathways $F_2 \rightarrow F_1$. As in the bottom-up adaptive filter, the top-down signals U are also gated by LTM traces and the LTM-gated signals are summed at F_1 nodes. The pattern U of output signals from F_2 hereby generates a pattern V of LTM-gated and summed input signals to F_1. The transformation from U to V is thus also an adaptive filter. The pattern V is called a *top-down template*, or *learned expectation.*

Two sources of input now perturb F_1: the bottom-up input pattern I which gave rise to the original activity pattern X, and the top-down template pattern V that resulted from



4. Search for a correct F_2 code: (a) The input pattern I generates the specific STM activity pattern X at F_1 as it nonspecifically activates A. Pattern X both inhibits A and generates the output signal pattern S. Signal pattern S is transformed into the input pattern T, which activates the STM pattern Y across F_2. (b) Pattern Y generates the top-down signal pattern U which is transformed into the template pattern V. If V mismatches I at F_1, then a new STM activity pattern X^* is generated at F_1. The reduction in total STM activity which occurs when X is transformed into X^* causes a decrease in the total inhibition from F_1 to A. (c) Then the input-driven activation of A can release a nonspecific arousal wave to F_2, which resets the STM pattern Y at F_2. (d) After Y is inhibited, its top-down template is eliminated, and X can be reinstated at F_1. Now X once again generates input pattern T to F_2, but since Y remains inhibited T can activate a different STM pattern Y^* at F_2. If the top-down template due to Y^* also mismatches I at F_1, then the rapid search for an appropriate F_2 code continues.

activating X. The activity pattern X^* across F_1 that is induced by I and V taken together is typically different from the activity pattern X that was previously induced by I alone. In particular, F_1 acts to match V against I. The result of this matching process determines the future course of learning and recognition by the network.

The entire activation sequence

$$I \rightarrow X \rightarrow S \rightarrow T \rightarrow Y \rightarrow U \rightarrow V \rightarrow X^* \tag{1}$$

takes place very quickly relative to the rate with which the LTM traces in either the bottom-up adaptive filter $S \rightarrow T$ or the top-down adaptive filter $U \rightarrow V$ can change. Even though none of the LTM traces changes during such a short time, their prior learning strongly influences the STM patterns Y and X^* that evolve within the network by determining the transformations $S \rightarrow T$ and $U \rightarrow V$. We now discuss how a match or mismatch of I and V at F_1 regulates the course of learning in response to the pattern I, and in particular solves the stability-plasticity dilemma (Section 2).

5. INTERACTIONS BETWEEN ATTENTIONAL AND ORIENTING SUBSYSTEMS: STM RESET AND SEARCH

In Figure 4a, an input pattern I generates an STM activity pattern X across F_1. The input pattern I also excites the orienting subsystem A, but pattern X at F_1 inhibits A before it can generate an output signal. Activity pattern X also elicits an output pattern S which, via the bottom-up adaptive filter. instates an STM activity pattern Y across F_2. In Figure 4b, pattern Y reads a top-down template pattern V into F_1. Template V mismatches input I, thereby significantly inhibiting STM activity across F_1. The amount by which activity in X is attenuated to generate X^* depends upon how much of the input pattern I is encoded within the template pattern V.

When a mismatch attenuates STM activity across F_1, the total size of the inhibitory signal from F_1 to A is also attenuated. If the attenuation is sufficiently great, inhibition from F_1 to A can no longer prevent the arousal source A from firing. Figure 4c depicts how disinhibition of A releases an arousal burst to F_2 which equally, or nonspecifically, excites all the F_2 cells. The cell populations of F_2 react to such an arousal signal in a state-dependent fashion. In the special case that F_2 chooses a single population for STM storage, the arousal burst selectively inhibits, or resets, the active population in F_2. This inhibition is long-lasting. One physiological design for F_2 processing which has these properties is a *gated dipole field* (Grossberg, 1980, 1984a). A gated dipole field consists of opponent processing channels which are gated by habituating chemical transmitters. A

nonspecific arousal burst induces selective and enduring inhibition of active populations within a gated dipole field.

In Figure 4c, inhibition of Y leads to removal of the top-down template V, and thereby terminates the mismatch between I and V. Input pattern I can thus reinstate the original activity pattern X across F_1, which again generates the output pattern S from F_1 and the input pattern T to F_2. Due to the enduring inhibition at F_2, the input pattern T can no longer activate the original pattern Y at F_2. A new pattern Y^* is thus generated at F_2 by I (Figure 4d). Despite the fact that some F_2 nodes may remain inhibited by the STM reset property, the new pattern Y^* may encode large STM activities. This is because level F_2 is designed so that its total suprathreshold activity remains approximately constant, or normalized, despite the fact that some of its nodes may remain inhibited by the STM reset mechanism. This property is related to the limited capacity of STM. A physiological process capable of achieving the STM normalization property is based upon on-center off-surround feedback interactions among cells obeying membrane equations (Grossberg, 1980, 1983).

The new activity pattern Y^* reads-out a new top-down template pattern V^*. If a mismatch again occurs at F_1, the orienting subsystem is again engaged, thereby leading to another arousal-mediated reset of STM at F_2. In this way, a rapid series of STM matching and reset events may occur. Such an STM matching and reset series controls the system's search of LTM by sequentially engaging the novelty-sensitive orienting subsystem. Although STM is reset sequentially in time via this mismatch-mediated, self-terminating LTM search process, the mechanisms which control the LTM search are all parallel network interactions, rather than serial algorithms. Such a parallel search scheme continuously adjusts itself to the system's evolving LTM codes. In general, the spatial configuration of LTM codes depends upon both the system's initial configuration and its unique learning history, and hence cannot be predicted *a priori* by a pre-wired search algorithm. Instead, the mismatch-mediated engagement of the orienting subsystem realizes the type of self-adjusting search that was described in Section 2B.

The mismatch-mediated search of LTM ends when an STM pattern across F_2 reads-out a top-down template which matches I, to the degree of accuracy required by the level of attentional vigilance, or which has not yet undergone any prior learning. In the latter case, a new recognition category is then established as a bottom-up code and top-down template are learned.

6. ATTENTIONAL GAIN CONTROL AND ATTENTIONAL PRIMING

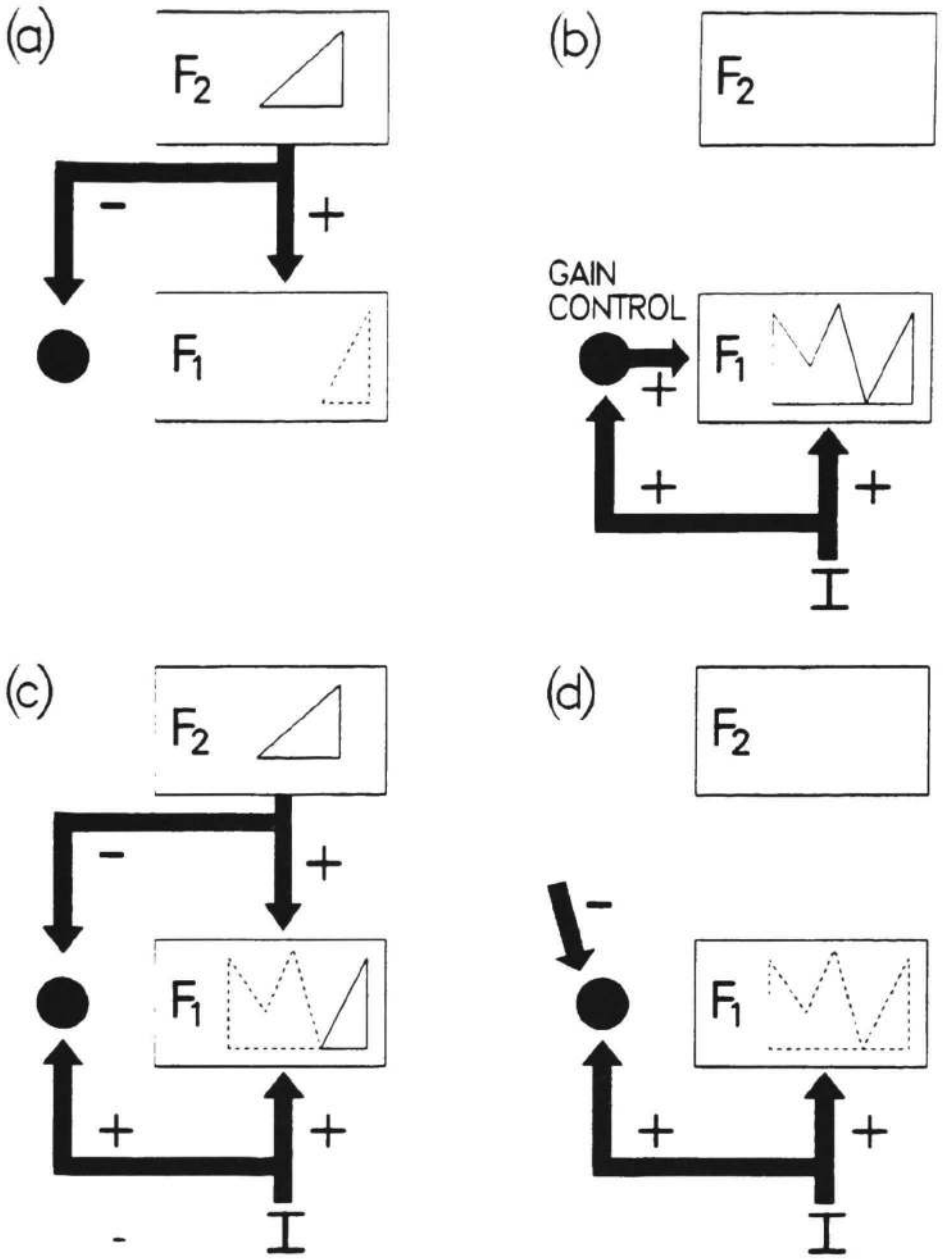
Further properties of the top-down template matching process can be derived by con-

sidering its role in the regulation of attentional priming. Consider, for example, a situation in which F_2 is activated by a level other than F_1 before F_1 can be activated by a bottom-up input (Figure 5a). In such a situation, F_2 can generate a top-down template V to F_1. The level F_1 is then primed, or sensitized, to receive a bottom-up input that may or may not match the active expectancy. As depicted in Figure 5a, level F_1 can be primed to receive a bottom-up input without necessarily eliciting suprathreshold output signals in response to the priming expectancy.

On the other hand, an input pattern I must be able to generate a suprathreshold activity pattern X even if no top-down expectancy is active across F_1 (Figures 4a and 5b). How does F_1 know that it should generate a suprathreshold reaction to a bottom-up input pattern but not to a top-down input pattern? In both cases, excitatory input signals stimulate F_1 cells. Some auxiliary mechanism must exist to distinguish between bottom-up and top-down inputs. This auxiliary mechanism is called *attentional gain control* to distinguish it from *attentional priming* by the top-down template itself (Figure 5a). While F_2 is active, the attentional priming mechanism delivers *excitatory specific learned* template patterns to F_1. The attentional gain control mechanism has an *inhibitory nonspecific unlearned* effect on the sensitivity with which F_1 responds to the template pattern, as well as to other patterns received by F_1. The attentional gain control process enables F_1 to tell the difference between bottom-up and top-down signals.

7. MATCHING: THE 2/3 RULE

A rule for pattern matching at F_1, called the 2/3 Rule, follows naturally from the distinction between attentional gain control and attentional priming. It says that two out of three signal sources must activate an F_1 node in order for that node to generate suprathreshold output signals. In Figure 5a, during top-down processing, or priming, the nodes of F_1 receive inputs from at most one of their three possible input sources. Hence no cells in F_1 are supraliminally activated by the top-down template. In Figure 5b, during bottom-up processing, a suprathreshold node in F_1 is one which receives a specific input from both the input pattern I and a nonspecific excitatory signal from the gain control channel. In Figure 5c, during the matching of simultaneous bottom-up and top-down patterns, the nonspecific gain control signal to F_1 is inhibited by the top-down channel. Nodes of F_1 which receive sufficiently large inputs from both the bottom-up and the top-down signal patterns generate suprathreshold activities. Nodes which receive a bottom-up input or a top-down input, but not both, cannot become suprathreshold: mismatched inputs cannot generate suprathreshold activities. Attentional gain control thus leads to a matching process whereby the addition of top-down excitatory inputs to F_1 can lead



5. Matching by the 2/3 Rule: (a) A top-down template from F_2 inhibits the attentional gain control source as it subliminally primes target F_1 cells. (b) Only F_1 cells that receive bottom-up inputs and gain control signals can become supraliminally active. (c) When a bottom-up input pattern and a top-down template are simultaneously active, only those F_1 cells that receive inputs from both sources can become supraliminally active. (d) Intermodality inhibition can shut off the F_1 gain control source and thereby prevent a bottom-up input from supraliminally activating F_1. Similarly, disinhibition of the F_1 gain control source by an "act of will" may enable a top-down prime to become supraliminal.

to an overall decrease in F_1's STM activity (Figures 4a and 4b). Figure 5d shows how competitive interactions across modalities can prevent F_1 from generating a supraliminal reaction to bottom-up signals when attention shifts from one modality to another.

8. CONCLUDING REMARKS: SELF-STABILIZATION AND UNITIZATION WITHIN ASSOCIATIVE NETWORKS

The qualitative properties discussed herein are elsewhere supplemented by a complete set of mathematical theorems and many computer simulations (Carpenter and Grossberg, 1986a, 1986b, 1986c). Two main conclusions of our work are especially salient. First, the code learning process is one of progressive refinement of distinctions. The distinctions that emerge are the resultant of all the input patterns which the network ever experiences, rather than of some preassigned features. Second, the matching process compares whole patterns, not just separate features. It may happen that two different input patterns to F_1 overlap a template at the same set of feature detectors, yet the network will reset the F_2 node in response to one input but not the other. The degree of mismatch of template pattern and input pattern *as a whole* determines whether coding or reset will occur.

Thus the learning of categorical invariants resolves two opposing tendencies. As categories grow larger, and hence code increasingly global invariants, the templates which define them become smaller, as they discover and base the code on sets of critical feature patterns, or prototypes, rather than upon familiar pattern exemplars. This work shows how these two opposing tendencies can be resolved within a self-organizing system, leading to dynamic equilibration, or self-stabilization, of recognition categories in response to an arbitrary list of arbitrarily many binary input patterns. This self-stabilization property is of major importance for the further development of associative networks and the analysis of cognitive recognition processes.

REFERENCES

Banquet, J.P., & Grossberg, S. (1986) Structure of event-related potentials during learning: An experimental and theoretical analysis. Submitted for publication.

Carpenter, G.A., & Grossberg, S. (1985a) Category learning and adaptive pattern recognition: A neural network model. *Proceedings of the Third Army Conference on Applied Mathematics and Computing*, ARO Report **86-1**, 37–56.

Carpenter, G.A., & Grossberg, S. (1985b) Neural dynamics of adaptive pattern recognition: Priming, search, attention, and category formation. **Society for Neuroscience Abstracts, 11**, 1110.

Carpenter, G.A., & Grossberg, S. (1986a) Neural dynamics of category learning and recognition: Attention, memory consolidation, and amnesia. In J. Davis, R. Newburgh, and E. Wegman (Eds.), **Brain structure, learning, and memory**. AAAS Symposium Series.

Carpenter, G.A., & Grossberg, S. (1986b) Neural dynamics of category learning and recognition: Structural invariants, reinforcement, and evoked potentials. In M.L. Commons, S.M. Kosslyn, and R.J. Herrnstein (Eds.), **Pattern recognition and concepts in animals, people, and machines**. Hillsdale, NJ: Erlbaum.

Carpenter, G.A., & Grossberg, S. (1986c) A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing.*

Cohen, M.A., & Grossberg, S. (1986a) Neural dynamics of speech and language coding: Developmental programs, perceptual grouping, and competition for short term memory. *Human Neurobiology*, **5**, 1–22.

Cohen, M.A., & Grossberg, S. (1986b) Unitized recognition codes for parts and wholes: The unique cue in configural discriminations. In M.L. Commons, S.M. Kosslyn, and R.J. Herrnstein (Eds.), **Pattern recognition and concepts in animals, people, and machines**. Hillsdale, NJ: Erlbaum

Cohen, M.A., & Grossberg, S. (1986c) Adaptive tuning of unitized perceptual groupings: Neural association, competition, and modulation.

Grossberg, S. (1976a) Adaptive pattern classification and universal recoding, I: Parallel development and coding of neural feature detectors. *Biological Cybernetics*, **23**, 121–134.

Grossberg, S. (1976b) Adaptive pattern classification and universal recoding, II: Feedback, expectation, olfaction, and illusions. *Biological Cybernetics*, **23**, 187–202.

Grossberg, S. (1978a) A theory of human memory: Self-organization and performance of sensory-motor codes, maps, and plans. In R. Rosen and F. Snell (Eds.), **Progress in theoretical biology**, Vol.5, pp.233–374. New York: Academic Press.

Grossberg, S. (1980) How does a brain build a cognitive code? *Psychological Review*, **87**, 1–51.

Grossberg, S. (1983) The quantized geometry of visual space: The coherent computation of depth, form, and lightness. *Behavioral Brain Sciences*, **6**, 625–692.

Grossberg, S. (1984a) Some psychophysiological and pharmacological correlates of a developmental, cognitive and motivational theory. In R. Karrer, J. Cohen, and P. Tueting (Eds.), **Brain and information: Event related potentials**, pp.58–151. New York: New York Academy of Sciences.

Grossberg, S. (1986a) The adaptive self-organization of serial order in behavior: Speech, language, and motor control. In E.C. Schwab and H.C. Nusbaum (Eds.), **Pattern recognition by humans and machines**, Vol.1. New York: Academic Press.

Grossberg, S., & Stone, G.O. (1986a) Neural dynamics of word recognition and recall: Attentional priming, learning, and resonance. *Psychological Review*, **93**, 46–74.

Grossberg, S., & Stone, G.O. (1986b) Neural dynamics of attention switching and temporal order information in short term memory, *Memory and Cognition.*

Posner, M.I. (1973) **Cognition: an introduction**. Glenview, IL: Scott, Foresman, and Co.

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

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

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

Representing Magnitude by Memory Resonance

A Hypothesis on Qualitative Judgment

Marten J. den Uyl

Psychological Laboraty
University of Amsterdam

ABSTRACT

Qualitative judgment, the ability to evaluate attributes that imply some degree of 'goodness' or preference, poses important problems for the information processing paradigm. In this paper one form of qualitative judgment, contextual judgment of magnitude, is analyzed in some detail.

The results of psychophysical experiments are consistent with the idea that human magnitude representation is based on a contextual coding process in which an actual stimulus is compared with a sample of traces of previously encountered similar stimuli. Such a coding process is hard to realize in a conventional memory system.

A distributed model for contextual magnitude judgment is described, in which this trace sampling process is feasible, when special provisions for the use of resonance information are made. Resonance coding involves the representation within a memory system of the memory activity caused by specific patterns of stimulation. A possible implementation of resonance coding, detection of dissonance, is briefly described. The hypothesis is put forward that evaluation of memory resonance plays an equally important role in other forms of qualitative judgment.

INTRODUCTION

People routinely make qualitative judgments: 'an interesting novel', 'a very elegant solution', 'a hot bath', 'a bitter disappointment', 'a highly complex problem'. Such judgments have in common that on the one hand some *classification* of a quality is indicated -interestingness, elegance, hotness, etc.- and on the other hand there is an *evaluation of the degree*, of the quality present.

Qualitative judgment poses pervasive problems to the information processing paradigm. One set of problems is apparent in the actual performance of presentday AI (expert) systems. It is generally recognized that AI systems show distinct patterns of strengths and weaknesses when compared to standards of human intelligence. Where AI systems may easily excel in computational power in formal domains, in 'verbatim' memory capacity, in deductive inference, it is a challenging task to approach human standards in heuristic inference, common sense, creativity, esthetic judgment, and robustness in the face of unexpected and deviant situations. Much of this pattern of weaknesses can be roughly summarized in the statement that AI systems have problems in dealing with *qualitative* aspects of tasks and situations.

Problems with qualitative judgment may be further illustrated in the relative neglect in cognitive psychology of affective processes such as motivation, valuation, emotion (e.g. Mandler 1985) in which qualitative judgment plays a central role.

Yet another set of problems prevail at the epistemological level. When people make qualitative judgments, they experience certain qualitative contents or *qualia*, a tomato 'looks bright red', we 'feel a pain' in a sore tooth. Qualia play a prominent role in debates on the philosophy of mind, and particularly in arguments on the adequacy of the functional (cognitive) paradigm in psychology. It has often been argued that the information processing paradigm cannot account for the phenomenal qualities of human experience.

The major aim of this paper is to outline and illustrate a global hypothesis on the nature of qualitative judgment. According to this hypothesis qualitative judgment is based on an evaluation of *memory resonance*, roughly the impact of stimulation on patterns of autonomous activity in the representational system. The body of the paper is devoted to an analysis of the *archetype* of qualitative judgment: the contextual judgment of magnitude.

QUALITATIVE JUDGMENT OF MAGNITUDE

Any system that operates in the real world must face the elementary task of representing the magnitudes or intensities of objects on continuous physical variables such as length, weight, sound- or light intensity. The question how human beings deal with this task is the proper domain of psychophysics, one of the oldest research traditions in modern experimental psychology. Research in psychophysics has demonstrated many peculiarities of human magnitude representation (For reviews see e.g. Carterette & Friedman 1974).

First of all, magnitude judgment often is a hard task for human beings. The point is perhaps appreciated best when it is realized how dependent we are on the many measurement instruments -from yardstick and balance to sophisticated electronic devices- that have been developed to overcome the limitations of sensory systems.

Problems with magnitude judgment arise in particular when memory representations of magnitudes are involved. Human capabilities in intensity resolution *per se* -i.e. discriminating between magnitudes when two or more are present simultaneously or in immediate succession- differ widely between sensory modalities, but in some cases human discrimination is fair to any standards and may be hard to equal by mechanical devices. One of the few general observations that can be made about intensity discrimination is known as 'Weber's Law': just notable differences (jnd's) between intensities tend to be constant fractions of stimulus intensity.

However, when a single stimulus at a time is judged in relation to a memory representation of one or more reference magnitudes, human performance is severely limited and, moreover, shows surprisingly little variation over sensory modalities. As Miller (1956) noticed, for many different sensory continua people cannot reliably distinguish more than 7 +/- 2 levels of intensity in identification tasks. Performance in intensity identification tasks, where reference levels must be remembered, contrasts sharply with performance in intensity discrimination tasks. For continua like loudness and brightness subjects may be able to partition the range from the weakest, just perceptible, to the largest, just bearable, intensity (the *dynamic range*) in well over a hundred jnd steps, when stimuli are pairwise compared; yet Ss can at most distinguish about 10 levels of intensity with single stimulus presentations.

It may be conjectured that these memory limitations are a direct consequence of the characteristics of the hardware in which human intelligence is implemented. Artificial recording devices store information in physically stable structures that may remain essentially unchanged over longer periods of time. The human brain is a soft kind of hardware, as living tissue grows, changes and decays over time. Human magnitude representation may be understood as a striving for constancy in an inherently unstable system.

A further characteristic of memory representations of magnitude is their dependency on context, particularly on the range of magnitudes over which judgments are made. The stability of memory representations of magnitude decreases with increasing judgmental range. Two intensity levels that can be easily distinguished when they form the extreme magnitudes in a small range stimulus set, may become highly confusable in a large range stimulus set.

Context dependency is also a striking feature of the way magnitudes are typically expressed in everyday language. Magnitudes are typically communicated in qualitative terms like 'big', 'small', 'many', 'few'. Such terms express magnitude in reference to a contextually determined norm; 'a lot of people' implies different numbers at cocktail parties or mass meetings.

How are contextual norms represented in memory? How are judgments made in relation to these norms? One approach, exemplified in the influential 'Adaptation Level' theory (Helson 1964), is to assume that judgments are made in relation to some neutral point that represents the central tendency of the stimulus distribution. Context effects then are represented in a single parameter, the adaptation level. However, it has repeatedly been demonstrated that not just the central tendency but the entire shape of the stimulus distribution is reflected in magnitude judgments (e.g. Parducci & Perrett 1971). Apparently people employ some 'multi parameter' representation of stimulus distributions.

Trace Sample Theory

The characteristics of human magnitude representation just described can be accounted for by a general model of qualitative magnitude judgment ('Trace Sample' theory, den Uyl 1981). Only a few elements of this theory need to be mentioned here.

Magnitude information may be represented in one of two different formats: primary or *trace code* and secondary or *contextual code* (cf. Durlach & Braida 1969). A trace code, the output of some perceptual system, is a direct or analogue representation of stimulus intensity. The precize nature of trace codes we leave unspecified for the moment; it is, however, assumed that trace codes are highly unstable and decay rapidly over time. Hence, the long term memory representation of individual trace codes will be subject to very large error.

The central assumption in trace sample theory is that contextual magnitude representations are formed by comparing the trace code for an actual stimulus magnitude to a sample of trace codes of previously encountered similar stimuli. The resulting contextual magnitude judgment is essentially the rank or percentile score of the actual stimulus magnitude in a subjective reference distribution formed by the trace sample. Thus, when we judge a dog to be 'very large', this essentially means that this dog appears to be larger than most dogs we have seen before. It should be noted that because of the large error in the memory

representation of traces, the subjective reference distribution will be systematically distorted with regard to the objective distribution of reference magnitudes.

Space does not permit to review the evidence here that this simple contextual coding mechanism can account for many findings in experimental psychophysics (cf. den Uyl 1981). We should turn attention to the implications of this coding scheme for the organization of memory.

Classification, Judgment and Memory

Trace Sample theory implies that quite a number of specific traces must be somehow represented in memory; in order to make qualitative judgments a sample of traces should be available for each dimension of each category in memory.

How is the representation of these traces integrated in the representation of categories in memory? One possibility is to assume that *fixed sets of traces* are stored with each category in memory, and can be accessed when the entry for the category is reached as a result of some classification process. The problem with this proposal is that it may easily lead to a proliferation of postulated trace samples when the aim is to account for the flexibility of human judgment. To just mention some potential problems:
-Often the norm for one dimension is dependent on the value on some other dimension, e.g. a heigth of three feet would be 'tall' for a two year old child, 'very tall' for a 1.5 year old, 'extremely tall' for a one year old. It may be possible to represent dependencies between correlated continuous variables in sets of discrete subcategories (e.g. Lebowitz 1985), but it would not seem a particularly elegant solution.
-Sometimes contextual judgments appear to be made in reference to *ad hoc categories* (Barsalou 1983) constructed on the spot. For example, the judgment that there are 'not so many people present' at a particular lecture, may be made in reference to an ad hoc norm for 'this kind of lecture', defined by a set of circumstances like size of the lecture hall, fame of the lecturer, time of day etc.

An alternative approach that could solve these problems in a principled way, would be to assume that *a trace sample is newly composed for each occasion* where a qualitative judgment is made. That is, for each judgment the memory system samples the traces from the previously judged objects *most similar* to the object presently being judged. Clearly, this is an attractive possibility in that it could conceivably give a system the flexibility apparent in human judgment[1]. However, the computational costs of this scheme would seem extravagant in a conventional computational architecture: the scheme implies that some similarity metric is computed between the judged object and each individual object stored in memory. In order to further explore this proposal, we need to consider an implementation of the trace sample model in a parallel-distributed processing architecture in which the implied computations are feasible.

[1] An elegant treatment of 'norm theory' based on related principles has recently been presented in (Kahneman & Miller 1986).

MAGNITUDE REPRESENTATION IN A DISTRIBUTED MEMORY

The distributed model I will outline here is based on the 'Harmonium' model developed by Smolensky & Riley (1984) with a few modifications inspired by related models (e.g. McClelland & Rumelhart 1985). Harmony theory relies on a formal mapping between parallel computation and thermal physics and is similar in this respect to the 'Boltzman machine' described by Ackley, Hinton & Sejnowski (1985).

A distributed memory consists of a -large- set of interconnected modules, each of which in turn consists of many interconnected simple processing units or 'nodes'. The activation state of nodes may vary, and the basic mode of operation of nodes is the passing of activation signals to other nodes within and between modules. Specific patterns of activity over the nodes in the memory network constitute active knowledge states. Information processing takes the form of chains of knowledge states, brought about by the units spreading activation through the network.

Each module contains two layers of nodes: *representation nodes* (R-nodes) define the active knowledge states of the system; *trace nodes* (T-nodes) contain information on past contingencies between R-nodes and may send activation signals to R-nodes on this basis. Connections are only between layers, nodes within a single layer are not connected. Both representation and trace nodes take only two activation values: nodes are either on/active or off/inactive.

The basic cognitive operation in a connectionist memory is pattern completion. Pattern completion takes place in (asyncronous) processing cycles in individual modules as follows: At the beginning of the cycle some subset of the R-nodes in the module is clamped into activation states by incoming connections from other modules. Other R-nodes are assumed to have random activation values at the beginning of the cycle. The task for the module is to reinstate a stable and complete pattern over all the representation nodes in the module. Note that in a distributed model 'psychological' stimulus features (e.g. shapes, color) do not correspond to individual processing units, but to activation patterns over collections of units.

Each trace node is connected to a set of representation nodes by bidirectional links. A T-node contains -as a result of past experience- a *key*, a set of weights on connections with R-nodes with values of either +1 or -1. This key defines a 'preferred pattern' of activation states over connected R-nodes.

A processing cycle is divided in discrete 'ticks' (McClelland & Rumelhart 1985). At each tick a T-node receives an activation signal from each R-node that is consistent with the key (the state of the R-node matches the sign of the weight) and a de-activation signal for each mismatch. A trace node is in the active state as long as the sum of the (de-)activation signals exceeds some variable threshold value. Active T-nodes send activation signals to R-nodes in accordance with the weights in the key, e.g. a de-activation signal is send when the weight is negative.

The thresholds of T-nodes are gradually raised in the course of a processing cycle. As a result many T-nodes may participate in the first stages of a cycle. At the end of a cycle only the best matching T-nodes remain active when the module 'freezes' into a stable completion.

The activation state of a R-node not clamped by external connections is a (probabilistic) function of the incoming signals from T-nodes; representation nodes will tend to conform to the 'majority vote' of incoming signals.

How do trace nodes come to represent contingencies between representation nodes? I will shortcut this complex issue here by postulating a global learning rule that suffices for present purposes: at the end of a completion cycle, key weigths of then active trace nodes that are inconsistent with the activation state of the connected representation node may change their sign with a certain probability so as to achieve consistency[2].

Harmony

In a 'harmonium model' processing is driven by a single principle: that of achieving completions with the highest harmony or 'self-consistency'. Harmony is the degree of consistency between a pattern of activation over the representation nodes, and a set of preferred patterns defined by the keys in the active trace nodes. More precisely, the harmony of a system state can be expressed as follows[3]:
Let a vector T represent the states of trace nodes with the values active=1 and inactive=0; a vector R over representation nodes takes the values on=+1 and off=-1; a vector K_i represents the key in trace node T_i with the values +1 and -1 for positive and negative weights, and 0 when T_i is not connected to R_j. The harmony of a state then is:

$$H_{(T,R)} = \sum_i T.R^*K_i \qquad (1)$$

It can readily be seen that (1) is equivalent to the sum of all consistent signals in the module minus the summed inconsistent signals ('*' denotes the dot product).

Primary Coding of Magnitude

Magnitudes on some dimension are represented on a subset of the R-nodes in a harmonium module, the 'magnitude representation nodes' or M-nodes. The primary code for a magnitude is simply the number of M-nodes active in the module. This primary coding may be brought about thus.
M-nodes are clamped into activation states by external connections carrying activation signals that have their origin in sensory systems. For each M-node a threshold parameter m_{it} is defined that randomly fluctuates over time. An M-node will be clamped in the active state upon presentation of a stimulus magnitude s_j when $s_j > m_{it}$ and will be clamped 'off' otherwise.
Weber's Law implies that the tresholds of M-nodes are spaced approximately geometrically over the dynamic range of the magnitude dimension. The cumulative distribution function, i.e. the expected number of active M-nodes, then is a logarithmic function of stimulus intensity.

[2] This postulate marks a transition from an 'enumeration of specific instances' principle in Smolensky & Riley (1984) to a 'superposition of traces' principle (McClelland & Rumelhart 1985).
[3] The present notation is slightly different from Smolensky & Riley (1984).

A rationale for the present instantiation of Fechner's time-honored 'Logarithmic Law' is that this coding scheme makes minimal demands on the stability of the units in which magnitude information is encoded. In this way reliable intensity resolution can be achieved by parallelling highly unreliable individual units.

Contextual Coding

In Trace Sample theory (den Uyl 1981) the contextual code for a stimulus magnitude s_j is the proportion of traces in the trace sample smaller than s_j. The present distributed model does not store traces (copies) of magnitudes. Stimuli may only have a lasting impact on memory by changing key weights. Yet, a contextual code can be computed in the harmonium model that is equivalent to the secondary code in Trace Sample theory.

Four kinds of *trace signals* t(R,K) from T-nodes to R-nodes contribute to system harmony:
- activation signals to active R-nodes: t(+,+),
- de-activation signals to active R-nodes: t(+,-),
- activation signals to inactive R-nodes: t(-,+),
- de-activation signals to inactive R-nodes: t(-,-).

Inconsistent trace signals (t(+,-) and t(-,+)) provide information concerning the position of s_j in the magnitude distribution of similar stimuli. A signal t(+,-) implies that the present stimulus dominates a threshold ($s_j > m_{it}$) while some past stimulus s_p which caused the negative key weight in the signal, had been below this threshold ($m_{it} > s_p$). Hence, it may be inferred from t(+,-) that s_j dominates at least one past stimulus. Analogously, a signal t(-,+) indicates that s_j is smaller than some past stimulus. Of course, error is introduced into these inferences because of the random fluctuations over time of the thresholds.

Consistent signals do not provide contextual information, these signals only indicate where both s_j and past stimuli stand with regard to threshold nodes.

We denote the frequency of the signal t(R,K) summed over M-nodes, given activation vectors T and M upon presentation of stimulus s_j by $f_{(T,M)j}(R,K)$. A contextual code C_j equivalent to the percentile code in Trace Sample theory can then be expressed as:

$$C_j = \frac{f_{(T,M)j}(+,-)}{f_{(T,M)j}(+,-) + f_{(T,M)j}(-,+)} \qquad (2)$$

C_j may range from 0 for extremely small stimuli to 1 for extremely large stimuli, the expected value for intermediate stimuli (the adaptation level) is 0.5.

It may be observed that in the present model norms for magnitude judgment, i.e. the distribution of trace signals in a harmonium module, are indeed composed anew for each judgment, as each stimulus pattern may activate a different T vector and hence sample a different set of key weights (cf. McClelland & Rumelhart 1985).

Dissonance as a Contextual Code

The contextual coding scheme in the preceding section raises one important problem: How can the contextual code in (2) be computed in a distributed system? All the information required for the computation is present in the M-nodes. It would appear a simple solution to have all M-nodes send counts of inconsistent incoming signals to special 'collector-units' where the necessary computations could be performed. However, this proposal would require M-nodes to spread other -more complex- information through the system than just their activation state. The proposal implies a major breach with connectionist design principles.

There are ways to approximate the contextual code using standard -though specialized- connectionist processors. One such scheme involves the detection of *dissonance* in activation patterns. A clamped representation node is dissonant, when a majority of incoming trace signals are inconsistent with the clamped activation state of the R-node.

Dissonance can be detected in the following way. Suppose a completion cycle in a harmonium module is followed by a *resonance cycle* in which the activation pattern on T-nodes remains unchanged, but all clamps on R-nodes are removed, thus allowing R-nodes to settle into preferred states. Dissonant nodes -i.e. R-nodes which change their activation state in the resonance cycle- could then be registrated in specialized units (e.g. units sensitive to time-contrast in activation signals).

The contextual code in (2) can be approximated in various ways from the dissonance in M. Large stimuli will tend to clamp high-threshold M-nodes -expected to be off- in dissonant active states, small stimuli will produce dissonance in inactive nodes. Intermediate magnitudes wil cause the lowest overall dissonance in M.

The precize form a dissonance-based contextual code may take is not important for the moment. The general point I want to make here is that although it is feasible to develop fairly simple coding schemes for contextual magnitude representations in a connectionist memory module, some special provisions are required to this end.

RESONANCE AND QUALITATIVE JUDGMENT

It may be conjectured that all qualitative judgments share the structure of contextual magnitude judgment, they all involve the contextual evaluation of *memory resonance* to an object description.

A crucial step must be taken in order to use resonance information in a memory system. Global characteristics of memory activity in response to a pattern of stimulation must be represented within the memory system itself, in order to be interpreted as providing information about the object that gave rise to this memory activity. For instance, in order to use the extent of inconsistent activation in a module as a measure of the extreemness of the magnitude of an object, inconsistent activation must in some way be detected and registrated.

It has been proposed that resonance evaluation takes the form of evaluating patterns of dissonance between a 'preferred' or expected resonance pattern, and the pattern externally imposed on the memory. A rationale for this form of evaluation can be found in the design principles of a distributed memory. In a distributed memory knowledge is hidden rather than

stored. Past events only leave traces in the form of changes in activation weights. It is not easy to systematically search or recover past episodes from such traces.

Some tasks, like magnitude judgment or evaluation of prototypicality, require a system to make use of knowledge concerning the *distribution of different stimulus patterns* in the past. Yet, a module in a distributed memory can only support one pattern of activation at a time. Traces of past events that do not correspond to an actual pattern in a module may only influence further processing, if it is in some way registered that they would have liked to see things different.

I have said little about 'goodness', preference and affective qualities in qualitative judgment. The reason is that liking does not occur -except in a metaphorical sense- on the level of memory modules. Only the system as a whole has likes and dislikes. In order to extend the present hypothesis to affective evaluation, the notion of memory resonance should be extended to higher levels of organization, i.e. resonance should be evaluated over collections of modules rather than within single modules. A suitable framework for such an extension can be found in Frijda's 'concern-realization' theory of emotion (Frijda 1986; den Uyl & Frijda 1984).

REFERENCES

Ackley, D., Hinton, G. E. & Sejnowski, T. J. (1985). Boltzmann machines: Constraint satisfaction networks that learn. *Cognitive Science* 9, 147-169.

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

Carterette, E. C. & Friedman, M. P. (1974). *Handbook of Perception Volume II*, Ac. Press, New York.

Durlach, N. I. & Braida, L. D. (1969). Intensity Perception I: Preliminary theory of intensity resolution. *Journal of the Acoustical Society of America*, 46, 372-383.

Frijda, N.H. (1986) *The Emotions*. New York: Cambridge U.P.

Helson, H. (1964). *Adaptation Level Theory*, New York, Harper and Rowe.

Kahneman, D. & Miller, D.T. (1986). Norm Theory: comparing reality to its alternatives. *Psychological Review*, Vol. 93, 136-153.

Lebowitz, M. (1985). Categorizing numeric information for generalization. *Cognitive Science*, 9, 285-308

Mandler, G. (1985). *Cognitive Psychology: an essay in cognitive science*. LEA, Hillsdale, N.J.

McClelland, J. L. & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General*, Vol. 114, 159-188.

Miller, G. A. (1956). The magical number 7 plus or minus two. *Psychological Review*, 63, 81-97.

Parducci, A. & Perrett, L.F. (1971) Category rating scales: effects of relative spacing and frequency of stimulus values. *Journal of Experimental Psychology, Monograph*, 89, 427-452.

Smolensky, P. & Riley, M. S. (1984). Harmony Theory: Problem solving, parallel cognitive models, and thermal physics. ICS Report 8404, UC San Diego.

den Uyl, M.J. (1981). On the Relativity of Human Judgment: Outline of a Theory of Contextual Magnitude Representation. Manuscript, Psych. Dept. Stanford University.

den Uyl M.J. & Frijda, N.H. (1984). Mood, Emotion and Action: A Concern-Realization Model. Proceedings of the Sixth Cognitive Science Conference, Boulder, Co. 137-141.

THE SPACING EFFECT ON NETTALK, A MASSIVELY-PARALLEL NETWORK

Charles R. Rosenberg
Cognitive Science Laboratory
Princeton University

Terrence J. Sejnowski
Department of Biophysics
Johns Hopkins University

ABSTRACT

NETtalk is a massively-parallel network that learns to convert English text to phonemes. In NETtalk, the memory representations are shared among many processing units, and these representations are learned by practice. In humans, distributed practice is more effective for long-term retention than massed practice, and we wondered whether learning in NETtalk had similar properties. NETtalk was tested on cued paired-associate recall using nonwords as stimuli. Retention of these target items was measured as a function of spacing, or the number of interspersed items between successive repetitions of the target. A significant advantage for spaced or distributed items was found for spacings of up to forty intervening items when tested at a retention interval of 64 items. Conversely, a significant advantage for massed items was found if testing immediately followed study. These results are strikingly similar to the results of many experiments using human subjects and suggest an explanation based on distributed representations in massively-parallel network architectures.

INTRODUCTION

CRR was supported in part by a research grant (487906) from IBM, by the Defense Advanced Research Projects Agency of the Department of Defense and by the Office of Naval Research under Contracts Nos. N00014-85-C-0456 and N00014-85-K-0465, and by the National Science Foundation under Cooperative Agreement No. DCR-8420948 and under NSF grant number IST8503968. TJS was supported by grants from the National Science Foundation, System Development Foundation, Sloan Foundation, General Electric Corporation, Exxon Education Foundation, Allied Corporation Foundation, Westinghouse, and Smith, Kline & French Laboratories.

We are indebted to Dr. T. Landauer for calling this problem to our attention. We also wish to thank George Miller and Stephen Jose Hanson for many helpful comments, Katherine Miller for expert editorial assistance, and Bell Communications Research for generously providing computational support.

In 1885, Ebbinghaus noted that "with any considerable number of repetitions a suitable distribution of them over a space of time is decidedly more advantageous than the massing of them at a single time" (Ebbinghaus, 1885/1964 p.89). Since then, the spacing effect has been found across a wide range of stimulus materials and tasks, semantic as well as perceptual/motor, and has even been found when the repetitions are across modality, or across languages, if bilinguals are employed as subjects (see Hintzman, 1974, for a review). The ubiquity of these results suggests that spacing reflects something of central importance in memory. However, despite over a hundred years of research, the spacing effect, as general as it is, continues to defy adequate, or at least, simple, explanation.

Perhaps the most popular account of the spacing effect is the encoding variability hypothesis (e.g. Melton, 1970; Martin, 1968; Glenberg, 1979). This hypothesis makes two major assumptions: (1) stimuli are encoded relative to the context, or environment, in which they occur, and (2) the probability of retrieval is positively correlated with the similarity of the context at retrieval to the context at encoding.[1] Given a continuously evolving environment, two trials that occur back-to-back will tend to share more context with each other than two trials widely separated in time.[2] Distributing practice will hence be advantageous to the extent that the two repeats of the to-be-remembered (TBR) item are encoded more independently, thus boosting the probability of the item's retrieval in a randomly chosen context presumably by increasing the number of possible retrieval routes.

Overall, the encoding variability hypothesis has found only limited empirical support (Hintzman, 1976). One recent failure of the hypothesis is the study by Postman and Knecht (1983) in which the encoding contexts of words were varied by embedding them in different sentences, either one sentence repeated three times or in three sentences each only once repeated. In the cued recall task, cueing was with one or three of the sentence frames (with the TBR word deleted). According to the encoding variability hypothesis, retrieval should be greater in the multiple context condition. Nevertheless, they found no difference in free recall of the target words, tested either immediately or after 24 hours. In fact, cued recall with a single sentence frame led to higher recall rates for targets that appeared in single contexts than targets that appeared in multiple contexts, a trend in a direction opposite to that predicted by the hypothesis.

---

[1] The notions underlying the encoding variability hypothesis were originally derived from Estes's stimulus sampling and fluctuation model (1959).

[2] The precise use of the term "context" has not always been consistent among investigators. See Maki & Hasher (1975) for a discussion and empirical investigation of this issue.

Postman and Knecht concluded that encoding items in different contexts does not necessarily improve retention, and may, in fact, lead to diminished retention. An overriding factor may be the strength of specific cue-target associations, built up by repeating the item in identical contexts. That is, many weak retrieval routes are not necessarily better than one strong one.

If we assume, as the Postman and Knecht study suggests, that the spacing effect depends to some extent upon the repetition of specific items, and not necessarily on the encoding of items relative to a continuously varying context, then the following question arises: Which repetition of the item, the first or the second, is less effectively processed or encoded when the two presentations occur back-to-back? Those theories that claim that the first presentation is deficient include the rehearsal-buffer theory (Atkinson & Shiffron, 1968; Rundus, 1971) and a version of consolidation theory (Landauer, 1969). In either case, the disadvantage found for massed practice is the result of the interruption of an ongoing encoding process by the immediate occurrence of the second item. Bjork and Allen (1970) found, however, that interposing a more difficult task between repetitions did not disrupt this encoding process, as both rehearsal-buffer theory and the consolidation hypothesis would predict. To the contrary, they found *improved* retention in the difficult task condition.[3] This result is hard to reconcile with either theory.

The other alternative is that the *second* massed presentation is more poorly processed or somehow less effective. It has been proposed that subjects habituate (e.g. Hintzman, Block, & Summers, 1973) and therefore cannot process the second massed presentation as effectively as they can the first. It is not clear how this proposal could explain the Bjork and Allen result, unless the intervening difficult task in some way releases the habituation from the first item. In addition, attempts to overhabituate to a target item by presenting the item for longer durations have been unsuccessful (Hintzman, Summers, & Block, 1975).

Another suggestion is that subjects do not attend as effectively (e.g. Shaughnessy, Zimmerman, & Underwood, 1972) to the second occurrence of an item if it closely follows an identical first item in time. But efforts to force subjects to attend to the second occurrence in various ways have indicated no reduction in the spacing effect (e.g. Hintzman, Summers, Eki, & Moore, 1975).

Jacoby (1978) has offered an account of spacing in terms of processing effort. That is, in the massed presentation condition, subjects have the first item consciously available when

[3] This result has been replicated by Tzeng (1973).

they are presented with the second item and consequently do not have to process the second item to the same degree. As a result, processing on the second massed item is not as great as that on the first and does not form as rich a code. This explanation seems to give a coherent account of all the evidence thus far mentioned: It does not depend on encoding items relative to a dynamic context, and it accounts for the Bjork and Allen results, since interposing a difficult task could plausibly make the second item less available, requiring more processing. However, it leaves unclear what "processing effort" (not to mention "consciousness") involves.

All these theories attempt to explain spacing in terms of concepts such as encoding, habituation, and consolidation, which make little reference to the actual form of the memory representation, although implicit in some of the explanations is the assumption that individual items have local representations. Another approach is to seek an explanation at the level of the representation: *It may matter how the information is stored in the system.* One way to explore this possibility is to construct explicit models that incorporate particular memory representations and learning mechanisms and to test them with the same experimental paradigms that have been used to study human memory.

It will be demonstrated that the spacing effect is a natural consequence of learning in a network with distributed memory representations and an incremental learning procedure. In this framework, learning consists of modifying connections in the network so that this information is retained as accurately as possible within the constraints imposed by the number of available connections. This way of storing information is fundamentally different from a local representation where individual items can be stored independently of one another, as in a computer memory. In a distributed representation a single connection can participate in the storage of many items, and conversely a single item is stored in many connections. Approaching memory in this way has already led to new insights in the domains of categorization and concept formation (McClelland & Rumelhart, 1985; Anderson, Silverstein, Ritz, & Jones, 1977; Eich, 1982; Amari, 1977; Kohonen, Oja, & Lehtio, 1981). These models of memory are inspired by the parallel architecture of the brain (Ballard, Hinton, & Sejnowski, 1983; Feldman & Ballard, 1982).

NETtalk

We have recently described NETtalk (Sejnowski & Rosenberg, 1986), a massively-parallel network that learns to translate letters in English text into phonemes and associated word stress. It achieves approximately 95% accuracy per letter without access to information about semantics or syntax. In NETtalk, the learning occurs by modifying the strengths of connections between a large number of simple and identical processors, or units. These

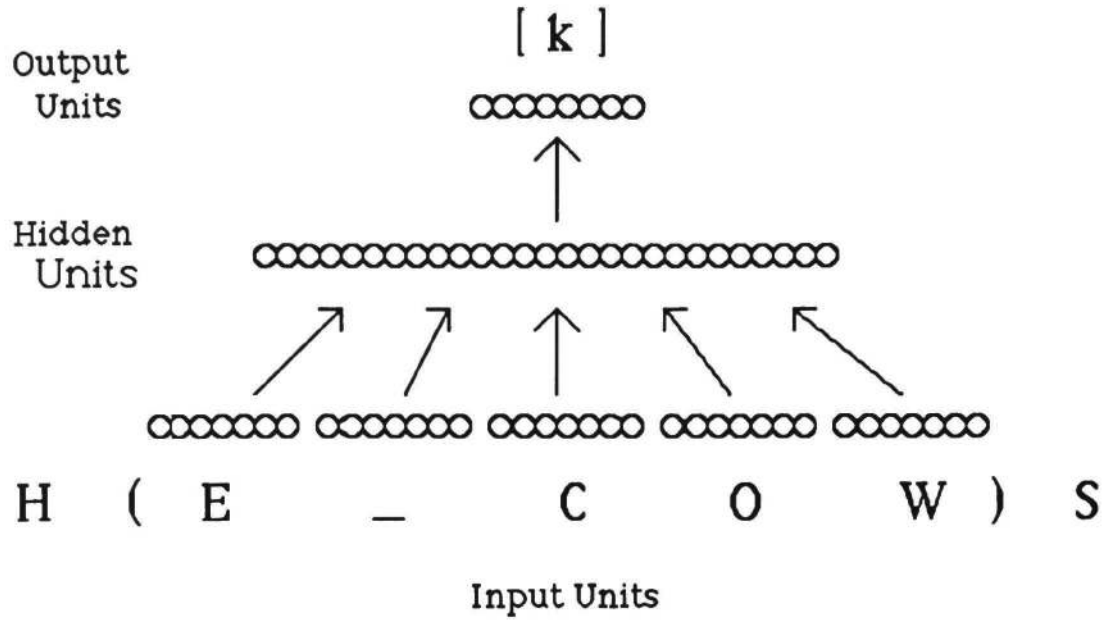


Figure 1. A schematic drawing of the NETtalk architecture. The little circles represent units (there are many more units than shown here) and the arrows represent bundles of connections, or weights, between the groups of units. Connectivity is complete between the connected groups, so each unit in each of the five input groups shown has connections to all of the units in the hidden layer, all of which are in turn connected to each of the output units. For the present experiments, there were 29 units in each of the input groups, 60 units in the hidden layer, and 26 units in the output layer. In addition, all units have a connection to a special unit that is always "on" (not shown), which serves as a variable threshold.

connections, which are real-valued and directional, determine how the activity of one unit affects the activity of another unit. If unit A is in an active state and there is a positively-valued (excitatory) connection going from unit A to unit B, then the activity level of unit B will be driven towards one. Conversely, a negatively-valued (inhibitory) connection between the two units will drive the activity level in unit B towards zero. A given unit typically has connections to a large number of other such units. The behavior of the network is the collective result of a large number of these simple, local, computations that are performed in parallel.

NETtalk has access to the correct pronunciations of the words during the learning, so it is "supervised" and akin to learning with a teacher. The back-propagation of error was used to adjust the values of the connection strengths (Rumelhart, Hinton, & Williams, 1986), which is a generalization of the perceptron learning rule (Rosenblatt, 1962) to multi-layered networks.

There are 231 units and 10,346 connections in the version of NETtalk used in the present experiments. As shown in Figure 1, the units that compose NETtalk are arranged in a layered hierarchy, consisting of three layers: an input layer, which encodes letters, an output layer, which encodes phonemes and stress, and a hidden layer that connects the input layer to the output layer. Each of the layers is completely connected to the layer just above and/or just below it. Letters are "clamped" at the input layer, and information (in the form of unit activity levels) passes up through the hidden layer, finally reaching the output layer where the pattern of activity on the output units is interpreted as a phoneme and stress.

The decision of how each letter is to be pronounced must be made on the basis of the surrounding letter context, since all letters can be pronounced in several ways. Using NETtalk, we have been able to examine how performance varies with window size. In this version, the network "sees" five letters at a time: the current letter, the preceding two letters, and the following two letters.[4] Each of these five letters is encoded simultaneously in a set or group of twenty-six dedicated units, locally representing each of the twenty-six English letters. Imposed on the network is a control structure that steps this five-letter window through the corpus, letter-by-letter.

More specifically, the value of each unit is a function of the values of all the units in the layer below it and the strength of the connection between the two units. The value of the ith unit is determined by first summing all of its inputs

$$E_i = \sum_j w_{ij} p_j \tag{1}$$

where p_j is the value of the jth unit and w_{ij} is the weight value of the connection between the two units, and then applying a sigmoidal transformation

$$p_i = P(E_i) = \frac{1}{1 + e^{-E_i}} \tag{2}$$

The resulting pattern of activity produced at the output layer is interpreted as the "guess" of how the middle letter in the window should be pronounced.[5] This output vector is then compared with the "correct" phoneme provided it, and the connection strengths in the network

[4] This window size has been reduced from seven in the original NETtalk in order to speed training.

[5] This was done by computing the projection of the output vector on all the possible phonemes (there are 55 of them) and selecting the phoneme with the highest overlap.

are recursively adjusted to minimize their differences (see Rumelhart, Hinton, & Williams, 1986, for details).

There were two adjustable learning parameters in the model: The rate of learning, ε, was set to 4.0, and the smoothing parameter, α, was set to zero (see Sejnowski & Rosenberg, 1986, for an explanation of these parameters). Continuous decay of the weight values towards zero has been experimented with, but was not used in the present experiments.

The purpose of the present experiment was to investigate the spacing effect in NETtalk, a network with two layers of modifiable weights. The design was modeled after Experiment 1 by Glenberg (1976). In this experiment, subjects were presented with paired associates, repeated twice at spacings of approximately 0, 1, 4, 8, 20, and 40 intervening items, and tested at retention intervals of approximately 2, 8, 32, and 64 items. Each pair was composed of two four-letter common nouns, "constructed to avoid common pre-experimental associations, rhymes, and orthographic similarities" (pg. 4). At test, just the stimulus word was presented, and the subject was to recall the associated response term. Glenberg's results are reproduced here as Figure 2. A significant interaction was found between spacing (lag) and retention interval. At short retention intervals, massed repetitions led to a higher probability of recall, whereas at long retention intervals, distributed repetitions were advantageous. Glenberg also noted that retention at the 64-item retention interval was a monotonic and negatively accelerating function of spacing.

As in Glenberg's experiment, the retention of target stimuli repeated a certain number of times at various spacing intervals was measured as a function of retention interval. If NETtalk exhibits the spacing effect, then long-term retention of these items should be better when a large number of other items intervene between successive repeats of the target (distributed practice). Conversely, short-term retention of the target items should be better when fewer items are presented between repeats (massed practice).

METHOD

Pre-Experimental Training

The network was first trained to pronounce a set of commonly occurring English words. These words were obtained by selecting the one thousand most frequent words from the *Webster's Pocket Dictionary,* based on frequency counts in the Brown corpus (Kucera & Francis, 1967). The network cycled through this one thousand word corpus a total of eleven times. The performance of the network at this point in training, as determined by the percentage of the correct phonemes "guessed", was 85%, and could have been improved with

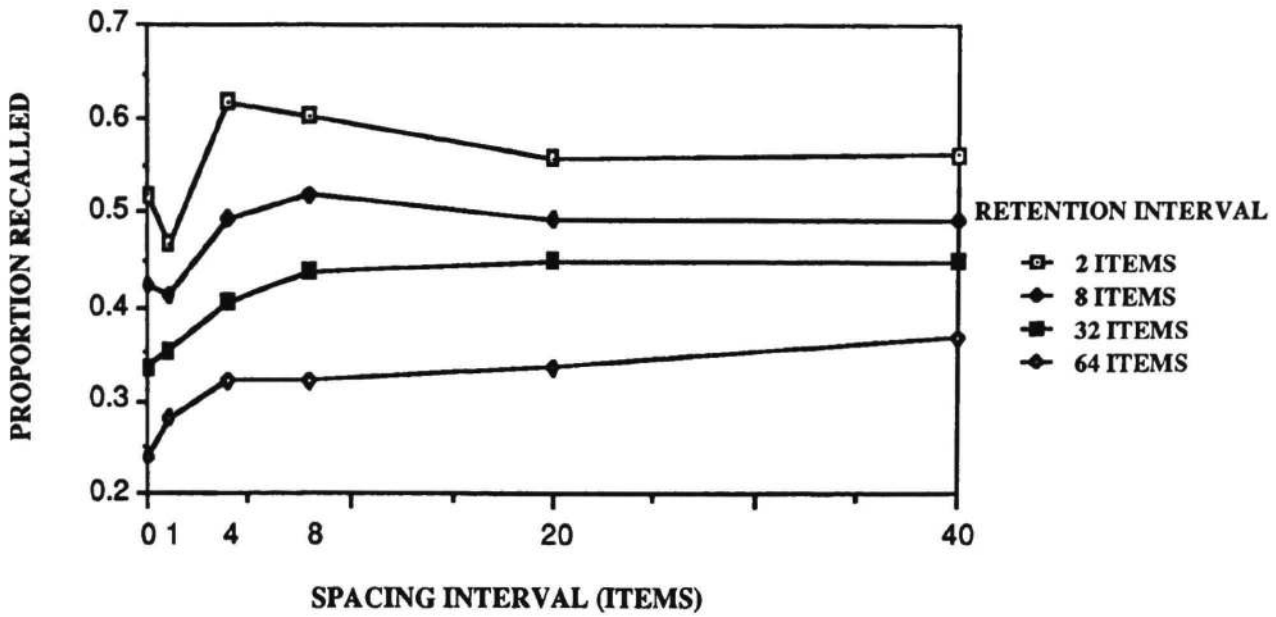


Figure 2. The proportion of response terms recalled as a function of spacing interval and retention interval. (After Glenberg, 1976.)

further practice.

The weight values of the network were stored following this initial training, and served as a common starting point for all of the subsequent experimental trials.

Target Stimuli

In order to force new learning to take place, random character strings of length six were employed as target stimuli. Thus there was no orderly relation between the cue and response. Whatever performance level NETtalk was able to reach on these items could not have been due to the utilization of rules acquired either prior or subsequent to study.

Twenty six-letter cues were generated by choosing six letters at random (with replacement) out of the twenty-six letters of the English alphabet. Likewise, the response terms associated with each of these cues were randomly generated phoneme and stress strings, also six characters each in length. There were 53 possible phonemes and five possible stress characters.[6] The frequency of occurrence of the characters in natural language were not taken into account in this selection process. Some of these items and several items from the

[6] In generating the target stimuli, two "phonemes", the space between words (_) and the period (.), were not possible choices.

Table 1. Examples of some training (distractor) and target items used.

DISTRACTORS

letters	phonemes	stress
file	fAl-	>1<<
all	cl-	1<<
second	sEkxnd	>1<0<<
take	tek-	>1<<
together	txgED--R	>0>1<<0<
neck	nEk-	>1<<
atmosphere	@tmxsf-Ir-	1<>0>>>2<<

RANDOM TARGET ITEMS

letters	phonemes	stress
fozepd	WdicnK	1<121>
sccfyk	p-UdSp	>202<1
bmyqcl	bzgTlz	0>><<>
grtufh	KCczOL	>1<010
eqhxxu	ANT|vM	>01<>2
ncssvr	zTSdWg	<<12>2
wxsale	RKpfI|	1<1110
djzxde	Yby^yI	20>>2>
kmfjqi	WGenGN	1><102

training corpus are presented in Table 1.

Procedure

The twenty target items were tested individually on separate trials. A trial consisted of first reading in the pre-experimental weights (described above), presenting a target item either two, ten, or twenty times, and then measuring the retention of the target as it was interfered with by subsequent learning. Furthermore, each target was presented at each of six spacing intervals, with either 0 (massed), 1, 4, 8, 20, or 40 (distributed) intervening items. Thus, eighteen trials were devoted to each target item (3 repetition groups x 6 spacing intervals). Between successive repeats of the target, words were presented from the original training corpus. Following the last repeat, the training corpus was again presented, and retention of the response terms of the target item was assessed after every item by presenting the cue term and measuring the mean squared difference between the output of the network and the correct

response

$$error = \frac{\sum_{j=1}^{J} (p_j^* - p_j)^2}{J} \tag{3}$$

for the J units in the output layer, where p_j^* is the target activation of the jth output unit, and p_j is it's actual value. Response accuracy was defined as one minus the mean error for the word

$$accuracy = 1 - \frac{\sum_{n=1}^{N} error}{N} \tag{4}$$

where N is the number of letters in the word. Learning was turned off (achieved by setting the learning rate to zero) for these tests, so that no changes were made to the strengths of the connections in the network.

RESULTS

Accuracy, as defined above, was averaged over the twenty target items and plotted as a function of retention interval for each repetition group (Figure 3). Following Glenberg (1976), values were selected from this curve at retention intervals of 2, 8, 32, and 64 items and re-plotted as a function of spacing interval (Figure 4).

A 6 (spacing intervals) x 4 (retention intervals) analysis of variance was performed on these selected values, treating target items as subjects. The main effect of retention interval was highly significant in all repetition groups, F (3, 57) = 32.82, 58.50, and 48.29, all $p <$ 0.001, for the two, ten, and twenty repetition groups, respectively, indicating that a considerable amount of forgetting of the target items did take place. The main effect of spacing was highly significant only in the twenty repetition condition, F (5, 95) = 5.10, $p <$ 0.001, and marginally significant in the ten repetition condition, F (5, 95) = 3.02, $p < 0.05$. Of interest, however, was the interaction between spacing and retention interval. This interaction was significant for all three repetition groups: F (15, 285) = 27.68 and 37.29, both $p < 0.001$, for the ten and twenty repetition conditions, respectively, and F (15, 285) = 2.73, p < 0.03, in the two repetition condition.

A trends analysis of the accuracy measures was performed across spacings for retention intervals of 0 (short-term) and 64 (long-term) items. The downward trend in retention for immediate retention as spacing increased was highly significant following ten and twenty repeats of the target, F (5, 95) = 17.14 for the ten, and F (5, 95) = 6.70 for the twenty repetition groups, both $p < 0.001$. In both cases, the linear trend was highly significant, F (1,

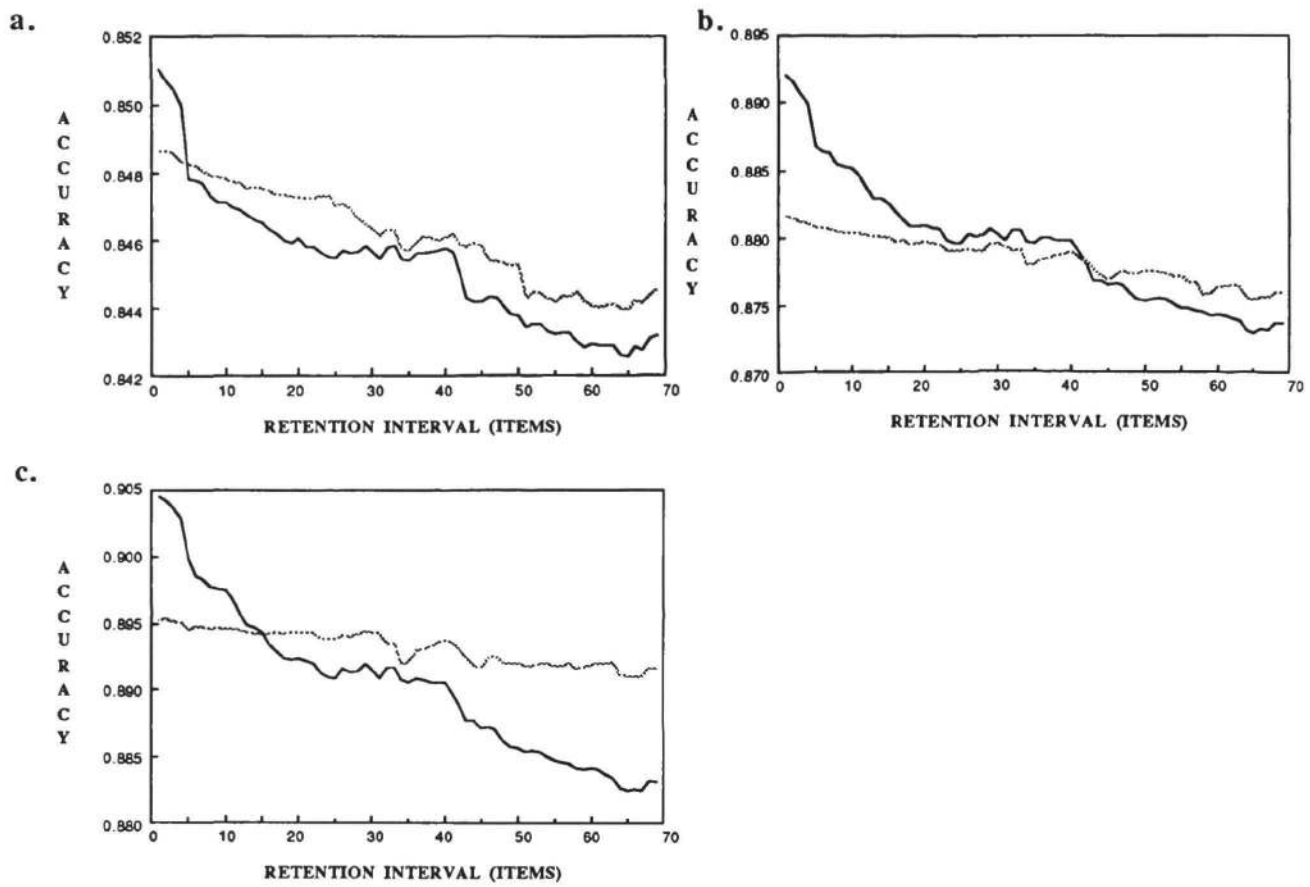


Figure 3. Mean response accuracy over all target items plotted as a function of retention interval following two (a), ten (b), and twenty (c) repetitions of the target item. Only spacing intervals of zero (solid) and forty (dots) items are shown.

19) = 38.43, $p < 0.001$, and F (1, 19) = 11.80, $p < 0.001$, for the ten and twenty repetition groups, respectively. This downward trend was not significant after only two repeats, however, F (5,15) < 0.5. Neither the quadratic nor the cubic trends reached significant levels in any of the three repetition groups.

The upward trend at the 64-item retention interval was significant for all repetition groups, F (5, 95) = 2.31, $p = 0.05$, for two repeats, and F (5, 95) = 8.92 and 22.40, both $p <$ 0.001 for the ten and twenty repetition groups. The shape of the curve varied, however. As the number of repetitions increased from two to ten to twenty, the trend varied from cubic, F (1, 19), $p < 0.05$, to linear, quadratic, and cubic, F (1, 19) = 6.06, 24.94, and 6.12, $p < 0.05$, 0.001, and 0.05, respectively, to only linear and quadratic, F (1, 19) = 16.21 and 66.63, both p < 0.001.

DISCUSSION

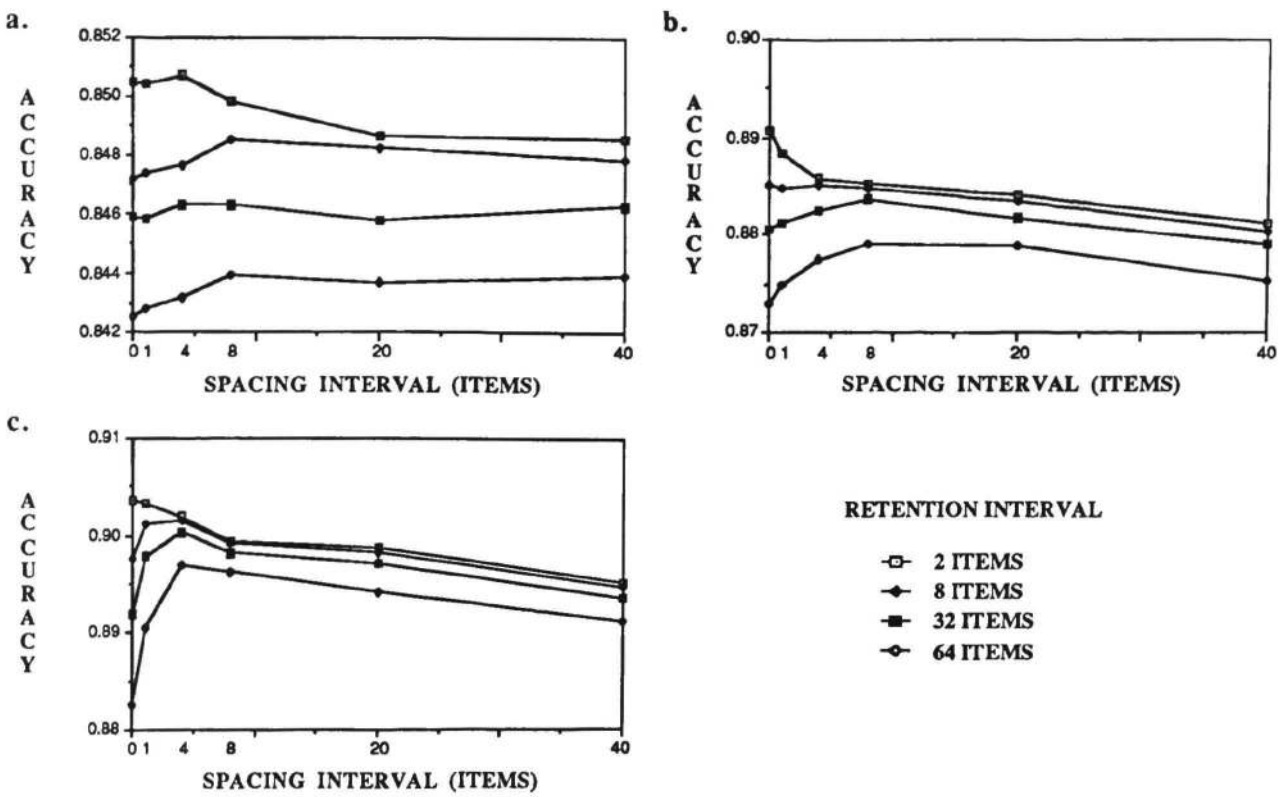


Figure 4. Mean response accuracy plotted as a function of spacing interval at 2, 8, 32, and 64 item retention intervals for the two (a), ten (b) and twenty (c) repetition groups.

A significant spacing effect was observed in NETtalk: Retention of nonwords after a 64-item retention interval was significantly better when presented at the longer spacings (distributed presentation) than at the shorter spacings. In addition, a significant advantage for massed presentations was found for short-term retention of the items. Although stimulus materials, response measures, and procedure differ sufficiently to make direct comparison impossible, the overall pattern of these results resembles that found by Glenberg (1976), in an experiment using human subjects. We obtained our results without making additional assumptions or including additional mechanisms such as consolidation, rehearsal, or attention. Nor were explicit assumptions made about a continuously changing context other than the context implicitly provided by the network.

Recency effects, similar to those reported here, are common in the human literature and have been reported in spacing experiments (e.g. Peterson, Wampler, Kirkpatrick, & Saltzman, 1963; Sperber, 1974). This short-term advantage for massed practice is commonly discussed with reference to a limited-capacity memory buffer. The present experiments indicate that some of the effects such a mechanism was designed to account for can be produced without such a device.

Why should NETtalk exhibit these characteristics? The answer, as we attempt to show, depends on the way in which learning and the resulting knowledge is represented in NETtalk.

An intuitive understanding of learning in NETtalk can be obtained by thinking of the set of n weight values, the many sites of learning in the network, as specifying a point in an n-dimensional hyperspace. The goal of learning is to move on a trajectory through this hyperspace towards a point where the error on the entire training corpus is minimized. The direction in which to move at any one point is determined by estimating the error gradient. If a global minimum for a given corpus can be reached, no further learning (i.e. weight adjustments) is required. A minimum is consequently a point of high stability, *until* a new item is presented that is irregular, or is for some other reason not like the other items in the training corpus. Our hypothesis is that distributing practice leads to a more stable position in this hyperspace upon the re-presentation of the training corpus.

For the sake of simplicity, suppose that NETtalk has only three connections, so that its state at any one time can easily be represented in a 3-dimensional space (see Figure 5). Suppose further that, as in the present simulations, this network has been trained on a large pre-experimental training corpus and that it has reached an optimum (where the error is at a global minimum) for these items (Point A). Now a new and unusual target item is presented in either a massed or spaced condition to our mini-network. If the target is presented several times back-to-back, as in the massed condition, minimizing the error following each presentation will lead us down a path toward a point that is optimal for this target item, perhaps even reaching this optimum (Point B). But because this voyage will have taken us quite out of the way from our starting point (A), this new position is not likely to be stable to the re-presentation of the training corpus, and so the massed learning of the new item will be lost quickly.

Assuming, however, that there *is* a point that is optimal for both the training corpus and the target item (Point C in the figure), alternating presentations of the target with items from the training set is one way of moving closer to this highly stable point.[7] Upon the first presentation of the target item, the error gradient for that item is estimated and the error is reduced by adjusting the weights in the direction of the steepest descent (to position 1). So far, this procedure has been identical to that for the massed condition, and so the network is at

[7] Another way is to update the weight values less frequently. Instead of learning in small increments, as in NETtalk, which updates after every word, one could also collect data over many trials and then take one big jump. Although this procedure (within its resolution) overcomes the problems associated with presentation order (such as the spacing effect), it may be a hazardous one, since new information is integrated at a slow rate.

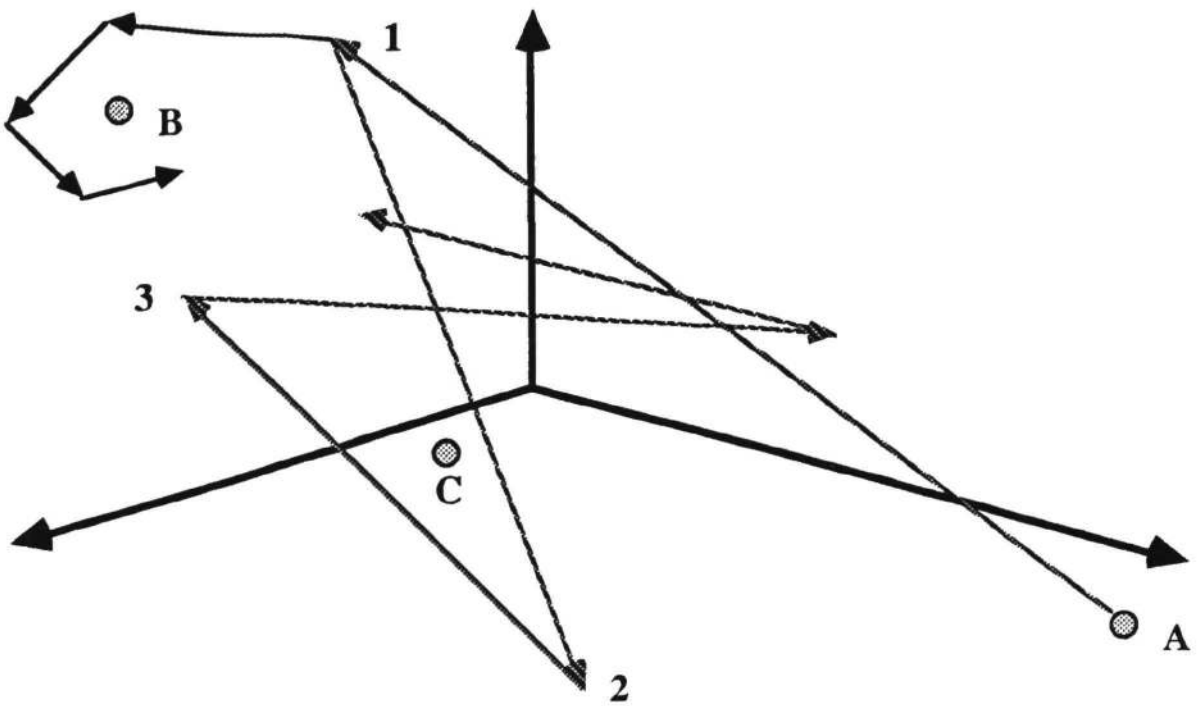


Figure 5. Movements in weight-space during learning for massed (solid) and distributed (textured) conditions. Point A is a global optimum for the pre-experimental training corpus (the assumed starting point for all experimental trials), Point B is an optimum for the target item, and Point C is an optimum for both the target and the training corpus. (See text for explanation.)

the same point in hyperspace. Now, however, instead of presenting the target again, an item from the original training corpus is presented. Again, the weights are adjusted to minimize the error on the item (to position 2), only this time the direction of movement is more likely to be towards Point A than Point B, since A was a global minimum for the training corpus. Presenting the target again will cause a movement back towards B (to position 3), and so on. We see that distributing practice causes the network to weave back and forth in this hyperspace, allowing it to perform a more complete search of the error space for *both* the training corpus *and* the target item. The network therefore has a better chance of finding the optimal position (Point C) than it would if practice were massed, and its encoding of the target item will consequently be more able to withstand interference due to further training on both types of material.

The explanation of the spacing effect that we offer here is not meant as an alternative to previous suggestions; it is a different type of explanation, relying as it does on the underlying structure of the representations. The decline in learning rate as local optima are approached is reminiscent of the process of habituation: less is effectively learned each time the item is

repeated. Other aspects of our model bear a resemblance to encoding variability to the extent that items are encoded relative to the current state of the network, which is in a state of continual flux. And if we identify Jacoby's processing effort with the degree of change required to construct a distributed representation, then our simulations can be considered support for this proposal as well. Nevertheless, while these concepts of habituation, encoding variability, and processing effort may be reinterpreted within the framework of connectionist models such as ours, they are at a different level of explanation.

Our results are limited to a particular network architecture in a particular domain. To what extent is this conclusion dependent on the details of our model? If the spacing effect is a direct consequence of incremental learning in memory systems that use distributed representations, as we suspect, then the same effects of massed and distributed learning should occur in other task domains and with other network architectures that also have learning algorithms with distributed representations, such as Boltzmann machines (Hinton & Sejnowski, 1983; Ackley, Hinton, & Sejnowski, 1985). We predict as well that the same general principles may underlie the spacing effect in human learning.

References

Ackley, D.H., Hinton, G.E. , & Sejnowski, T.J. (1985). A Learning Algorithm for Boltzmann Machines. *Cognitive Science, 9*, 147-169.

Amari, S. (1977). Neural Theory of Association and Concept Formation. *Biological Cybernetics, 26*, 175-185.

Anderson, J.A., Silverstein, J.W., Ritz, S.A., & Jones, R.S. (1977). Distinctive Features, Categorical Perception, and Probability Learning: Some Applications of a Neural Model. *Psychological Review, 84*, 413-451.

Atkinson, R.C., & Shiffron, R.M. (1968). Human Memory: A Proposed System and Its Control Processes. In K.W. Spence & J.T. Spence (Eds.), *The Psychology of Learning and Motivation: Advances in Research and Theory* (Vol. 2). New York: Academic Press.

Ballard, D.H., Hinton, G.E. , & Sejnowski, T.J. (1983). Parallel visual computation. *Nature, 306*, 21-26.

Bjork, R.A., & Allen, T.W. (1970). The Spacing Effect: Consolidation or Differential Encoding. *Journal of Verbal Learning and Verbal Behavior, 9*, 567-572.

Ebbinghaus, H. (1964). *Memory: A Contribution to Experimental Psychology* (originally published, 1885) . New York: Dover.

Estes, W.K. (1959). The Statistical Approach to Learning Theory. In S. Koch (Ed.), *Psychology: A Study of a Science* (Vol. II). New York: McGraw-Hill.

Feldman, J.A., & Ballard, D. (1982). Connectionist Models and Their Properties. *Cognitive Science, 6*, 205-254.

Glenberg, A.M. (1976). Monotonic and Nonmonotonic Lag Effects in Paired-Associate and Recognition Memory Paradigms. *Journal of Verbal Learning and Verbal Behavior, 15*, 1-16.

Glenberg, A.M. (1979). Component-levels Theory of the Effects of Spacing of Repetitions on Recall and Recognition. *Memory and Cognition, 7*, 95-112.

Hinton, G.E., & Sejnowski, T.J. (1983). *Optimal Perceptual Inference.* Washington, D. C.: Proceedings of the IEEE Computer Society Conference on Computer Vision & Pattern Recognition.

Hintzman, D.L. (1974). Theoretical Implications of the Spacing Effect. In R.L. Solso (Ed.), *Theories in Cognitive Psychology: The Loyola Symposium.* Hillsdale, N.J.: Erlbaum.

Hintzman, D.L. (1976). Repetition and Memory. In G.H. Bower (Ed.), *The Psychology of Learning and Motivation* (Vol. 10). New York: Academic Press.

Hintzman, D.L., Block, R.A., & Summers, J.J. (1973). Modality Tags and Memory for Repetitions: Locus of the Spacing Effect. *Journal of Verbal Learning and Verbal Behavior, 12*, 229-238.

Hintzman, D.L., Summers, J.J., & Block, R.A. (1975). What Causes the Spacing Effect? Some Effects of Repetition, Duration, and Spacing on Memory for Pictures. *Memory and Cognition, 3*, 287-294.

Hintzman, D.L., Summers, J.J., Eki, N.T., & Moore, M.D. (1975). Voluntary Attention and the Spacing Effect. *Memory and Cognition, 3*, 576-580.

Jacoby, L.L. (1978). On Interpreting the Effects of Repetition: Solving a Problem Versus Remembering a Solution. *Journal of Verbal Learning and Verbal Behavior, 17*, 649-667.

Kohonen, T., Oja, E., & Lehtio, P. (1981). Storage and Processing of Information in Distributed Associative Memory Systems. In G.E. Hinton & J.A. Anderson (Eds.), *Parallel Models of Associative Memory*. Hillsdale, N.J.: Erlbaum.

Kucera, H., & Francis, W.N. (1967). *Computational Analysis of Modern-Day American English*. Providence, R.I.: Brown University Press.

Landauer, T.K. (1969). Reinforcement as Consolidation. *Psychological Review, 76*, 82-96.

Maki, R.H., & Hasher, L. (1975). Encoding Variability: A Role in Immediate and Long-term Memory?. *American Journal of Psychology, 88*, 217-231.

Martin, E. (1968). Stimulus Meaningfulness and Paired-Associate Transfer: An Encoding Variability Hypothesis. *Psychological Review, 75*, 421-441.

McClelland, J.L., & Rumelhart, D.E. (1985). Distributed Memory and the Representation of General and Specific Information. *Journal of Experimental Psychology: General, 114*, 159-188.

Melton, A.W. (1970). The Situation with Respect to the Spacing of Repetitions and Memory. *Journal of Verbal Learning and Verbal Behavior, 9*, 596-606.

Peterson, L.R., Wampler, R., Kirkpatrick, M., & Saltzman, D. (1963). Effect of Spacing Presentations on Retention of a Paired-Associate Over Short Intervals. *Journal of Experimental Psychology, 66*, 206-209.

Postman, L., & Knecht, K. (1983). Encoding Variability and Retention. *Journal of Verbal Learning and Verbal Behavior, 22*, 133-152.

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

Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). Learning Internal Representations by Error Propagation. In D.E. Rumelhart & J.L. McClelland (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Cambridge, Mass.: MIT

Press.

Rundus, D. (1971). Analysis of Rehearsal Processes in Free Recall. *Journal of Experimental Psychology, 89*, 63-77.

Sejnowski, T.J., & Rosenberg, C.R., NETtalk: A Parallel Network that Learns to Read Aloud, The Johns Hopkins University Electrical Engineering and Computer Science Technical Report JHU/EECS-86/01, 1986.

Shaughnessy, J.J., Zimmerman, J., & Underwood, B.J. (1972). Further Evidence on the MP-DP Effect in Free-Recall Learning. *Journal of Verbal Learning and Verbal Behavior, 11*, 1-12.

Sperber, R.D. (1974). Developmental Changes in Effects of Spacing of Trials in Retardate Discrimination Learning and Memory. *Journal of Experimental Psychology, 103*, 204-210.

Tzeng, O.J.L. (1973). Stimulus Meaningfulness, Encoding Variability, and the Spacing Effect. *Journal of Experimental Psychology, 99*, 162-166.

Processing Verb Phrase Anaphors[1]

Michael K. Tanenhaus
University of Rochester

and

Greg N. Carlson
University of Iowa

In this paper we present three experiments which investigate the hypothesis proposed by Hankamer and Sag (1976) that there are two distinct kinds of anaphors in natural language--"deep" and "surface" anaphors. "Deep" anaphors in English include among other things definite pronouns, "One"-pronominals, and Null Component anaphora, exemplified in (1) below.

1. a. John left. He was angry. (Definite pronoun)
   b. Mary bought a green car. Frank bought a red one. (One pronominal)
   c. Mary knew who was guilty. But she wouldn't tell___. (Null Component)

"Surface" anaphors, on the other hand, include examples of Verb Phrase Ellipsis, Gapping, and Sluicing.

2. a. Sander built a new house. Max did___, too. (Verb Phrase Ellipsis)
   b. William caught a barracuda, and Harry,___, a shark.
   c. Someone just called you. But I don't know who___.

All anaphors, we assume, fall into one of these two classes.

One of the primary differences between the two categories is that deep but not surface anaphors may find their antecedents in the general context of use (e.g. something pointed at or otherwise made salient). Surface anaphors, unlike deep anaphors, require the presence of a linguistically-expressed antecedent. Consider, for instance, a situation in which two people are watching a fisherman reel in a fish. Under those circumstances, one of the bystanders could turn to the other and say, "Do you think he'll eat it?" (using a deep anaphor), but not "And Bill's nephew,___a rainbow trout" (infelicitiously using a surface anaphor, meaning Bill's nephew caught a rainbow trout). Thus, deep anaphors take antecedents which may or may not be linguistically-introduced antecedents. We will assume that deep anaphors find their antecedents in some non-linguistic form of representation, while surface anaphors seek antecedents among linguistic representations.

Moreover, we follow Sag and Hankamer (1984) in assuming that the level of linguistic structure in which surface anaphors find their antecedents is more abstract than surface structure; in fact, a level of logical form seems to be the most appropriate level at the present time (this is a level of linguistic representation in which scope of operators, such as quantified NP's, is unambiguously represented). Although multiple levels of linguistic representation are assumed by many formal linguistic and AI theories, little if any processing evidence supports the need for such abstract levels. Deep and surface anaphors offer an ideal contrast for studying the processing of these representations, because in the same context of interpretation, they may index different aspects of mental representations on the way to the same final interpretation.

One of the more compelling arguments for a deep-surface distinction is that surface anaphors seem to require that their antecedents be constituents at the appropriate level of linguistic representation, whereas deep anaphors do not. Consider the examples in (3).

3. a. Someone has to take out the garbage.
   b. The garbage has to be taken out.
   c. But Bill refused to___.
   d. But Bill refused___.

Sentence (3a) may be felicitously followed by either (3c) or (3d) equally well, though (3c) is a case of surface VP-Ellipsis, whereas (3d) is an example of a deep Null Complement anaphor. However, if both are preceded in a discourse by (3b) instead, (3c)--the surface anaphor--becomes infelicitous, whereas (3d)--the deep anaphor--remains perfectly acceptable and easily interpretable. This is because in the logical form of (3a) the verb phrase ("take out the garbage") is a constituent, as it is on the surface, whereas the logical form of (3b) has no constituent assigned the meaning "take out the garbage," again the same as surface form in which "the garbage" and "(be) taken out" do not form a single constituent. Following Hankamer and Sag, we will refer to this as the "parallelism" requirement of surface anaphors, reflecting the intuition that the antecedent of a surface anaphor requires structure parallel to that required at the site of the anaphor itself; in the case of (3c) a VP, missing at the site of the anaphor, is required as an antecedent. On the other hand, deep anaphors find antecedents based on knowledge of situations and other conceptual phenomena, where the linguistic notions of category and constituency do not come directly into play. Hence there is no "parallelism" requirement for deep anaphors, like (3d).

In three experiments we manipulated the parallelism of the antecedent. Using a "makes sense" judgment task (henceforth, "the judgment task"), we asked subjects to read a context sentence and then decide whether a subsequently presented target sentence made sense given the context. This task was chosen because it provides both judgment and reaction time data, and because it requires the subjects to integrate the anaphor with preceding discourse in order to make the judgment. All of the context and target sentences were grammatical sentences in English, but some of the filler targets did not make sense given the context (e.g. "Bill won first prize. He was glad that he didn't___," or "Tom took out the garbage willingly. He objected to doing it.")

In our first experiment, we created non-parallel antecedents by changing active sentences into passives (e.g. "Somebody had to take out the garbage" vs. "The garbage had to be taken out.") The passive creates a non-parallel antecedent because the VP in the active (the only reasonable antecedent in the context) is no longer a constituent in the passive. We reasoned that if deep anaphors find their antecedents in conceptual representations, they should be equally comprehensible with both parallel and non-parallel antecedents as the context sentences should give rise to the same or nearly the same conceptual representations (though see the general discussion below).

Sample materials for this experiment are illustrated in (4). The parallel antecedent is represented in context sentence (4a), the non-parallel antecedent in context sentence (4b). The surface and deep anaphors (with a definite pronoun) were presented as exemplified in target sentences (4c) and (4d), respectively.

4. a. Someone had to take out the garbage.
   b. The garbage had to be taken out.
   c. But Bill refused to___.
   d. But Bill refused to do it.

Twenty sets of materials, similar to (4), with fillers, were counterbalanced across four presentation lists. The results for 32 subjects are presented in Table 1. An interaction was obtained between parallelism and type of anaphor in the judgment data, with parallelism affecting judgments to surface but not deep anaphors, and in the reaction time data. However, contrary to our expectations, parallelism did significantly affect comprehension times to deep anaphors as well as surface anaphors.

In the second experiment, we created non-parallel antecedents in a different way, by presenting antecedents in a nominalized form. This presents the non-parallel antecedent as a constituent (unlike in experiment 1), but a constituent of the wrong category for the surface anaphor--an *N* instead of a *VP*. Sample materials are presented in (5). The parallel and non-parallel antecedents are presented in the context sentences (5a) and (5b), respectively, and the surface and deep anaphors in target examples (5c) and (5d), respectively.

5. a. It always annoys Sally when anyone mentions her sister's name.
   b. The mention of her sister's name always annoys Sally.
   c. However, Tom did anyway out of spite.
   d. However, Tom did it anyway out of spite.

The results for 28 subjects are presented in Table 2. Again, parallelism interacted with type of anaphor in the judgment task, with parallelism having no effects on the proportion of deep anaphors judged to make sense, but strong effects on the surface anaphors. There was a main effect of parallelism in the reaction time data and no interaction with type of anaphor.

Summarizing the results of the first two experiments, we find that parallelism did not affect judgments to deep anaphors, whereas it had robust effects on judgments to surface anaphors. In contrast, reaction time (when the anaphor was judged to make sense) was affected by parallelism for both types of anaphors. Why should non-parallelism increase reaction times to comprehend deep anaphors, but have no effect on judgments? One explanation, proposed by Murphy (1982), is that deep anaphors in some way contain more clues to the nature of the antecedent, and are thus less dependent on surface form than surface anaphors. The partial dependence on surface forms accounts for the increased reaction times, while the additional "clues" facilitate judgment.

In order to test this hypothesis, as well as to test the more general observation that the critical factor in our experiments could be the contrast between null anaphors and phonologically-realized anaphors (e.g. "it"), we conducted an experiment that contrasted Null Complement anaphors (e.g. "Bill refused ___" as in (3d)) with Verb Phrase Ellipsis (e.g. "Bill refused to___" as in (3c)). Both anaphors are null. On Murphy's analysis, though, Null Complement anaphors contain fewer cues than VP-Ellipsis anaphors, although Null Complement anaphors fall into the category of deep, rather than surface, anaphors.

Twelve sets of materials were constructed in which parallel and non-parallel antecedents preceded either a Null Complement anaphor or a VP-Ellipsis anaphor, as exemplified above in (3). The materials were counterbalanced across four presentation lists, and 40 subjects were tested using the judgment task. Only twelve sets of materials were used because the number of English verbs that allow the construction of a Null Complement/VP-Ellipsis contrast is limited. The results are presented in Table 3. Again we see a robust interaction between type of anaphor and parallelism of the antecedent in the percentage of sentences judged to make sense, and a main effect of parallelism in the comprehension time data.

General Discussion

Summarizing the results of the three experiments, we find that the parallelism did not affect judgments to deep anaphors, while it had a robust effect on judgments to surface anaphors. In contrast, reaction time (when the anaphors were judged to make sense) was affected by parallelism for both types of anaphors. Our hypotheses account for the difference in judgments, but do not explain why parallelism should not differentially affect reaction times. In particular, why should lack of parallelism increase reaction time to deep anaphors? There would appear to be two possible accounts consistent with our hypotheses. One is that subjects were making grammaticality judgments. In checking for grammaticality, attention is paid to the linguistic form antecedents are expressed in. The non-consistency of the antecedents of the passives requires extra steps in checking plausibility, and the nominalized forms represent the less usual sort of antecedent for pronouns standing for activities, verbal phrases being the more expected type of antecedent. But this hypothesis seems implausible given that all the sentences given subjects were in fact grammatical, and that subjects were basing judgments on understanding the sentences in context instead of attending to their formal properties. A more likely possibility, which we plan to pursue in further research, is that the manipulations we used to create non-parallel antecedents at the level of linguistic representation also had effects at the conceptual or discourse model level. More specifically, passivizing a sentence shifts focus to the underlying direct object (the surface subject), now in contrast to the backgrounded verb and possibly other remaining VP material emphasizing disunity, unlike the active counterpart. Similarly, the nominalized forms used in the second experiment present the actions as if presupposed or otherwise backgrounded, in contrast to the verbal forms (Kiparsky and Kiparsky (1971)). In both cases, using a (deep) VP anaphor requires a search which requires a shift of focus. These focus shifts should take time, but result in antecedents that are completely comprehensible. Thus, the focus hypothesis promises an account of the complete dissociation between the effects of non-parallelism on reaction time, and the effects on comprehensibility as indexed by the proportion of sentences judged to make sense.

In summary, our experiments support a processing difference between deep and surface anaphors. They further suggest that surface anaphors take linguistic antecedents. These results do not, however, address the question of what level or aspects of linguistic representation are important for the online interpretation of surface anaphors. In work in progress we are addressing this issue by manipulating parallelism in different ways in order to see which aspects of representation are important for the comprehension of surface anaphors.

Table 1

Results of the Passive Experiment

	Antecedent Type			
	Active (Parallel)		Passive Non-Parallel)	
Anaphor	% judged to make sense	(RT)	% judged to make sense	(RT)
Deep	94%	2181 msec	91%	2381 msec
Surface	89%	2165 msec	70%	2848 msec

Note: Judgments are to the sentence with the anaphor and reaction times are to those sentences judged to make sense.

Table 2

Results of the Nominalization Study

	Antecedent Type			
	Active (Parallel)		Passive (Non-Parallel)	
Anaphor	% judged to make sense	(RT)	% judged to make sense	(RT)
Deep	87%	2686 msec	87%	2952 msec
Surface	89%	2557 msec	71%	2923 msec

Note: Judgments are to the sentence with the anaphor and reaction times are to those sentences judged to make sense.

Table 3

Results of Null Complement Verb Phrase Ellipsis Experiment

	Antecedent Type			
	Parallel		Non-Parallel	
Anaphor	% judged to make sense	RT(msec)	% judged to make sense	RT(msec)
Null Comp. (Deep)	93%	18250	89%	21525
VPE	95%	20153	77%	21208

Note: Judgments are to the sentence with the anaphor. Reaction time data are to sentences judged to make sense.

References

Hankamer, J. & Sag, I. (1976) Deep and surface anaphora. Linguistic Inquiry, 7, 391-426.

Kiparsky, P., & Kiparsky, C. (1971) Fact. In D. Steinberg & L. Jakobovits, Semantics. Cambridge, England: Cambridge University Press, 345-369.

Murphy, G. (1982) Understanding anaphora. Stanford University doctoral dissertation.

Sag, I., & Hankamer, J. (1984) Toward a theory of anaphoric processing. Linguistics & Philosophy, 7, 325-345.

[1]This research was supported by Grant BNS-8217378 from the National Science Foundation. We thank Ed Sclafani and Sania Hamilton for collaborating with us on Experiment 3.

The effect of the discourse center on the local coherence of a discourse[1]

Susan B. Hudson, Michael K. Tanenhaus, and Gary S. Dell
University of Rochester

This paper presents two experiments that test the notion of a discourse center introduced by Grosz, Joshi, and Weinstein (1983). The discourse center is a central component of a larger theory of discourse structure being developed by Grosz and Sidner (1985).

Grosz et al. (1983) have defined two levels of discourse coherence. Large segments of a discourse are related to one another by a process, namely focusing, which maintains the global coherence. Centering is an additional process which aids in the local coherence of a discourse. Local coherence is defined as the coherence between adjacent utterances. A forward-looking center provides entities to which the remaining discourse may be tied. A backward-looking center (hereafter the center) connects the current sentence with the immediately preceding discourse. The center is that element from all the focused elements that the utterance is about.

Grosz et al. suggest different roles played by pronouns and nouns in discourse coherence. Pronouns most often serve in identifying the single entity the discourse is about. Noun phrases, on the other hand, are most often used to shift the focus of the discourse and as such they are related to the global coherence of the discourse. Thus the use of a noun phrase rather than a pronoun to refer to the center is somewhat unnatural as seen in the fact that sentence (1b) more naturally follows (1a) than (1c).

(1) a. Who did Max see yesterday?
b. He saw Rosa.
c. Max saw Rosa.

We tested two predictions generated from the discourse center hypothesis. First, an ambiguous pronoun, e.g., the pronoun "she" when there are two female antecedents in the discourse will be interpreted immediately with the antecedent of the pronoun assumed to be the discourse center. If the correct antecedent later turns out not to be the center, the pronoun will have to be reinterpreted resulting in increased processing time. The second prediction is that a pronoun will be more rapidly understood than a noun phrase when both refer to the center of the preceding sentence.

The materials were two-sentence discourses in which the first sentence, hereafter the context sentence, introduced two possible antecedents and the second sentence, hereafter the target sentence, began with either a pronoun or a proper noun which

referred to either the centered or noncentered entity in the context sentence. The subject of the context sentence was established as the discourse center by using verbs in which the subject was likely to be the perceived cause of the event described by the verb (Brown and Fish, 1983; Newman, 1984). Implicit causality of verbs has been shown to control antecedent assignment in sentences with ambiguous pronuouns in studies by Caramazza and colleagues (Garvey & Caramazza, 1974; Garvey, Caramazza, & Yates, 1975; Caramazza, Grober, Garvey, & Yates, 1977) and more recently by Newman (1984).

Example materials are presented in (2). The context sentence is given in (2a). Target sentences beginning with a pronoun and noun that refer to the centered entity are given in (2b) and (2c), respectively, and target sentences beginning with a pronoun and noun that refer to the noncentered entity are given in (2d) and (2e), respectively.

(2) a. Jack apologized profusely to Josh.
b. He had been rude to Josh yesterday.
c. Jack had been rude to Josh yesterday.
d. He had been offended by Jack's comment.
e. Josh had been offended by Jack's comment.

Twenty sets of materials similar to those given in (2) were used in Experiment 1. The four target sentences for each context sentence were counterbalanced across four presentation lists. The test sentences were intermixed with sensible and nonsensible fillers. An example of a nonsensible filler context and continuation sentence is presented in (3a) and (3b), respectively.

(3) a. John couldn't mail Tim a check.
b. He was proud that Tim was able to.

Twenty University of Rochester volunteers served as subjects. Their task was to read each sentence carefully and, when cued, to indicate whether or not the two-sentence texts were comprehensible. The sentences were presented visually on a CRT. On each trial the context sentence was displayed. Upon reading the context sentence the subject pressed a button. The context sentence was removed from the screen and the target sentence was immediately presented. Upon reading the target sentence the subject again pressed the button. The target sentence then disappeared from the screen, and a question mark appeared. The subject then judged whether or not the text made sense by pressing one of the two response buttons.

The results are presented in Table 1. The conditions are labeled according to whether a pronoun or noun was used as an anaphor and whether the pronoun or noun referred to the centered or noncentered discourse entity. The percentage of sentences judged to make sense are in parentheses. Reading times are for

target sentences judged to make sense.

A 2X2 ANOVA with Type of Antecedent and Type of Anaphor as factors was conducted on both the reading times for the target sentences and the percentage of cases in which the texts were judged "sensible." For the judgment data, there was a significant interaction both by subjects ($F(1,15)=17$, $p<.001$) and by items ($F1,18)=8.69$, $p<.008$), and for the reaction data, there was a significant interaction by subjects ($F(1,15)=5.643$, $p<.03$) and a nearly significant interaction by items ($F1,18)=3.69$, $p<.08$).

There were two primary hypotheses. The first was that the initial phase of the target sentence would be read more rapidly when it began with a pronoun than when it began with a noun, because the pronoun's antecedent would be immediately interpreted as the discourse center. Second, if the pronoun's antecedent turns out not to be the center, readers will have been led down the "garden path" and they will be forced to reprocess the sentence. Thus we expect that in these cases subjects will either judge these sentences to be nonsensical or will take a relatively longer time to determine that the sentences do make sense.

The main predictions were confirmed. Target sentences that began with a pronoun were read more rapidly and judged to be sensible more often when the antecedent was not the center. Sentences in which the subject of the target sentence referred to the centered entity were read more rapidly when the sentence began with a pronoun than when the sentence began with a noun.

The second experiment was conducted to replicate the first experiment and to provide more local information about the pronoun assignmment. We used materials similar to those in the first experiment but divided the target sentence into two phrases so that the disambiguating information always came in the second phrase. Example materials are presented in (4). The context sentence is presented in (4a). Target sentences were either consistent or inconsistent with the centered agent in (4b-c) and (4d-e), respectively. Target sentences beginning with a pronoun are presented in (4b) and (4d) and target sentences beginning with a proper name are presented in (4c) and (4e). The slash mark indicates where the sentences were broken into phrases.

(4) a. Jack apologized to Josh.
b. He hadn't even/ noticed Josh.
c. Jack hadn't even/ noticed Josh.
d. He hadn't even/ noticed Jack.
e. Josh hadn't even/ noticed Jack.

The experimental materials were counterbalanced over four presentation lists. The test sentences were intermixed with sensible and nonsensible fillers. Twenty-four University of Rochester volunteers served as subjects. The context sentence was

displayed on a CRT followed by a target sentence. When the first phrase of the target sentence was displayed subjects pressed a button which resulted in the presentation of the second phrase. The subject then decided whether or not the entire target sentence was comprehensible with respect to the context sentence. This judgment was indicated by a YES or NO response.

Table 2 presents the proportion of sentences judged to make sense and the reading times to those sentences. A 2X2 ANOVA with Type of Anaphor and Type of Antecedent as factors was conducted on the reading time data for the first phrase and both the judgment and reading time data for the second phrases.

The reading time data for the first phrase support the first prediction from the center hypothesis. Sentences with pronouns were read faster than sentences beginning with nouns. This was reflected in a significant effect of Type of Anaphor ($F(1,19)=7.615$, $p<.01$ by subjects and $F(1,19)=5.738$, $p<.03$ by items). Somewhat surprisingly, the reading times to the noncentered noun were faster than the centered noun, suggesting that beginning the target sentence with the noncentered noun violated reader's expectations. The comparison between centered pronouns and nouns was in the right direction, with phrases begining with centered nouns taking longer to read, but the difference did not reach significance ($F(1,19)=2.573$, $p<.12$ by subjects and $F(1,19)=2.79$, $p<.11$ by items).

The second phrase judgment data strongly support the second prediction. There was a robust interaction between Type of Anaphor and Type of Antecedent in the subject and item analyses ($F(1,19)=19.54$, $p<.0004$ by subjects and $F(1,19)=38.94$, $p<.00003$ by items). As in the first experiment, subjects frequently rejected sentences with pronouns that referred to the noncentered entity. The reading time data are less clear. As expected, reading times were longest when the second phrase indicated that the pronoun in the first phrase referred to the noncentered entity. However, contrary to our expectations, the fastest second phrase reading times obtained when the first phrase contained a centered noun. Overall, there was a significant effect of centeredness, ($F(1,19)=7.77$, $p<.01$ by subjects only), indicating that having a noncentered entity as the subject of the sentence interfered with processing, even when the noncenter was unambiguously introduced as a noun.

Conclusion

The results of both experiments lend strong support to the discourse center hypothesis proposed by Grosz et al (1981). When readers encounter an ambiguous pronoun, they immediately assume that its antecedent will be the discourse center of the previous sentence. Moreover, sentences which continue with the same center as the preceding sentence are read more rapidly and judged to be

more comprehensible when the center is mentioned as a pronoun than when the center is mentioned as a noun. However, the mechanics of center shifting remain unclear, and our future research will focus on these issues.

TABLE 1

	Type of Antecedent	
Type of Anaphor	Center	Noncenter
Pronoun	2158 (97%)	2644 (78%)
Noun	2475 (94%)	2422 (95%)

Reading times for the target sentences in Experiment 1 with the percentage of sentences judged to be sensible in parentheses.

TABLE 2

	Type of Antecedent	
	Center	
Type of Anaphor	Phrase 1	Phrase 2
Pronoun	He hadn't even 961	noticed Josh. 2026 (85%)
Noun	Jack hadn't even 1075	noticed Josh. 1687 (75%)
	Noncenter	
	Phrase 1	Phrase 2
Pronoun	He hadn't even 970	noticed Jack. 2368 (60%)
Noun	Josh hadn't even 1187	noticed Jack. 2019 (80%)

Reading times for the first and second phrases of the target sentences in Experiment 2. The percentage of sentences judged to make sense is in parentheses after the second phrase reading times.

References

Brown, R. & Fish, D. (1983). The psychological causality implicit in language. Cognition, 14, 237-273.

Caramazza, A., Grober, E.H., Garvey, C., & Yates, J. (1977). Comprehension of anaphoric pronouns. Journal of Verbal Learning and Verbal Behavior, 16, 601-609.

Garvey, C., Caramazza, A., & Yates, J. (1975). Factors influencing assignment of pronoun antecedents. Cognition, 3, 243-277.

Grosz, B.T., Joshi, A.K., & Weinstein, S. (1983). Providing a unified account of definite noun phrases in discourse. Proceedings of the Association of Computational Linguistics, MIT, 44-50.

Grosz, B.T. & Sidner, C. (1985). Discourse structure and the proper treatment of interruptions. International Journal of the Conference of Artificial Intelligence, August, 832-838.

Newman, J.E. (1984). Finding the referent: Focus and the interpretation of anaphoric pronouns. Unpublished manuscript.

[1]The research was partially supported by NSF Grant BNS-8217378.

THE TIME COURSE OF PRONOUN COMPREHENSION

Rosemary J. Stevenson

Department of Psychology
University of Durham
England.

ABSTRACT

Two experiments investigated the time course of pronoun comprehension, and addressed the following questions. How soon, on reading a pronoun, is the retrieval of potential antecedents initiated? And at what point in the sentence is selection of an appropriate antecedent completed, reflecting successful comprehension?

Subjects were presented with sentences one phrase at a time on a computer screen. At certain points during the presentation of the phrases, a single test word appeared. Subjects were required to indicate, by pressing one of two keys, whether the test word had already appeared in the sentence being read. The test word appeared either before a phrase containing a pronoun or after a phrase containing a pronoun. In addition, the test word named either the antecedent of the pronoun or the non-antecedent. Response times to the test words indicated that unambiguous pronouns, which have a unique antecedent, are interpreted very quickly - at least by the time the verb following the pronoun has been read. Conversely, ambiguous pronouns, which require inferences from general knowledge, are still unresolved at the end of the phrase containing the pronoun. The results suggest that the search for antecedents is initiated very quickly, but that selection of a unique antecedent may continue during the reading of subsequent words.

INTRODUCTION

The assignment of pronouns to appropriate antecedents is essential for successful comprehension. Such assignments typically require inferences based on general knowledge (e.g. Ehrlich, 1980; Hirst & Brill, 1980). Hence in the following sentence:

> The mother punished her daughter because
> she stayed very late at the party.

The pronoun she is linguistically ambiguous. Linguistic rules can only state that, in this sentence, neither the mother nor the daughter can be ruled out as potential antecedents for the pronoun. To interpret the sentence appropriately, the reader has to make inferences from general knowledge about some possible reasons for punishment, to infer that she refers to the

daughter. Thus, despite the linguistic ambiguity of the sentence, the reader may use pragmatic inferences to derive a plausible and unambiguous interpretation of it. Of course, if the mother were replaced by the father (and her by him) in the sentence above, then the pronoun could be interpreted on linguistic grounds, without the need for pragmatic inferences. Indeed, it has been shown that linguistically ambiguous pronouns do take longer to comprehend than unambiguous pronouns (Ehrlich, 1980).

Despite the interest in the use of such inferences (e.g. Hirst & Brill, 1980), there has been little direct investigation of the time course for the retrieval and selection of an appropriate antecedent. The present experiments address this issue. In particular, they address the following questions: When a reader encounters a pronoun, how soon is retrieval of the antecedent initiated? And at what point in the sentence is a pronoun assigned to the antecedent, thus reflecting successful retrieval and interpretation?

These considerations raise the general question of the extent and kind of processing that may be carried out by a reader while a particular word is being read. Current research indicates that lexical retrieval of a word and some syntactic parsing occur while a word is being read (Frazier & Rayner, 1982; Swinney, 1979). However, in the case of pronouns, all that can be retrieved by reading the word itself is information about number and gender. To interpret a pronoun appropriately, a reader must access and integrate information from another portion of the text in order to select an antecedent that matches the pronoun in number and gender. If the pronoun is linguistically unambiguous, this will yield a unique antecedent. But if the pronoun is linguistically ambiguous, the selection of a single, appropriate antecedent can only be made by the use of inferences based on general knowledge. Thus, although the search for an antecedent may be initiated very quickly when a pronoun is encountered, the selection of a single, appropriate antecedent may not be completed until some point after the pronoun has been read. Further, we might expect that completion of assignment will take longer with ambiguous pronouns than with unambiguous pronouns.

Ehrlich & Rayner (1983) measured eye movements while sentences containing pronouns were being read. They found that with increasing distance between the pronoun and its antecedent there was an increasing delay in the time taken for pronoun comprehension to be completed. However, Ehrlich & Rayner only considered unambiguous pronouns, which may be interpreted without the need for pragmatic inferences. By contrast, Corbett & Chang (1983) investigated sentences containing ambiguous pronouns. They asked subjects to identify test words presented at the end of sentences containing linguistically ambiguous pronouns. Their results suggest that both the antecedent and the non-antecedent are activated at the end of the sentence. However, this result may be due to end-of-sentence integration effects, and not to a failure of successful interpretation of the pronoun.

The present experiments investigated the activation of antecedent information during the reading of the sentences containing either linguistically ambiguous pronouns or unambiguous pronouns. The basic procedure was as follows: Subjects were presented with a series of sentences.

Each sentence appeared on a computer screen one phrase at a time. In the following example each phrase is on a new line:

John apologised to Anne
at the end of the class (1)
because he regretted (2)
having caused so much trouble.

There were 56 experimental sentences. In 28 of them the first noun phrase was the antecedent of the pronoun. These were called NP1 sentences. (See the example above.) In the remaining 28 sentences, the second noun phrase was the antecedent of the pronoun. These were called NP2 sentences, for example:

Anne scolded John
three times during the class (1)
because he dropped (2)
several books on the floor.

The presentation of the phrases was self-paced: Subjects pressed the space bar of the computer keyboard when they had read and understood the current phrase. Depression of the space bar removed the current phrase and replaced it with the next phrase of the sentence. At some point during the presentation of the phrases, when the subject pressed the space bar, instead of the next phrase coming up immediately, the subject saw a single capitalised word surrounded by asterisks (e.g. ***JOHN***). When presented with this word, the subject had to press one of two keys to indicate whether or not the word had already occurred in the sentence currently being read. This test word was either the antecedent (John in the sentences above) or the non-antecedent (Anne). The time taken to respond to it was the dependent variable. In the experimental sentences, the test word appeared either after the second phrase (position number (1) in the examples above) or after the third phrase (position number (2) in the examples above). These positions correspond to the EARLY and LATE conditions respectively. Thus, there were four experimental conditions: The test word was either the antecedent or the non-antecedent; and the test word position was either early or late. There were 76 filler sentences, 66 of which required "no" responses to the test word, and all of which varied the position of the test word. To encourage comprehension of the sentences, 40% of them were followed by a yes/no question.

The test words in the late position occurred after the phrase containing the pronoun; hence the speed of recognition should indicate whether or not the pronoun's antecedent has been activated. The test words in the early position occurred before the pronoun, and hence served as controls against which to measure the effects of reading the pronoun.

One assumption underlying the use of this task is that the time taken to recognise the test words reflects their presence or absence in the reader's currently active working memory. (See, for example, Ratcliffe, Hockley & McKoon, 1985). A second assumption underlying the measurement of recognition

times is that the primary task for the reader is the interpretation of the pronoun (and hence the sentence). Thus, the effects of this primary task will also be reflected in the time taken to recognise the test words.

EXPERIMENT ONE: UNAMBIGUOUS PRONOUNS

In experiment one, unambiguous pronouns were used. With these sentences, we would expect that the antecedent would be retrieved relatively quickly after the pronoun has been read, and certainly by the end of the phrase containing the pronoun. Thus, when the test word is the antecedent, recognition times should be faster than when the test word is the non-antecedent, but only in the late condition. The results are shown in Table 1. Table 1 also shows the percentage of recognition errors in parentheses.

Analyses of variance on the recognition times indicated that, overall, test words in NP2 sentences were recognised faster than test words in NP1 sentences. $F1=4.45$, $df=1,31$, $p<.05$; $F2=3.39$, $df=1,54$, $p<.07$. The only other significant result was an interaction between type of test word (antecedent vs. non-antecedent) and position of test word (early vs. late). $F1=4.5$, $df=1,31$, $p<.05$; $F2=3.12$, $df=1,54$, $p<.08$. Observation of Table 1

TABLE 1: MEAN RECOGNITION TIMES (IN MSECS) AND RECOGNITION ACCURACY FOR ANTECEDENT AND NON-ANTECEDENT PROBE WORDS FOR THE TWO TYPES OF SENTENCE IN THE TWO PROBE POSITIONS (Unambiguous Pronouns)

	Early				Late			
	Antecedent		Non-Antecedent		Antecedent		Non-Antecedent	
	RT	PC	RT	PC	RT	PC	RT	PC
NP1 Sentences	1014	(6.7)	1020	(2.7)	1011	(8.5)	1052	(4.5)
NP2 Sentences	1000	(2.7)	982	(6.3)	978	(1.3)	1035	(8.0)
Means	1007	(4.7)	1001	(4.5)	994	(4.9)	1043	(6.2)

indicates that, with both types of sentence, recognition times to the antecedent test words are faster than recognition times to the non-antecedent test words, but only in the late position. Thus the results support the initial prediction. Unambiguous pronouns, which can be interpreted without the use of inferences, appear to be fully interpreted by the time the verb following the pronoun has been read. This seems to render the non-antecedent less accessible. Hence, the basic finding supports the view that the search for an antecedent is initiated very quickly. The selection of a unique antecedent is completed by the time one word after the pronoun has been read.

However, this observation was only made on unambiguous pronouns. Since ambiguous pronouns invariably require the use of inferences from general knowledge for the selection of a unique antecedent, we might expect that this process would be completed less quickly than with unambiguous pronouns. Experiment two investigated this issue.

Observation of the errors in Table 1 presents a slightly different picture. Most errors seem to occur with the antecedent test word in NP1 sentences and with the non-antecedent test word in NP2 sentences. (The first noun phrase in both cases). Analyses of variance indicated that this was, indeed, the case. The only significant effect was an interaction between type of sentence (NP1 vs. NP2) and type of test word (antecedent vs. non-antecedent). $F1=21.6$, $df=1,28$, $p<.01$; $F2=10.76$, $df=1,48$, $p<.01$. (Three subjects were excluded from the F1 analysis because they made no errors. Three sentences were excluded from the F2 analysis because they elicited no errors). Thus it appears that there are more errors when the first noun phrase is the test word. One interpretation of this finding is that there is a greater likelihood that the first noun phrase will have been forgotten by the time the test word appears.

EXPERIMENT TWO: LINGUISTICALLY AMBIGUOUS PRONOUNS

Experiment two used linguistically ambiguous pronouns. Thus, in the example sentences above, John was replaced by Joan, and the pronoun was changed to she. In prior pilot work, the first three phrases of each sentence were presented to five independent judges. In all the experimental sentences, the intended antecedents were unanimously selected as the referents of the pronouns by these judges. Thus, despite the linguistic ambiguity, there was complete concensus on the interpretation of each pronoun by the end of the phrase containing the pronoun. The pronouns were, therefore, "pragmatically" unambiguous. Nevertheless, with these sentences, the pronouns cannot be interpreted without the use of pragmatic inferences.

If the use of these inferences involves the retrieval of both potential antecedents which are checked for pragmatic plausibility, then we might expect that the activation of these two alternatives would interfere with the recognition of the test words (see, e.g. Corbett, 1984). The inferencing required to determine the interpretation of the pronoun is likely to interfere

TABLE 2: MEAN RECOGNITION TIMES AND RECOGNITION ERRORS FOR ANTECEDENT AND NON-ANTECEDENT PROBE WORDS FOR THE TWO TYPES OF SENTENCE IN THE TWO PROBE POSITIONS (Ambiguous Pronouns)

	Early				Late			
	Antecedent		Non-Antecedent		Antecedent		Non-Antecedent	
	RT	PC	RT	PC	RT	PC	RT	PC
NP1 Sentences	1001	(2.6)	988	(4.6)	1009	(9.2)	994	(5.1)
NP2 Sentences	926	(4.1)	973	(4.1)	990	(5.6)	1043	(6.1)
Means	963	(3.4)	980	(4.3)	999	(7.4)	1018	(5.6)

with the secondary task of identifying the test word. Hence recognition times should be slow in the late position (after the pronoun) relative to the early position (before the pronoun). Table 2 shows the results. Again, the percentage of recognition errors are presented in parentheses.

Analyses of variance on the data in Table 2 indicated that, as predicted, test words in the late position took longer to recognise than test words in the early position. $F1=16.84$, $df=1,27$, $p<.01$; $F2=5.56$, $df=1,54$, $p<.05$. Analyses of the error data also support the prediction. For both types of sentences, it was more difficult to identify the test word when it occurred in the late position as opposed to the early position: More errors of recognition occurred when the test word occurred late rather than early. $F1=14.1$, $df=1,22$, $p<.01$; $F2=7.5$, $df=1,52$, $p<.01$. (Five subjects were discarded from the F1 analysis because they made no errors. One sentence was excluded from the F2 analysis because it elicited no errors). An analysis on the number correct rather than the number of errors yielded the same significant finding. The effect of test word position on the recognition times appears larger in the NP2 sentences than in the NP1 sentences. However, the interaction between test word position and type of sentence is only significant in the subjects analysis ($F1=4.73$, $df=1,27$, $p<.05$), and not in the sentences analysis ($F2 < 1$).

The data in Table 2 also suggest an overall benefit in recognition times for antecedent test words relative to non-antecedent test words.

However, this difference was not reliable. F1=2.07, df=1,27; F2=1.11, df=1,54. It is possible, though, that such a benefit for antecedent test words is confined to the NP2 sentences. However, the interaction between type of sentence and type of test word was only marginally significant. F1=3.09, df=1,27, p<.09. F2=2.90, df=1,54, p<.09.

Thus the pattern of results supports the view that ambiguous pronouns are not comprehended as readily as unambiguous pronouns. This is the case even though the prior pilot studies indicated that the sentences were always given the intended interpretation. It thus appears that the inferencing required to select a unique antecedent continues while subsequent words in the sentence are being read.

DISCUSSION

Overall, the results of Experiment 1 suggest that the interpretation of unambiguous pronouns is completed quite rapidly, and this reflects the ease with which the antecedent can be retrieved. Thus, initiation of the search process must begin very quickly since, with these unambiguous pronouns, comprehension appears to be complete by the time the verb following the pronoun has been read. Conversely,the results of Experiment 2 suggest that the interpretation of ambiguous pronouns is still unresolved when the verb following the pronoun has been read. In this instance, therefore, it appears that interpretation of the pronoun continues while subsequent words are being read. However, there are two points that remain unresolved by these experiments. The first concerns the precise point in the text when the assignment of an unambiguous pronoun is completed. The present Experiment 1 only tested for the activation of antecedent information after the verb. It may well be the case, though, that assignment is completed as soon as an unambiguous pronoun (which requires a minimal search for an antecedent) is encountered. However, it is also possible that these unambiguous pronouns are comprehended by means of inferences from general knowledge, even though such inferences are logically unnecessary for the selection of a unique antecedent. (See, for example, Hirst & Brill, 1980; Stevenson & Vitkovitch, to appear). Hence, there may also be some delay in the completion of assignment for these unambiguous pronouns. The second point is that it remains unclear exactly how long it takes for ambiguous pronouns to be fully comprehended. All we can say from Experiment 2 is that the process is <u>not</u> completed when the verb following the pronoun has been read. Investigation of these two issues is currently in progress.

In general, these results support the proposition that processing is not completed immediately whenever the retrieval of relevant information is sufficiently complex (for example, when there is a need for inferences as in Experiment 2, or when a lengthy search process is required as in Ehrlich and Rayner, 1983). In these circumstances, we are likely to find cases where the processing continues while subsequent words are being read. More generally, it may well be the case that processing time increases whenever higher order integrative processes are involved; processes, for example, which are

necessary to construct a discourse model. There is no reason to suppose that such higher order processes, which are necessary for the full comprehension of a particular word, will be completed while that word itself is still being read.

REFERENCES

Corbett, A. T. (1984). Prenominal adjectives and the disambiguation of anaphoric pronouns. Journal of Verbal Learning and Verbal Behaviour, 23, 683-695.

Corbett, A. T. & Chang, F. R. (1983). Pronoun disambiguation: accessing potential antecedents. Memory and Cognition, 11, 283-294.

Ehrlich, K. (1980). Comprehension of pronouns. Quarterly Journal of Experimental Psychology, 32, 247-255.

Ehrlich, K. & Rayner, K. (1983). Pronoun assignment and semantic integration during reading. Journal of Verbal Learning and Verbal Behaviour, 22, 75-87.

Frazier, L. & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. Cognitive Psychology, 14, 178-210.

Hirst, W. & Brill, G. A. (1980). Contextual aspects of pronoun assignment. Journal of Verbal Learning and Verbal Behaviour, 19, 168-175.

Ratcliffe, R., Hockley, W. & McKoon, G. (1985). Components of activation: repetition and priming effects in lexical decision and recognition. Journal of Experimental Psychology: General, 114, 435- 450.

Stevenson, R.J. & Vitkovitch, M. (to appear). The comprehension of anaphoric relations. Language and Speech.

Swinney, D.A. (1979). Lexical access during sentence comprehension: (re)consideration of context effects. Journal of Verbal Learning and Verbal Behaviour, 18, 645-659.

THE COMPREHENSION OF CONCEPTUAL ANAPHORA IN DISCOURSE

Morton Ann Gernsbacher
University of Oregon

Abstract

A primary constraint on using a pronominal anaphor is that it must agree with its antecedent in number. However, there are situations in which pronouns act as conceptual anaphors. For example, in the discourse, "I think I'll order a frozen margarita. I just love them.", the pronoun "them" does not refer to a single margarita, but perhaps all the margaritas the speaker has ever tasted. When anaphors operate in this way, they are often mismatched with their literal antecedent in number. Three situtations when conceptual anaphora occurs are identified: when referring to the members of a Collective Set (as opposed to the set per se), a Multiply occurring Item or Event (versus a Unique Item/Event), or a Generic Type (versus a Specific Token). Two experiments are reported. The first demonstrated that subjects consider a mismatched, plural pronoun more natural than a matched, singular pronoun when it follows a Collective Set, Multiple Item/Event, or Generic Type noun. Conversely, subjects consider a matched, singular pronoun more natural when it follows an Individual Member of a set, Unique Item/Event, or Specific Token noun. The second experiment demonstrated that subjects comprehend a mismatched, plural pronoun faster than a matched, singular pronoun when it follows a Collective Set, Multiple Item/Event, or Generic Type noun, but they comprehend a matched, singular pronoun faster when it follows an Individual Member, Unique Item/Event, or Specific Token noun. This suggests that when comprehenders encounter conceptual--though mismatched anaphors--they do not have to reinstate the multiple entities into their mental representations.

A convenient feature of language is that it provides mechanisms for referring back to people or things previously mentioned. One such mechanism is anaphora. Over the past few years, many cognitive psychologists have been interested in understanding how comprehenders resolve discourse anaphora.[1] That is, how do comprehenders access from their mental representations the correct referent for an anaphoric expression? This question is also of interest to Artificial Intelligence specialists, particularly those working on Natural Language Processing (NLP) systems.

Anaphora resolution in many NLP systems is accomplished via certain heuristics, presumably the same heuristics employed by human comprehenders. Tyler and Marslen-Wilson (1981) have identified four possible constraints that guide this heuristic process. They are (a) Lexical Constraints, cued by lexical markings such as number, gender, and case, (b) Syntactic Constraints, (c) Thematic Constraints, cued by discourse markings such as topic, focus, or foregrounding, and (d) Pragmatic Constraints, provided by the comprehender's knowledge and inferential reasoning about the real world.

Heuristics which follow lexical constraints--number, gender, and case--are most easily incorporated into NLP systems. They are also the heuristics which human comprehenders acquire earliest (Palermo & Molfese, 1972) and which novice writers are most successful at applying (Bartlett, 1984).

This paper focuses on a particular use of pronominal anaphora, a use that one might assume would cause difficulty for comprehenders. At least, it is known that this type of anaphoric expression creates problems for virtually all extant

NLP systems (cf. Webber, 1984). The reason is that this type of construction clearly violates one of the most elementary, lexical constraints. An example, given by Sidner (1984), is the following:

(1a) My neighbor rides a monster Harley 1200.
(1b) They are really huge but gas-efficient bikes.

In this discourse, there is a blatant mismatch between the number of the pronoun and its supposed antecedent. The anaphor in (1b) clearly requires a plural antecedent; yet there are only singular nouns available in (1a). However, such mismatches occur rather frequently. Consider the following utterances overheard in a bar:

(2a) I think I'll order a frozen margarita.
(2b) I just love them.

Or the following comments overhead on a university campus:

(3a) My roommate was so excited. She actually made an A.
(3b) She doesn't make them very often.

Or the following exchange between the author (A) and a friend (F):

(4a) F: I can't believe you have a Fiat.
(4b) A: Why is that?
(4c) F: They're so temperamental.

Or the following statements the author made a few days after the exchange in (4a-c).

(5a) I need to call the garage [where her car was being serviced].
(5b) They said they'd have it ready by five o'clock, but I'm sure they won't.

In each of these instances, the mismatch occurs because the pronominal anaphor is not intended to map literally onto a preceding noun; rather these anaphors are intended to refer in a more conceptual manner. The speaker in (2) was not proclaiming her affection for one specific frozen margarita; rather she appeared to be proclaiming affection for all frozen margaritas in the universe (or at least those the speaker had tasted). Similarly, the author's friend in (4) was not diagnosing the personality of the specific token of Fiats that the author owns, but the generic type of automobile. And when the author stated that she needed to call the garage, she was not literally referring to a physical structure or place of business, but the mechanics who work there.

Such cases of conceptual anaphora can be simply classified--albeit roughly--as occurring in at least three situations. In example (5), the literal antecedent is a collective noun, a noun that refers to a collection or set of individuals. The mismatched plural pronoun is intended to refer to the individual members of the collection rather than the set per se. Conceptual anaphors are used frequently to refer to the individual members of what are traditionally considered Collective Sets (e.g., team, group, musical band) as in the examples below:

(6a) The substitute teacher begged the class to stop misbehaving.

(6b) But they didn't pay any attention to her.

Conceptual anaphors are also used to refer to the members of less traditional Collective Sets, for example:

(7a) After college, my sister went to work for IBM.
(7b) They made her a very good offer.

(8a) You wouldn't believe how bad it is to work for the city of Eugene.
(8b) They can never tell you whether your job will be covered in the next month's budget.

(9a) I need to call Sears.
(9b) They made a mistake on my last credit card bill.

A second situation when conceptual anaphora is used is when referring to things one is likely to have multiples of, or events one is likely to experience repeatedly, for example:

(10a) I need a plate.
(10b) Where do you keep them?

(11a) Yesterday was my birthday.
(11b) I used to really dread them, but yesterday I didn't care.

(12a) I just spilled something. Would you go get me a paper towel?
(12b) They're in the kitchen.

In the above examples, the literal antecedent is a sole item or event; however, because most households posses more than one plate (and presumably keep those plates together), most people have more than one birthday, and paper towels are usually dispensed in a roll of many, the intended reference is to these Multiple Items or Events. Thus, a conceptual anaphor, resulting in a mismatched pronoun, is used.

A third situation arises when conceptual anaphors are used to refer to Generic Types as in the following:

(13a) My mother's always bugging me to wear a dress.
(13b) She thinks I look good in them, but I don't.

(14a) Carla is downstairs watching a soap opera.
(14b) If she had her way, she'd watch them all afternoon.

(15a) I enjoy having a pet.
(15b) They are such good companions.

In this situation, the mismatched plural pronoun is intended to refer to a concept in general. For instance, it is soap operas in general, rather than the specific one Carla is currently watching, that the speaker in (16) believes Carla could watch all afternoon.[2]

The present classification scheme is not presented as a formal distinction. It is possible that the boundaries between these three situations are actually fuzzier or that stricter boundaries are needed. However, what is common among

these sentence pairs is that the pronoun in the second sentence refers to something more than what is explicitly mentioned in the first sentence. In other words, these pronouns are operating as conceptual (or implicit) anaphors as opposed to literal (or explicit) anaphors.

On the other hand, there are situations when a literal mapping between an antecedent and its anaphor is intended. For example, this occurs when an Individual Member of a collective set is singled out. In this situation, a matched, singular pronoun is used, as in the examples below:

(6c) The substitute teacher begged the student to stop misbehaving.
(6d) But he didn't pay any attention to her.

(7c) After college, my sister went to work for the vice president of IBM.
(7d) He made her a very good offer.

(8c) You wouldn't believe how bad it is to work for the mayor of Eugene.
(8d) He can never tell you whether your job will be covered in the next month's budget.

Similarly, there are situations in which a literal mapping between an anaphor and its antecedent is intended because the item or event being referred to is Unique (i.e., one is likely to have only one of such an item, or experience such an event only once). In this situation, a matched, singular pronoun is used. Compare, for example, the following three sentence pairs with (10a&b), (11a&b), and (12a&b), respectively:

(10c) I need an iron.
(10d) Where do you keep it?

(11c) Yesterday was my fortieth birthday.
(11d) I used to really dread it, but yesterday I didn't care.

(12c) I just spilled something. Would you go get me a mop?
(12d) It's in the kitchen.

Finally, there are situations when a literal mapping between an antecedent and its anaphor is intended because the preceding, coreferential noun has been identified so distinctly that it represents a Specific Token of a class of items, for example:

(13c) My mother's always bugging me to wear a dress that she bought me last year for Christmas.
(13d) She thinks I look good in it but I don't.

(14c) Carla is downstairs watching a soap opera that stars Michael Lewis.
(14d) If she had her way, she'd watch it all afternoon.

(15c) I enjoy having a pet canary named "Chatty".
(15d) She is such a good companion.

The present research was undertaken to answer two major questions about the comprehension of conceptual anaphora. The first question was this: How natural do comprehenders find references to conceptual antecedents via mismatched pronouns? That is, are comprehenders disturbed by these mismatches? Or do they

find them comprehensible <u>because</u> the antecedent noun represents a Collective Set, Multiple Item/Event, or Generic Type? If so, then presumably comprehenders would find mismatched pronouns less natural when the antecedent noun represents an Individual Member, Unique Item/Event, or a Specific Type. To empirically investigate this question, an experimental approach was taken in which the same sentence was presented in one of four different conditions.

Method

Sixteen sets of four sentence pairs were constructed for each of the three discourse situations when conceptual vs literal anaphora is used (i.e., reference to Collective Sets vs Individual Members, Multiple Events/Items vs Unique Events/Items, and Generic Types vs Specific Tokens). Two of the four sentence pairs were formed by preceding a sentence containing either a Plural or a Singular pronoun by a sentence with a Collective Set, Multiple Event/Item, or Generic Type noun. The other two sentence pairs were formed by preceding either a Plural or a Singular pronoun by a sentence containing an Individual Member of a Collective Set, a Unique Event/Item, or a Specific Token noun. An example set of four sentence pairs of each situation is shown in Table 1.

Table 1

The substitute teacher begged the class to stop misbehaving.	Collective Noun
But they didn't pay any attention to her.	Plural Pronoun
The substitute teacher begged the class to stop misbehaving.	Collective Noun
But it didn't pay any attention to her.	Singular Pronoun
The substitute teacher begged the student to stop misbehaving.	Individual Noun
But they didn't pay any attention to her.	Plural Pronoun
The substitute teacher begged the student to stop misbehaving.	Individual Noun
But he didn't pay any attention to her.	Singular Pronoun
I need a plate.	Multiple Noun
Where do you keep them?	Plural Pronoun
I need a plate.	Multiple Noun
Where do you keep it?	Singular Pronoun
I need a iron.	Unique Noun
Where do you keep them?	Plural Pronoun
I need an iron.	Unique Noun
Where do you keep it?	Singular Pronoun
My mother's always bugging me to wear a dress.	Generic Type
She thinks I look good in them but I don't.	Plural Pronoun

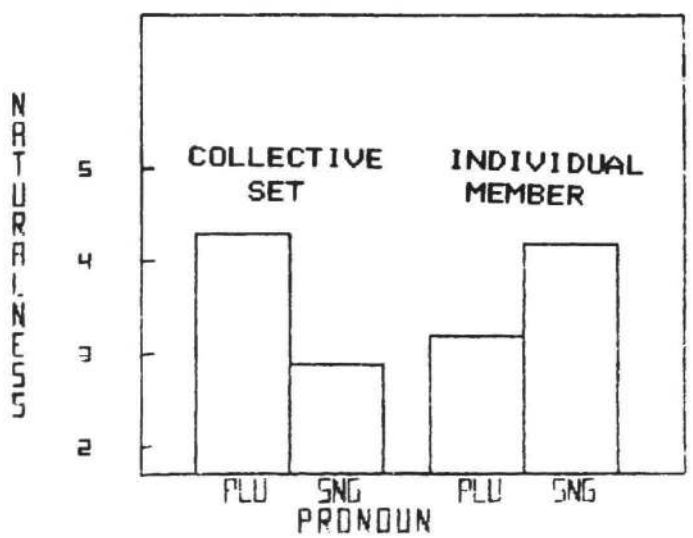
My mother's always bugging me to wear a dress. She thinks I look good in it but I don't.	Generic Type Singular Pronoun
My mother's always bugging me to wear a dress that she bought me last year for Christmas. She thinks I look good in them but I don't.	Specific Token Plural Pronoun
My mother's always bugging me to wear a dress that she bought me last year for Christmas. She thinks I look good in it but I don't.	Specific Token Singular Pronoun

These sentences were presented to 65 college-aged subjects. To minimize the subjects' exposure to similar sentences, each subject was presented with only two members of each set of four sentence pairs: one of the 2 sentence pairs with a Collective Set, Generic Type, or Multiple Noun and one of the 2 sentence pairs with an Individual Member, Specific Token, or Unique Noun. Thus, each subject was presented with 96 of the 192 sentence pairs. The subjects' task was to read each sentence pair and to rate "how natural" the second sentence seemed in reference to the first. The meaning of "natural," the subjects were told, was "how likely it is that you might hear such a sentence or produce such a sentence." To indicate their ratings, subjects used a 5-point scale where 5 meant "Very natural" and 1 meant "Not very natural."

Results

Collective Sets vs Individual Members. The mean ratings for the sentences following sentences with Collective Set vs Individual Member nouns are shown in Figure 1. The two bars on the left represent the mean ratings of the sentences when they contained either Plural or Singular pronouns, respectively, and they followed sentences with Collective Set nouns. The two bars on the right represent the mean ratings of the sentences when they contained either Plural or Singular pronouns, respectively, and they followed sentences with Individual Member nouns. An analysis of variance (ANOVA) revealed neither a main effect of pronoun number (Plural vs Singular) nor one of preceding noun (Collective vs Individual) [both *ps* > .4]. There was, however, a significant interaction between these two variables [$\underline{\text{minF}}'(1,24) = 52.77$].[3]

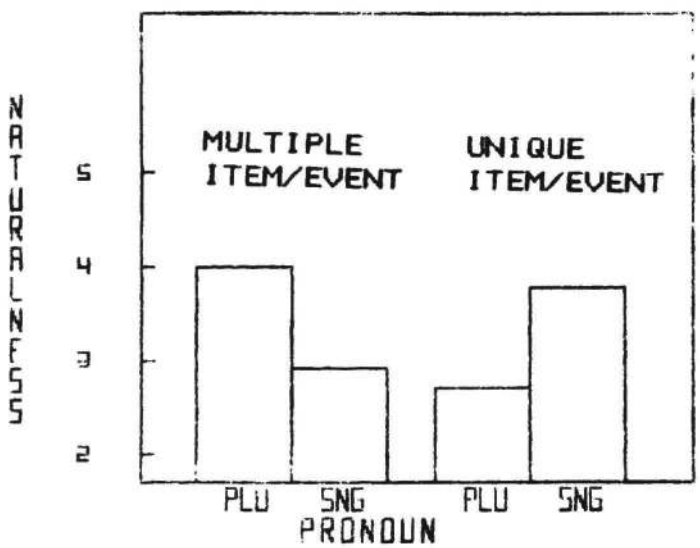
Figure 1



Additional planned comparisons revealed the following: When the sentences followed sentences with Collective nouns, they were rated significantly more natural when they contained Plural than Singular pronouns [minF'(1,22) = 37.46]. In contrast, when the sentences followed sentences with Individual nouns, they were rated significantly more natural when they contained Singular than Plural pronouns [minF'(1,26) = 21.63]. In addition, when the sentences contained Plural pronouns, they were rated considerably more natural when they followed sentences with Collective than Individual nouns [minF'(1,24) = 28.55]. In contrast, when the sentences contained Singular pronouns, they were rated considerably more natural when they followed sentences with Individual than Collective nouns [minF'(1,21) = 27.20].

Multiple Items/Events vs Unique Items/Events. The mean ratings for the sentences containing Plural vs Singular pronouns following sentences with Multiple vs Unique nouns are shown in Figure 2. The two bars on the left represent the mean ratings of the sentences when they contained either Plural or Singular pronouns, respectively, and they followed sentences with Multiple Items/Events nouns. The two bars on the right represent the mean ratings of the sentences when they contained either Plural or Singular pronouns, respectively, and they followed sentences with Unique Items/Events nouns. An ANOVA again revealed no main effect of pronoun number [minF' <1.0], although there was a marginally significant effect of preceding noun: Sentences following Multiple nouns were rated slightly more natural (M = 3.45) than sentences following Unique nouns (M = 3.23) [minF'(1,24) = 3.97; p < .07]. More interestingly, there was a significant interaction between these two variables [minF'(1,23) = 44.51].

Figure 2

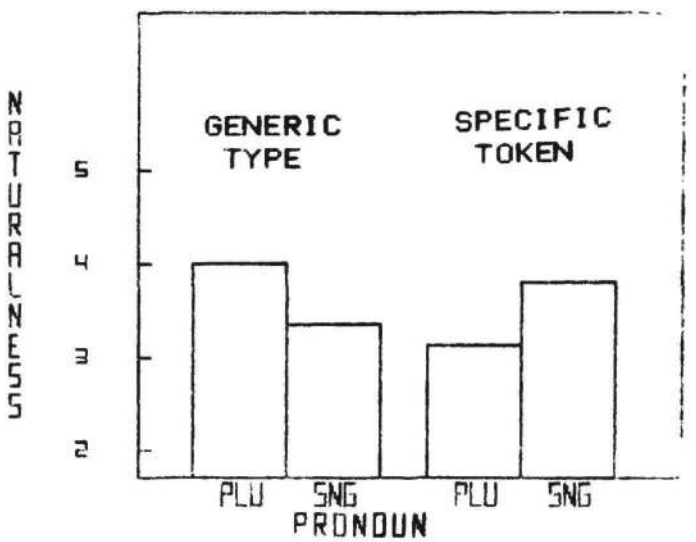


Again, planned comparisons revealed the following pattern: When the sentences followed sentences with the Multiple nouns, they were rated significantly more natural when they contained Plural pronouns [minF'(1,20) = 24.33]. In contrast, when the sentences followed Unique nouns, they were rated significantly more natural when they contained Singular pronouns [minF'(1,36) = 48.50]. In addition, when the sentences contained Plural pronouns, they were rated more natural when they followed sentences with Multiple nouns [minF'(1,32) = 60.16]. In contrast, when the sentences contained Singular pronouns, they were

rated considerably more natural when they followed Unique nouns [minF'(1,20) = 15.23].

Generic Types vs Specific Tokens. The mean ratings for the sentences containing Plural vs Singular pronouns following sentences with Generic Type vs Specific Token nouns are shown in Figure 3. An ANOVA again revealed no main effect of pronoun number (Plural vs Singular) or preceding noun (Generic Type vs Specific Token) [both ps > .4], only a significant intereaction between these two variables [minF'(1,20) = 16.80].

Figure 3



Additional planned comparisons revealed a familiar pattern: When the sentences followed sentences with Generic Type nouns, they were rated significantly more natural when they contained Plural pronouns [minF'(1,20) = 9.318]. In contrast, when the sentences followed sentences with Specific Token nouns, they were rated significantly more natural when they contained Singular pronouns [minF'(1,24) = 13.87]. In addition, when the sentences contained Plural pronouns, they were rated considerably more natural when they followed Generic Type nouns [minF'(1,19) = 13.07]. In contrast, when the sentences contained Singular pronouns, they were rated considerably more natural when they followed Specific Type nouns [minF'(1,19) = 5.410].

In summary, these results suggest strongly that comprehenders find references to conceptual antecedents via mismatched pronouns very natural. In fact, they find the use of a mismatched pronoun more natural than a matched pronoun. Yet it is because the preceding noun represents a Collective Set, Multiple Item/Event, or Generic Type that subjects find these mismatches acceptable. That is, they find mismatched pronouns considerably less natural when the preceding noun represents an Individual Member, Unique Item/Event, or a Specific Type.

The second question motivating this research was how difficult is it to map conceptual vs literal anaphors onto their intended antecedents? One prediction is that it is always difficult to map a plural pronoun onto a singular noun because on encountering a singular noun, only a single entity is established in the comprehender's mental representation of the discourse (e.g., the "discourse model" of Webber 1984, "discourse file" of Givon, 1979, or "mental model" of

Johnson-Laird, 1983). According to this prediction, when one subsequently encounters a mismatched pronoun, additional entities have to be reinstated. An opposite prediction is that it is only difficult to map a mismatched pronoun when it is used as a literal anaphor (i.e., it refers to an Individual Member, Unique Item/Event, or a Specific Token noun). But when a mismatched pronoun is used as a conceptual anaphor (i.e., it refers to a Collective Set, Multiple Item/Event, or Generic Type noun), it is no more difficult to map than mapping a matched pronoun to a literal anaphor. This would be the case if on encountering Collective Set, Multiple Item/Event or Generic Type nouns, comprehenders automatically incorporate multiple entities into their mental representations, though when encountering Individual Member, Unique Item/Event, or Specific Token nouns, only a single entity is instantiated.

An experimental paradigm frequently used to investigate anaphoric mapping is to measure the amount of time required for a comprehender to read a sentence containing the anaphoric reference (Clark & Sengul, 1979; Garrod & Sanford, 1977; Garrod & Sanford, 1983; Garnham, 1980, Garnham, 1984; Haviland & Clark, 1974; Malt, 1985; Sanford & Garrod, 1981; Yekovitch & Walker, 1978; Yekovitch, Walker, & Blackman 1979). Presumably, the more time required to read the sentence, the more difficult the mapping process. If on encountering a plural pronoun that refers to a Collective Set, Multiple Item/Event, or Generic type, comprehenders have to reinstate these multiple entities, then their reading time for these sentences should be longer than when a singular pronoun is used. On the other hand, if comprehenders automatically incorporate multiple entities into their mental representations, then their reading times should be shorter when a plural as opposed to singular pronoun is used. And the opposite would be true of situtations employing literal anaphora (i.e., references to an Individual Member, Unique Item/Event, or Specific Token).

Method

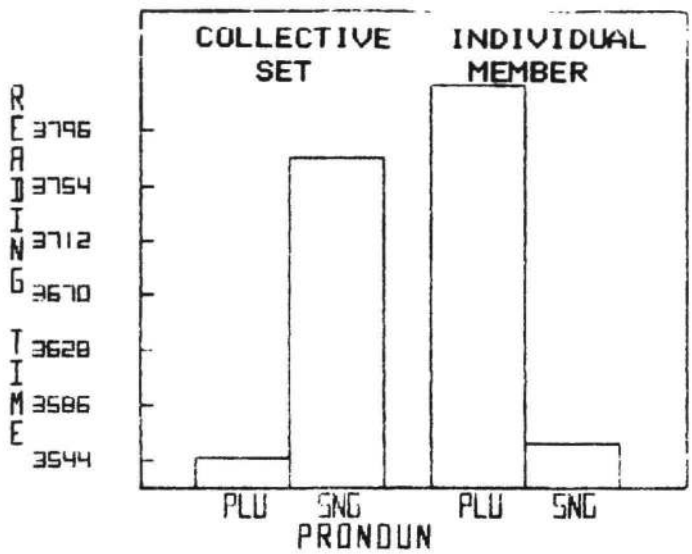
The same materials were used as in the first experiment. Each of 72 subjects was tested individually. The subject was seated in front of a video display monitor. For each pair of sentences, the first sentence appeared toward the top of the video display screen and remained there for a period of time proportionate to the number of characters it contained. After this first sentence disappeared, the second sentence appeared toward the bottom of the screen. This second sentence remained visible until the subject pressed a key to indicate that he/she was finished reading the sentence. Immediately after the second sentence of the pair disappeared, the word "Paraphrase" appeared on the screen. At this point the subject "retold [aloud] the sentence in his/her own words." The paraphrase task was included to ensure that subjects would read the sentence pairs for comprehension, rather than simply pressing the key as rapidly as possible. Each subject's paraphrases were recorded on audio tape.

Results

Collective Sets vs Individual Members. The mean reading times for the sentences containing Plural vs Singular pronouns following sentences with Collective vs Individual nouns are shown in Figure 4. Note that the interpretation of the direction of the reading time figures should be opposite that of the naturalness rating figures: A shorter bar on the graph represents a faster reading time, which is interpreted as greater ease in comprehending the sentence.[4] An ANOVA revealed no main effect of pronoun number (Plural vs

Singular) or of preceding noun (Collective vs Individual) [both minF's < 1.0], only a significant interaction between the two [minF'(1,33) = 8.76].

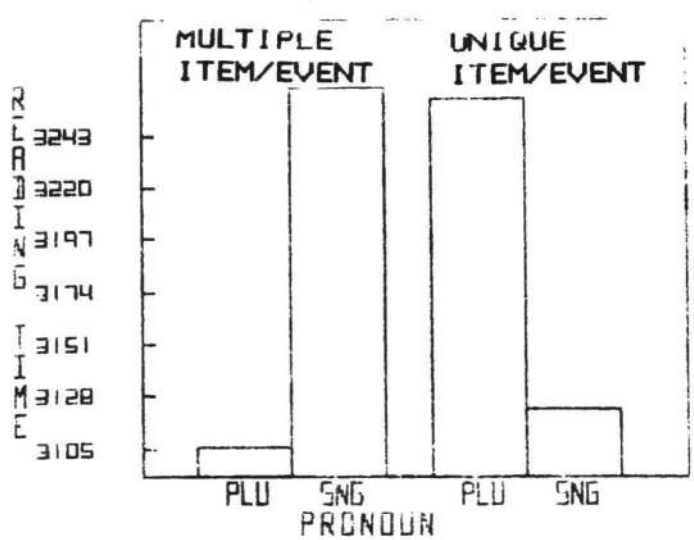
Figure 4



Additional planned comparisons revealed the following: When the sentences followed sentences with Collective nouns, they were read faster when they contained Plural pronouns [minF'(1,28) = 4.008]. In contrast, when the sentences followed sentences with Individual nouns, they were read significantly faster when they contained Singular pronouns [minF'(1,33) = 4.277]. In addition, when the sentences contained Plural pronouns, they were read considerably faster when they followed Collective nouns [minF'(1,29) = 4.961]. In contrast, when the sentences contained Singular pronouns, they were read faster when they followed Individual nouns [$F_1(1,68) = 10.38$; $F_2(1,15) = 5.99$; minF'(1,33) = 3.637, $p < .07$].

Multiple vs Unique Items/Events. The mean reading times for the sentences containing Plural vs Singular pronouns following sentences with Multiple vs Unique nouns are shown in Figure 5. An ANOVA again revealed no main effect of pronoun number (Plural vs Singular) or preceding noun (Multiple vs Unique) [both minF's < 1.0], only a significant interaction between the two [minF'(1,59) = 7.203].

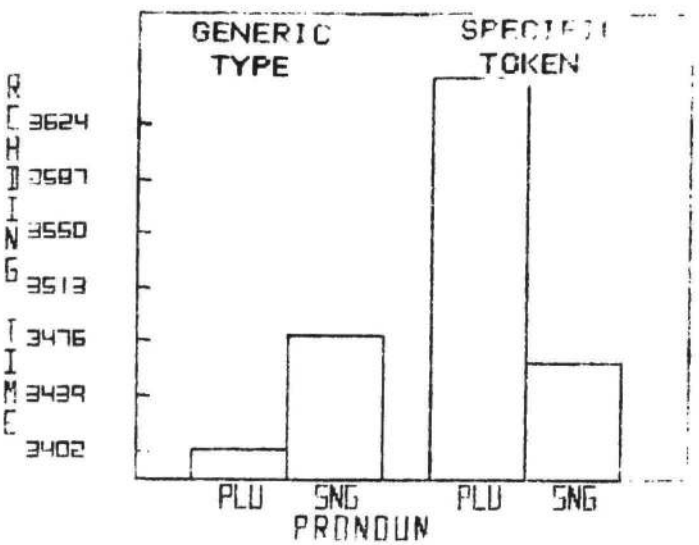
Figure 5



Again, planned comparisons revealed the following pattern: When the sentences followed sentences with the Multiple nouns, they were read significantly faster when they contained Plural pronouns [$F_1(1,68) = 5.903$; $F_2(1,15) = 6.578$]. In contrast, when the sentences followed Unique nouns, they were read significantly faster when they contained Singular pronouns [$F_1(1,68) = 4.559$; $F_2(1,15) = 7.689$]. In addition, when the sentences contained Plural pronouns, they were read faster when they followed sentences with Multiple nouns [$F_1(1,68) = 5.114$; $F_2(1,15) = 7.272$]. In contrast, when the sentences contained Singular pronouns, they were read considerably faster when they followed Unique nouns [$F_1(1,68) = 6.098$; $F_2(1,15) = 4.106$].

Generic Types vs Specific Tokens. The mean reading times for the sentences containing Plural vs Singular pronouns following sentences with Generic Type vs Specific Token nouns are shown in Figure 6. Again the main effects of pronoun number (Plural vs Singular) and preceding noun (Generic Type vs Specific Token) [both $ps > .40$] were not significant, but the interaction between the two was [$minF'(1,35) = 4.223$].

Figure 6



Additional planned comparisons revealed that the interaction differed from the pattern observed in the reading time data for the other two conceptual-vs-literal anaphora situations, as well as the naturalness ratings for this situation: When the sentences followed sentences with Generic Type nouns, they were read just as fast when they contained Plural as Singular pronouns; that is, there was no significant difference between their mean reading times [$minF' < 1.0$]. In contrast, when the sentences followed sentences with Specific Token nouns, they were read significantly faster when they contained Singular pronouns [$F_1(1,68) = 8.776$; $F_2(1,15) = 5.864$]. In addition, when the sentences contained Plural pronouns, they were read significantly faster when they followed Generic Type nouns [$minF'(1,26) = 4.819$]. However, when the sentences contained Singular pronouns, they were read just as fast when they followed Specific Token as Generic Type nouns [$minF' < 1.0$]. In other words, subjects comprehended the Generic Type-Singular sentences at the same rate as they comprehended the Generic Type-Plural or the Specific Token-Singular sentences.

It is curious why subjects had little difficulty when the singular pronouns referred to Generic Type nouns. Because this was the only effect in the reading time data that failed to mirror what was observed in the naturalness ratings, it was examined further. An additional source of data which could illuminate how subjects interpreted these Generic Type-Singular pronoun sentences was their paraphrases. These paraphrases indicated that for several of the sentences in this condition, instead of interpreting the singular pronoun as an awkward coreferent subjects interpreted it as a dummy subject. Their resulting paraphrases were formed via extraposition of a participial clause and It-Insertion. For example, a majority of the subjects paraphrased the following Generic Type-Singular pronoun sentence pair

(16a) My neighbor rides a moped.
(16b) I think it's dangerous.

as

(16c) I think it's dangerous to ride mopeds.

This was in contrast to the same sentence pair presented with a plural pronoun

(16d) My neighbor rides a moped.
(16e) I think they're dangerous.

for which the modal paraphrase was

(16f) Mopeds are dangerous.

It was also in contrast to the same sentence pair presented with a singular pronoun, and preceded by the Specific Token noun

(16g) My neighbor rides a moped that doesn't even have a light.
(16h) I think it's dangerous.

for which the modal paraphrase was

(16i) My neighbor's moped is dangerous because it doesn't have a light.

Thus, subjects had little difficulty mapping a singular pronoun onto a Generic Type because they simply did not attempt to. Rather, they handled this awkward coreference by interpreting it as a different construction.

In summary, it does not appear that on encountering a conceptual anaphor comprehenders have to reinstate multiple entitites into their mental representation. However, it is possible that such mismatched pronouns do cause momentary processing difficulties but these momentary difficulties are quickly resolved when integrating the two sentences. Perhaps the reading time paradigm used in this experiment only demonstrates integration processes, and other on-line measures (e.g., Gernsbacher, 1986) would better demonstrate immediate mapping processes and any incurred mapping difficulties.

Conclusions

These data suggest that not only are conceptual anaphors considered natural but they are relatively easily comprehended. Although currently problematic for most NLP systems, the use of conceptual anaphora for human comprehenders is a convenience. Unlike other types of anahors, they provide more than verbal

shorthand. That is, they operate beyond simply saving a speaker's breath of a writer's pens: They allow extension. Although the function of coreference has been suggested throughout this paper, these anaphors more closely resemble the function of cospecification suggested by Sidner (1984). (Indeed, there are some who might suggest that these situations are not cases of anaphora at all; yet, I am in agreement with Stenning (1978), that these situtations are a bonafide use of anaphora.)

This research provides only a demonstration, not an explication. Several questions remain about the use of conceptual anaphora. For example, what are the boundary conditions for interpreting nouns as Multiple Item or Events? What features of the following sentence

(17a) I just washed a plate.

identifies uniqueness so that (17b) is an appropriate sequiter?

(17b) Where should I put it?

How does the knowledge that the speaker is in a store convert the following reference to a Unique Item/Event into a reference to a Multiple Item/Event?

(18a) I need an iron.
(18b) Where aisle are they on?

These and other questions deserve further investigation.

References

Anderson, A., Garrod, S.C., & Sanford, A.J. (1983). The accessibility of pronominal antecedents as a function of episode shifts in narrative text. Quarterly Journal of Experimental Psychology, 36A, 1-12.

Bartlett, E.J. (1984). Anaphoric reference in written narratives of god and poor elementary school writers. Journal of Verbal Learning and Verbal Behavior, 123, 540-552.

Caramazza, A., Grober, E., Garvey, C., & Yates, J. (1977). Comprehension of anaphoric pronouns. Journal of Verbal Learning and Verbal Behavior, 16, 601-609.

Clark, H.H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. Journal of Verbal Learning and Verbal Behavior, 12, 335-359.

Clark, H.H., & Sengul, C.J. (1979). In search of referents for nouns and pronouns. Memory & Cognition, 7, 35-41.

Corbett, A.T. (1984). Prenominal adjectives and the disambiguation of anaphoric nouns. Journal of Verbal Learning and Verbal Behavior, 23, 683-695.

Corbett, A.T., & Chang, F.R. (1983). Pronoun disambiguation: Accessing potential antecedents. Memory & Cognition, 11, 283-294.

Dell, G.S., McKoon, G., & Ratcliff, R. (1983). The activation of antecedent information during the processing of anaphoric reference in reading. Journal of Verbal Learning and Verbal Behavior, 22, 121-132.

van Dijk, T.A., & Kintsch, W. (1983). Strategies of discourse comprehension. New York: Academic Press.

Erhlich, K. (1980). Comprehension of pronouns. Quarterly Journal of Experimental Psychology, 32, 247-255.

Garrod, S., & Sanford, A. (1977). Interpreting anaphoric relations: The integration of semantic information while reading. Journal of Verbal Learning and Verbal Behavior, 16, 77-90.

Garrod, S., & Sanford, A. (1983). Topic dependent effects in language processing. In G.B. Flores d'Arcais, & R.J. Jarvella (Ed.), The process of language understanding (pp. 271-296). New York: Wiley.

Garnham, A. (1981). Anaphoric reference to instances, instantatied and non-instantiated categories: A reading time study. British Journal of Psychology, 72, 377-384.

Garnham, A. (1984). Effects of specificity on the interpretation of anaphoric noun phrases. Quarterly Journal of Experimental Psychology, 36A, 1-12.

Gernsbacher, M.A. (1986). Mechanisms of referential accessibility in discourse. Manuscript in preparation.

Givon, T. (1979). On understanding grammar. New York: Academic Press.

Haviland, S.E., & Clark, H.H. (1974). What's new? Acquiring new information as a process in comprehension. Journal of Verbal Learning and Verbal Behavior, 13, 512-521.

Hirst, W., & Brill, G.A. (1980). Contextual aspects of pronoun assignment. Journal of Verbal Learning and Verbal Behavior, 19, 168-175.

Johnson-Laird, P. (1983). Mental models. Cambridge, MA: Harvard University Press.

Malt, B.C. (1985). The role of discourse structure in understanding anaphora. Journal of Memory and Language, 24, 271-289.

McKoon, G., & Ratcliff, R. (1980). The comprehension processes and memory structures involved in anaphoric reference. Journal of Verbal Learning and Verbal Behavior, 20, 204-215.

Palermo, D., & Molfese. D. (1972). Language acquisition from age five onward. Psychological Bulletin, 78, 409-428.

Prince, E.F. (1981). On the inferencing of thje indefinite-this NPs. In A. Joshi, B. Webber, & I. Sag (Eds.), Elements of discourse understanding (pp. 231-250). Cambridge: Cambridge University Press.

Sanford, A.J., & Garrod, S. (1981). Memory and attention in text comprehension: The problem of reference. In R. Nickerson (Ed.), Attention and performance VII (pp. 459-474). Hillsdale, NJ: Erlbaum.

Shillcock, R. (1982). The on-line resolution of pronominal anaphora. Language and Speech, 25, 385-401.

Sidner, C.L. (1984). Focusing in the comprehension of definite anaphora. In M. Brady, & R.C. Berwick (Eds.), Computational models of discourse (pp. 267-330). Cambridge, MA: MIT Press.

Stenning, K. (1978). Anaphora as an approach to pragmatics. In M. Halle, J. Bresnan, & G.A. Miller (Eds.), Linguistic theorry and psychological reality (pp. 162-200). Cambridge, MA: MIT Press.

Tyler, K., & Marlsen-Wilson, W. (1982). The resolution of discourse anaphors: Some on-line studies. Text, 2, 263-291.

Webber, B.L. (1984). So what can we talk about now? In M. Brady, & R.C. Berwick (Eds.), Computational models of discourse (pp. 331-371). Cambridge, MA: MIT Press.

Yekovich, F.R., & Walker, C.H. (1978). Identifying and using referents in sentence comprehension. Journal of Verbal Learning and Verbal Behavior, 17, 265-277.

Yekovich, F.R., Walker, C.H., & Blackman, H.S. (1979). The role of presupposed and focal information in integrating sentences. Journal of Verbal Learning and Verbal Behavior, 18, 535-548.

Footnotes

Acknowledgements: This research was supported by National Science Foundation grant BNS 85-10096. The author thanks Ellen Galloway for help in constructing the experimental sentences.

1. See, for example, Anderson, Garrod, & Sanford (1983); Caramazza, Grober, Garvey, & Yates (1977); Clark & Sengul (1978); Corbett (1984); Corbett & Chang (1983); Dell, McKoon, & Ratcliff (1983); van Dijk & Kintsch (1983, Ch. 5); Erhlich (1980); Garrod & Sanford (1977, 1983); Garnham (1981, 1984); Hirst & Brill (1980); Malt (1985); McKoon & Ratcliff (1980); Sanford & Garrod (1981); Yekovich, & Walker (1978); Yekovich, Walker, & Blackman (1979).

2. It has been suggested that the Generic Type vs Specific Token distinction is similar to the traditional distinction between "nonspecific" and "specific" nouns (Chafe, personal communication). Yet the more traditional distinction is too broad (see Prince, 1981, for a similar view). In particular, the traditional distinction fails to capture the present distinction between nouns representing Generic Types and Multiple Events/Items.

3. Two parallel sets of analyses were conducted on each effect; in one set of analyses, "subjects" were considered a random factor, and in the other, "items" were considered a random factor. The results reported are based on the minF' statistics (Clark, 1973) when significant at the .05 level or lower. When the

minF' statistic is only marginally significant ($.05 > p < .08$), the separate "subjects" F_1 statistic and "items" F_2 statistic are reported.

4. The figures for the reading time data have been scaled individually for each conceptual-vs-literal anaphora situation. The reason for this is that the average number of characters in the sentences differed across the three situations. The sentences following sentences with Collective Set vs Individual Member nouns were an average 44.81 characters long; those following Multiple vs Unique Items/Events nouns were an average 45.09 characters, whereas those following Generic Type vs Specific Token nouns were 35.81 characters. Because number of characters is one of the factors affecting reading time (Haberlandt & Grasser, 1985), the reading time for the sentences following Generic Type vs Specific Token nouns were, on the average, faster than those following the other two siuations. Of course, the sentences for the four conditions within each of the three situations did not differ in number of characters because the same sentences were cycled through each of the four conditions.

Conditioning and categorization
Some common effects of informational variables in animal and human learning

Mark A. Gluck & Gordon H. Bower
Stanford University

To what extent do the processes of human learning emerge from complex configurations and elaborations of the "elementary" learning processes observed in animals? Research in the two areas of human and infra-human learning share a long history which focussed on elementary associative learning (Ebbinghaus, 1885; Pavlov, 1927). About twenty years ago, however, animal and human learning research became divorced from each other. Animal research continued to be primarily concerned with elementary associative processes (Mackintosh, 1983; Mackintosh & Honig, 1969; Rescorla & Holland, 1982); while human learning (or "memory") tended to be characterized in terms of information-processing and rule-based, symbol-manipulation, an approach borrowed from artificial intelligence. Few current theories of learning attempt to bridge the gap between human and infra-human learning (some exceptions include Alloy & Tabachnick, 1984; Estes, 1985; Dickinson & Shanks, 1985; Medin, 1984; Holland, Holyoak, Nisbett, & Thagard, in press). Recently, however, interest in relating human cognition to configurations of elementary associative connections has revived. Among theorists using parallel-distributed processing models, the works of McClelland, Rumelhart, Hinton, Sejnowski, and James Anderson are notable for demonstrating the computational power and psychological verisimilitude of these "connectionist" networks (see e.g., Hinton & Anderson, 1981; McClelland & Rumelhart, 1981; Ackley, Hinton, & Sejnowski, 1985, Rumelhart & McClelland, 1986).

Given the voluminous studies of learning in animals alongside current attempts to model cognition with elementary associative processes, it would seem particularly timely to search for and exploit any correspondences which might exist between animal and human associative learning. This was our goal in these experiments.

Informational Variables in Classical Conditioning

A simple but powerful theory describing animals' learning in classical Pavlovian conditioning was presented by Rescorla and Wagner in the early 1970's (Rescorla & Wagner, 1972; Wagner & Rescorla, 1972). In classical conditioning, a previously neutral stimulus, the *conditioned stimulus* (CS), such as a bell, comes to be associated with a biologically significant stimulus, the *unconditioned stimulus* (US), such as food or an electric shock. Early learning theories assumed that the simple temporal contiguity or joint occurrence of a CS and US was *sufficient* for associative learning (e.g. Hull, 1943; Spence, 1956). Later experiments made clear, however, that simple contiguity was not sufficient. The ability of a CS to become conditioned to a US depended on its imparting reliable and non-redundant information about the occurrence of the US (Kamin, 1969; Rescorla, 1968; Wagner, 1969).

To illustrate, suppose that a light, the CS, has already been conditioned to predict a shock, the US. If a compound stimulus consisting of a light and a tone is then paired with the shock, learning of the *tone→shock* association hardly occurs at all compared to control subjects who received no pretraining to the light (Kamin, 1969). This result, similar to Pavlov's work on the overshadowing of one cue by another, is called "blocking" because prior training of the *light→shock* association blocks later learning of the *tone→shock* association during the second, *(light + tone)→shock* stage of training.

The Rescorla-Wagner Model

The blocking effect suggested that the effectiveness of a US for producing associative learning depends on the relationship between the CS and the *expected outcome* (Rescorla, 1968; Wagner, 1969; Kamin, 1969). Rescorla and Wagner provided a precise formulation of this proposal (Rescorla & Wagner, 1972; Wagner & Rescorla, 1972). Their formulation assumes that the association which accrues between a stimulus and its outcome on a trial is proportional to the degree to which the outcome is unexpected (or unpredicted) given *all* the stimulus elements that are present on that trial. We let V_i denote the strength of association between stimulus element CS_i and the *US*. If CS_i is followed by a reinforcing unconditioned stimulus, *US*, then the change in the association strength between CS_i and the *US*, ΔV_i, can be described by Equation (1):

$$\Delta V_i = \alpha_i \beta_1 (\lambda_1 - \sum_{k \in S} V_k), \tag{1}$$

where α_i reflects the intensity or salience of CS_i, β_1 reflects the rate of learning on trials with US presentations, λ_1 is the maximum possible level of association strength conditionable with that US intensity, and $\sum_{k \in S} V_k$ is the sum of the associative strengths between all the CS stimulus elements occuring on that trial and the US. If CS_i is presented on a trial without the US, then the association between CS_i and the US decreases analogously, viz.,

$$\Delta V_i = \alpha_i \beta_2 (\lambda_2 - \sum_{k \in S} V_k), \tag{2}$$

where λ_2 is the level of associative strength supported by non-presentation of the US (usually taken to be zero), and β_2 reflects the rate of change of the association due to nonreinforcement. Generally β_1 is assumed to be larger than β_2, but this is not critical for most predictions (see Rescorla & Wagner, 1972).

The Rescorla-Wagner model is the most widely accepted description of associative changes during classical conditioning. The wealth of confirmed implications arising from this deceptively simple model has been substantial. This model accounts for the blocking effect as follows: When in Phase 1, CS_1 has been initially conditioned to the US, V_1 approaches λ_1. If the associative strength of the novel stimulus, V_2 is assumed to be zero,

then the *compound stimulus strength*, $V_1 + V_2 = \lambda_1$. By Equation 1, the incremental learning accruing to the novel stimulus, ΔV_2, when the compound is paired with the US is thus predicted to be zero--as observed.

Learning in Associative Networks

A learning rule used in many of the "connectionist" network models of cognition is the *delta rule*, a variant of the perceptron convergence procedure (Rosenblatt, 1961) first proposed as a learning mechanism for adaptive networks by Widrow and Hoff (1960). Such networks connect a set of input nodes to some output nodes with "connection weights" w_{ij} from node i to node j. Given a training trial relating an input vector to an output vector, the weights are changed according to (3):

$$\Delta w_{ij} = \beta(z_j - \sum_{k=1}^{n} w_{kj} a_k) a_i, \tag{3}$$

where i is an input node, j is an output node, a_i is the activation on input node i, the summation is over all the input nodes to node j, and z_j is a special "teaching" input signal to output node j indicating what the activation of that node should be to get the correct response. The delta rule provides an iterative solution to a set a linear equations which will converge on discriminating weights if they exist. Otherwise, the algorithm will converge on weights which minimize the "least-squares" error between the resulting and desired output patterns (Kohonen, 1977).

Recently, Rumelhart, Hinton, and Williams (1986) have generalized the delta rule so it may be applied to perform learning in a multi-layered net of feed-forward elements with some "hidden" units between the input and output layers. They show how the delta rule, combined with back-propagation of weight adjustments, can learn many difficult discriminations such as parity, exclusive-or, and symmetry relationships.

As Sutton and Barto (1981) noted, the delta rule is essentially identical to the Rescorla-Wagner equations (with $\beta_1 = \beta_2$). If, in Equation 3, we let $V_i = w_{ij}$, set the training signal in the delta rule, z_j, equal to λ_1 when the US is present and to zero otherwise, and let a_i=1 when CS_i is present and 0 otherwise, then the delta rule reduces to Equations 1 and 2 of the Rescorla-Wagner model. Curiously, associative network theorists have adopted the delta rule because of its computational power, convergence properties, and generalizability to multi-layered networks. Nonetheless, associative networks which implement the delta rule can be viewed as a framework for modeling the emergent properties of complex configurations of elementary associative processes observed in animals. However, few studies have asked whether the delta rule is an appropriate characterization of the algorithm underlying human associative learning.

Some earlier investigators have noted the need for bridging experiments. Rudy (1974) noted a parallel between human paired-associate learning and animal associative learning and pointed to a form of blocking in human learning. Specifically, when

redundantly relevant cues are compounded with stimuli that are already sufficient to cue the associated response, the added cues are unlikely to become associated with the response (Trabasso & Bower, 1968). Dickinson and Shanks (1985) demonstrated some conditioning phenomena in human learning: They showed that human judgments of event correlation were influenced by the conditional status of other events that are present, in a manner reminiscent of blocking or overshadowing phenomena in animal conditioning. Schank (1982) has recently postulated a similar "expectation failure" as the driving force behind learning; EPAM used a similar rule long ago (Feigenbaum, 1959; Feigenbaum & Simon, 1961).

Experiment 1

Because category learning is a currently active area in cognitive research, we decided to test out the delta rule as it applied to subjects' learning to classify stimulus patterns into categories. In our experiment, university students served as hypothetical medical diagnosticians. They saw a series of 250 "patients," each described by the presence or absence of each of four symptoms. The student diagnostician classified each patient as having one or the other of two fictitious diseases, received feedback about that patient's correct diagnosis. Over training, subjects learned which symptoms are more or less diagnostic of which diseases.

Figure 1 illustrates a simple associative network to represent this category learning. Each of the four symptoms is represented by an input node at the left, and the two disease categories by nodes at the right. The connections from symptom i to category j has weight, w_{ij}, reflecting the strength of evidence that presence of symptom i provides towards disease j. The w_{ij} will be adjusted trial by trial according to the delta rule.

The pattern of features presented on a trial causes a pattern of activation of the features. If the presence or absence of each feature is represented by activations of 1 and 0, respectively, the activation at a given category node will equal the sum of the weights from presented features to that category node. This reflects the model's expectation for that category given the symptom pattern. Once activation values are computed for the category nodes, the next step is for the model to select a response. Several measures of associative strength are possible. One we have used is to ask subjects to judge directly the probability that a given patient has one disease or the other. We will suppose that the greater the difference in net strength of evidence for category 1 vs. 2, the higher will be subject's estimate that the patient has disease 1 rather than disease 2. A second measure asks subjects to choose disease 1 or 2 for a particular patient. For this case, we use the ratio response rule of Luce (1963) which says that the probability of choosing Category 1 is the ratio $\frac{V_1}{(V_1 + V_2)}$. Qualitative aspects of the predictions do not depend on the details of the response rule.

The "training signal" provided to each category node (Figure 1) is the experimenter's feedback (after the subject's response) regarding the correct response. We assume that if category j is the correct classification, then z_j will be set equal to one on that trial; if an alternative category is correct on a given trial, then z_j will be 0 for that

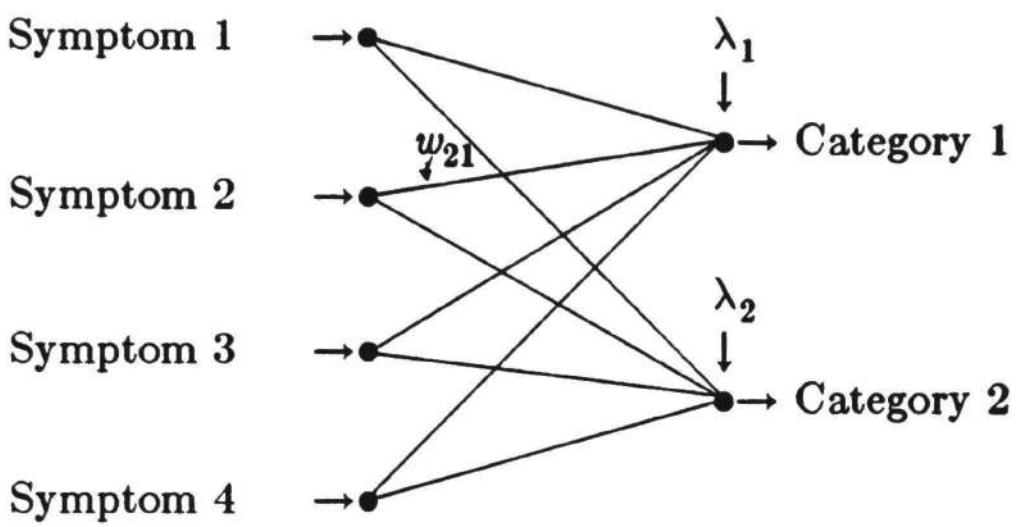


Figure 1. A simple "connectionist" network which learns to diagnose patterns of up to four symptoms as having one of two diseases using the Rescorla-Wagner/delta rule.

trial. We assume a single learning rate parameter, β, for adjusting the weights. However, if a fixed set of training patterns is presented many times in random order to the learning model, the convergence properties of the delta rule lead to *parameter free* predictions about the expected asymptotic levels of the w_{ij}'s, feature-to-category associations (Rescorla & Wagner, 1972; Stone, 1986).

We will compare the predictions of the delta rule in our category learning task with the predictions of three competing models of category learning (Estes, 1986): 1) *exemplar* models which presume that the learner stores all the exemplars of each category, and then classifies a new instance according to its similarity to the stored exemplars of each category (e.g. Medin & Schaffer, 1978; Nosofsky, 1984), 2) *feature-frequency* which presume that the learner stores relative frequencies of occurrence of cues within the categories, and then classifies an instance according to the relative likelihood of its particular pattern of features arising from each of the categories (Reed, 1972; Franks & Bransford, 1971), and 3) *prototype* models which presume the learner abstracts the central tendency (modal description) of each category and then classifies instances according to their similarity to this central prototype (Fried & Holyoak, 1984).

Applying models to our task where subjects estimate the probability of each category given each feature, the models make one of two predictions. Exemplar models and feature-frequency models predict that subjects' estimates will simply reflect the observed conditional feature-to-category probabilities of the training sequence, a form of "probability matching." On the other hand, prototype models and feature-frequency models which ignore variations in category base-rate frequencies would predict that subjects' estimates of the probability of the category given the feature will reflect simply the relative likelihood of the feature given the alternate categories, viz.,

$$\frac{P(f|c_1)}{P(f|c_1)+P(f|c_2)}.$$

In our experiment, we arranged to have the ordinal relationships among the conditional probabilities for different cues differ from the ordinal relationships among the expected asymptotic association strengths predicted by the Rescorla-Wagner/delta rule. This was achieved by unbalancing the relative frequency of the two diseases, making the common disease far more likely than the rare disease. The question was whether people's probability estimates would be more closely predicted by the Rescorla-Wagner/delta rule than by the alternative models.

Procedure

Nineteen subjects were trained to classify medical charts of hypothetical patients into one of two mutually exclusive disease categories. Disease names were fictitious but we will refer to them as the rare (R) disease and the common (C) disease. Among the training exemplars, patients with the common disease were three times as frequent as patients with the rare disease. A patient chart consisted of one to four symptoms drawn from a set of four possible symptoms: bloody nose, stomach cramps, puffy eyes, and discolored gums. In the training phase subjects were shown a set of symptoms corresponding to a patient, asked to make a diagnosis, and then given feedback as to the correct diagnosis. Figure 2a shows the probability of each of the four symptoms occurring in patients

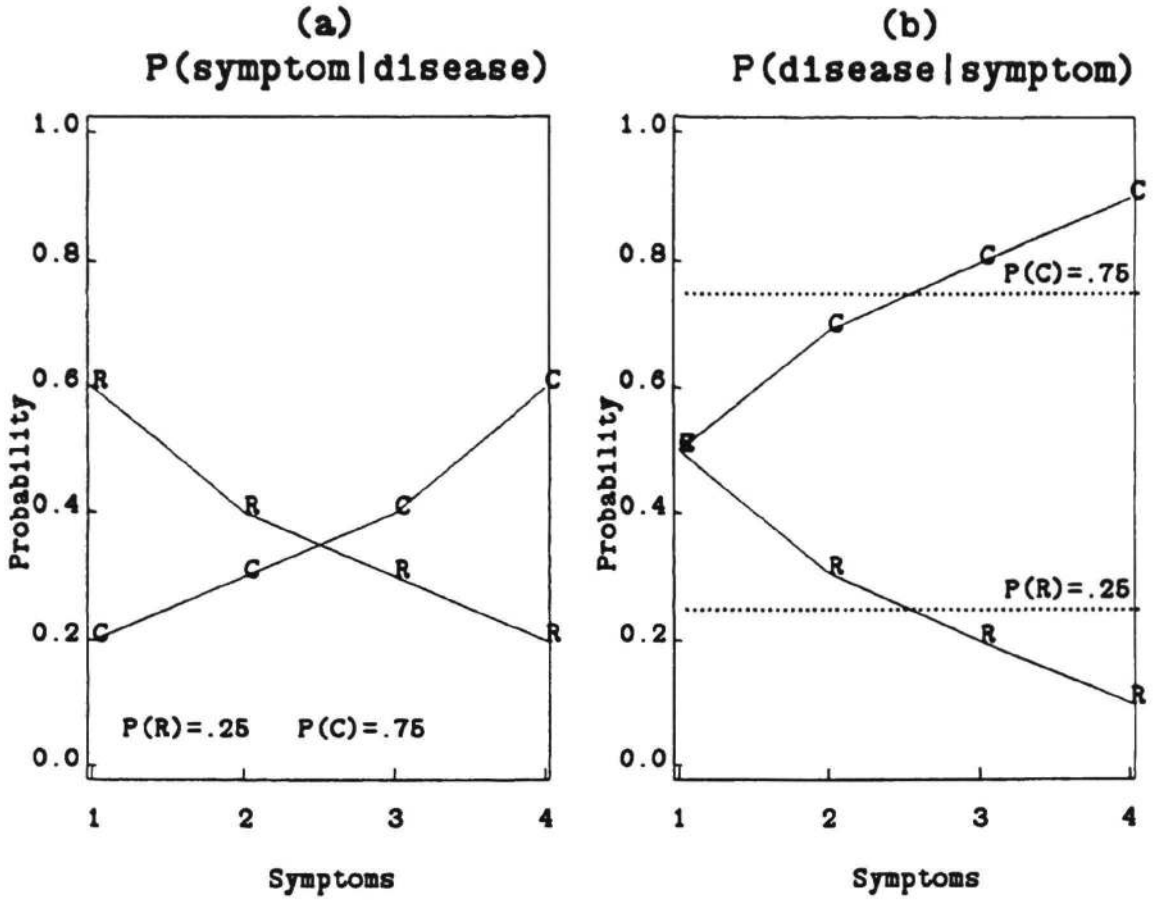


Figure 2. Experiment 1 design: (a) The probabilities of each of the four symptoms occurring in patients suffering from each of the two diseases. The lower numbered symptoms were more typical for the rare disease while the higher numbered symptoms were more typical of the common disease. (b) The conditional probabilities of each of the two diseases given the presence of each of the symptoms computed from (a) using Bayes Theorem.

suffering from each of the two diseases. The lower numbered symptoms were more typical for the rare disease while the higher numbered symptoms were more typical of the common disease. All symptoms, however, occurred in some patients with both diseases. Symptoms 1, 2, 3, and 4 were assigned actual symptom names randomly for each subject. Each subject received a novel set of training patients which were generated during the experiment according to a probabilistic procedure. First, each patient was randomly designated as suffering from either the rare disease (with probability .25) or the common disease (with probability .75). Second, given his disease, a patient's symptom chart was generated by choosing symptoms according to the *independent* probabilities shown in Figure 2a. Thus, if the patient suffered from the rare disease, then with probability .6, the chart would include symptom 1; with probability .4, symptom 2; with probability .3, symptom 3; and with probability .2, symptom 4 (and analogously, but inversely, for patients suffering from the common disease). From one to four symptoms were presented on a single chart (patients with no symptoms were eliminated from the training sequence). For the subjects, the diseases were identified by fictitious names which were counterbalanced across subjects in being assigned to the rare or common disease. Subjects were instructed that there was no simple rule for making the diagnosis and that the order of presentation of the symptoms within a patient's chart was irrelevant.

Using the base rates of $P(R) = .25$ and $P(C) = .75$ and the probabilities in Figure 2a, Bayes Theorem provides the conditional probability of the two diseases given the four symptoms considered seperately (see Figure 2b). For any single symptom the normative probability of the rare disease was always less than or equal to the probability of the more common disease.

Following 250 training trials of predicting diseases and receiving feedback, subjects were finally asked to estimate directly the probability that a patient exhibiting a particular symptom was suffering from one or the other disease. They gave a numerical estimates of $P(R|s_i)$ and $P(C|s_i)$ on a 0 to 100 scale for each of the four symptoms. These estimates are the data of primary interest in this report.

Results and Predictions

Because the conditional probabilities of the two diseases sum to 1 for any particular symptom, we will combine these conditional probabilities into a single *probability difference measure*, $P(R|s_i) - P(C|s_i)$, for each of the four symptoms. This measure, shown in Figure 3, reflects both the actual (normative) probabilities in the training patterns as well the probability matching behavior predicted by exemplar-storage and feature-frequency models. But, the Rescorla-Wagner/delta rule predicts that following training, subjects' estimates of the probability differences will follow a different pattern, reflecting the underlying strengths of the feature-to-category associative connections. These asymptotic connection weights can be calculated by deriving equations for the expected trial-by-trial weight change in each of the feature-to-category connections, setting these expected changes to zero, and solving the resulting four simultaneous equations in four variables for each of the two categories. The resulting asymptotic association strengths to the rare disease are .45, .18, .06, and -.09 for symptoms 1 through 4, respectively, and for the associations to the common disease, .02, .22, .37, and .68. The differences between these asymptotic strengths are plotted in Figure 3b; this is the theoretical index to be compared

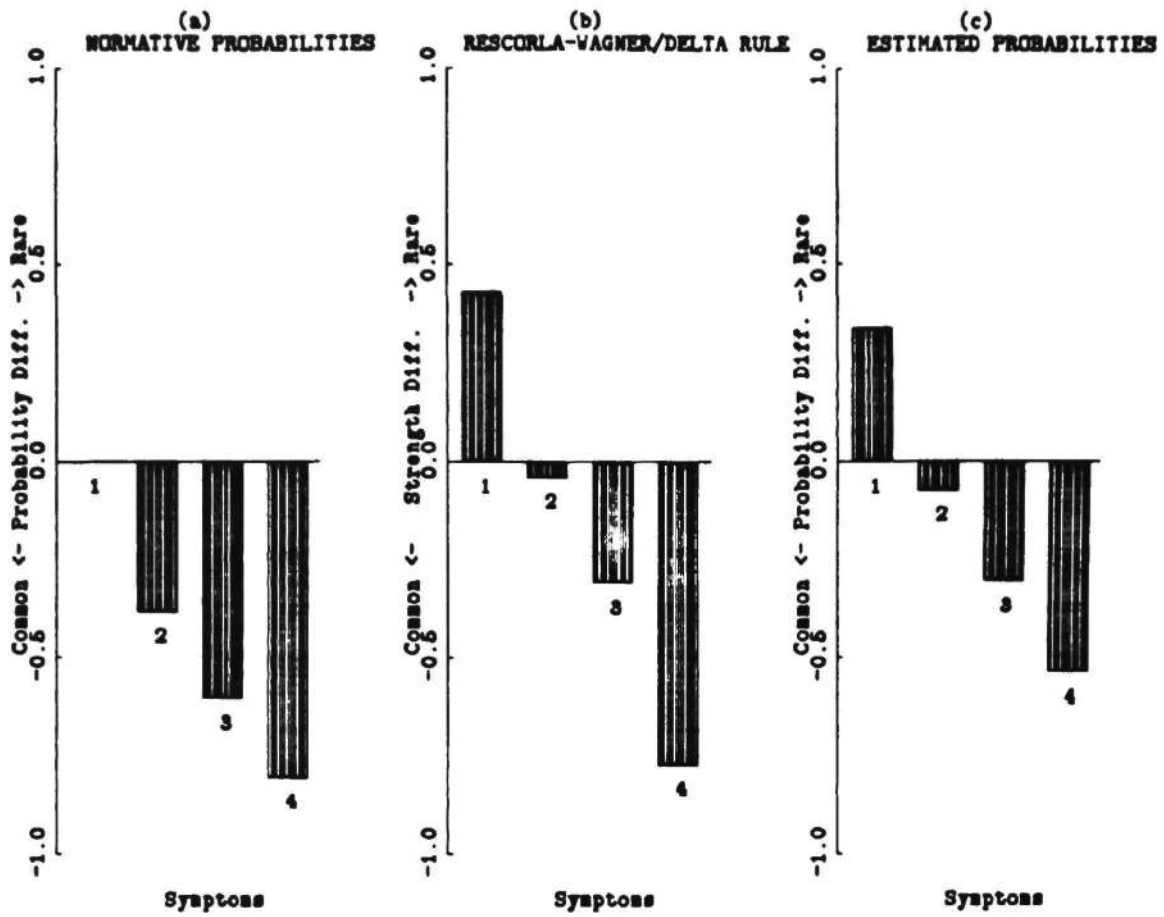


Figure 3. Results and predictions for Experiment 1. (a) Normative probability differences for each symptom. These correspond to the predictions of exemplar and feature-frequency learning models. (b) Predictions of the Rescorla-Wagner/delta rule based on asympotic levels of associations. (c) Subject's estimates of the probability differences.

to the observed probability difference measures.

The most striking difference between the normative probability measures in Figure 3a and the predicted associative weights in Figure 3b is evident in symptom 1, s_1: This symptom was paired equally often with the rare disease, R, as with the common disease, C, and hence the difference between the conditional probabilities of R versus C to this cue is zero. However, the delta-rule predicts that the the $s_1 \rightarrow R$ association will be considerably stronger than the $s_1 \rightarrow C$ association.

This prediction of the delta rule is understandable when one appreciates the *competitive* nature of the learning algorithm. The overall magnitude of the *symptom→disease* weight reflects the degree to which a symptom has been an informative and reliable predictor of a disease, *relative* to the predictive value of other co-present symptoms for that same disease. Although symptom 1 has the same predictive value for the two diseases, relative to the predictive value of the other symptoms for the common disease, it is not a very informative predictor. However, for the rare disease symptom 1 is a relatively better predictor than the other symptoms. It is this relative validity of a symptom for the two categories that determines its relative degrees of association to them.

Having described the model's predictions, we turn now to the data. Comparing the actual with the estimated conditional probabilities indicated that while subjects correctly learned the relative strengths of the conditional probabilities *within a particular disease category*, they considerably *overestimated* the conditional probability of the rare disease given each of the symptoms. Subjects' estimates of the probability of disease R versus C were converted into differences and graphed in Figure 3c. Our preceding analyses suggested that the data for symptom 1 are would be most critical for distinguishing between the models. As predicted by the delta rule, the data indicate that subjects believed that patients with symptom 1 were significantly more likely to be suffering from the rare disease than from the common disease ($p < .01$, 1 tailed, $t=2.76$, $df=18$). This simple result disconfirms the alternative models. By the same token, the delta rule expects the probability difference for symptoms 2, 3, and 4 to be much less than predicted by the probability matching theories, and this data pattern was also observed.

It would appear that our learners fell prey to a common form of " base rate neglect" in making predictive judgments: They erroneously judged that the presence of a symptom, s_1, highly representative of the rare disease was strong evidence for diagnosing the rare as opposed to the common disease. This result, predicted by the delta rule, is consistent with many results in research on judgment: People consistently overestimate the degree to which evidence that is representative or typical of a rare event is actually predictive of it (Tversky & Kahneman, 1972). When answering questions such as: "What is the probability that object A belongs to class B", people often resort to a *representativeness heuristic* in which their judgment reflects the degree to which A resembles a prototype of B (Tversky & Kahneman, 1982b). For example, in estimating the probability that a particular student is an English major in a classroom known to consist of 80% computer science majors, people base their predictions largely on the degree to which the personality characteristics of the student are representative of their stereotypes of computer wizards, thus neglecting the influence that base rate should have. (Kahneman & Tversky, 1972). Most studies demonstrating neglect of base rate in classification

judgments have used natural categories with familiar prototypes (e.g. feminists or engineers); base rate information is generally presented to subjects as additional numerical information (Tversky & Kahneman, 1982a). Ours is one of the first studies to demonstrate base rate neglect in a category learning experiment in which information about categories and base rates was induced by subjects from examples.

One might try explaining our results by supposing that subjects ignore base rates of the two categories in making their judgments. Earlier we indicated how a prototype model or a feature-frequency model could be interpreted as insensitive to base rates of the two categories. But this explanation fails because if subjects had been ignoring base rates, then they should have judged symptoms 1 and 2 to be as diagnostic of the rare disease as symptoms 3 and 4 were for the common disease, respectively (see Figure 2a). But as Figure 3c shows, this pattern was not obtained. Only symptom 1 was judged to be a significantly stronger predictor of the rare disease than the common disease. Though subjects' probability estimates reflected less attention to base rates than normatively required, the estimates show sensitivity to the differing base rates. These results are consistent with current decision making studies which suggest that in most situations base-rate information is not ignored, only underutilized (Borgida & Brekke, 1981; Kassin, 1979). Alternative category learning models, which predict either total neglect of base rates or full use of base rate information, do not provide a satisfactory account for these data. The Rescorla-Wagner/delta rule, however, correctly predicts that in our situation only symptom 1 will be perceived as stronger evidence for the rare versus the common disease; the other symptoms are predicted to be stronger evidence for the common disease.

Discussion

We believe that our results provide discriminating evidence in favor of the Rescorla-Wagner/delta rule as applied to a simple cue-to-category learning task for human adults. The unique nature of the predictions depends on the competitive nature of the learning algorithm. A cue that is paired with category 1 will acquire relatively little associative strength towards that category if the cue occurs in the company of others that already strongly predict that category. As indicated, this echoes the many results in animal conditioning on overshadowing, blocking, and CS-US correlations (see Prokasy, 1965; Rescorla, 1968). We are currently conducting further tests of implications of the delta rule in the symptom-disease learning paradigm, and those tests provide even further confirmation of the rule. We are encouraged that the delta rule of connectionist theories not only links up to the Rescorla-Wagner model of conditioning, but that they also imply the phenomenon of base-rate neglect which has proven to be a robust effect in the literature of judgment and decision. Such theoretical connections across disparate research areas are especially encouraging to the goals of cognitive science.

References

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). Boltzman machines: Constraint satisfaction networks that learn. *Cognitive Science, 9,* 147-169.

Alloy, L. B. & Tabachnik, N. (1984). Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information. *Psychological Review, 91,* 112-149.

Borgida, E. & Brekke, N. (1981). The base-rate fallacy in attribution and prediction. In J. H. Harvey, W. J. Ickes, & R. F. Kidd (Eds.), *New directions in attribution research (Vol. 3).* Hillsdale, N.J.: Erlbaum.

Dickinson, A. & Shanks, D. (1985). Animal conditioning and human causality judgment. In L. Nilsson & T. Archer (Eds.), *Perspectives on learning and memory.* Hillsdale, N.J.: Lawrence Erlbaum.

Estes, W. K. (1985). Some common aspects of models for learning and memory in lower animals and man. In L. Nilsson & T. Archer (Eds.), *Perspectives on learning and memory.* Hillsdale, N.J.: Lawrence Erlbaum.

Estes, W. K. (1986). Array models for category learning. *Cognitive Psychology, in press.*

Feigenbaum, E. A. (1959). *An information processing theory of verbal behavior.* Santa Monica, CA: RAND Corporation.

Feigenbaum, E. A. & Simon, H. A. (1961). Forgetting in an associative memory. *Proceedings of the ACM National Conference, 16,* 202-205.

Franks, J. J. & Bransford, J. D. (1971). Abstraction of visual patterns. *Journal of Experimental Psychology, 90,* 65-74.

Fried, L. S. & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory and Cognition, 10,* 234-257.

Hinton, G. E. & Anderson, J. A. (1981). *Parallel models of associative memory.* New Jersey: LEA.

Holland, J., Holyoak, K. J., Nisbett, R. E., & Thagard, P. (in press). *Induction: Processes of Inference, Learning and Discovery.*

Hull, C. L. (1943). *Principles of behavior.* New York: Appleton-Century-Crofts.

Kahneman, D. & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3,* 430-454.

Kamin, L. J. (1969). Predictability, surprise, attention and conditioning. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behavior* (pp. 279-296). New York: Appleton-Century-Crofts.

Kassin, S. M. (1979). Base rates and prediction: The role of sample size. *Personality and Social Psychology Bulletin, 5,* 210-213.

Kohonen, T. (1977). *Associative memory: A system-theoretic approach.*

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology.* New York: Wiley.

Mackintosh, N. J. (1983). *Conditioning and Associative Learning.* Oxford: Oxford Univ. Press.

Mackintosh, N. J. & Honig, W. K. (1969). *Fundamental Issues in Associative Learning.* Halifax: Dalhousie Univ. Press.

McClelland, J. L. & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review, 88.*

Medin, D. L. & Dewey, G. I. (1984). Learning of ill-defined categories by monkeys. *Canadian Journal of Psychology, 38,* 285-303.

Medin, D. L. & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85,* 207-238.

Nosofsky, R. M. (1984). Choice, Similarity, and the Context Theory of Classification. *Journal of Experimental Psychology: Learning, Memory and Cognition, 10,* 104-114.

Pavlov, I. (1927). *Conditioned Reflexes.* London: Oxford University Press.

Prokasy, W. F. (1965). Classical eyelid conditioning: Experimental operations, task demands, and response shaping. In W. F. Prokasy (Ed.), *Classical conditioning* (pp. 208-225). New York: Appleton-Century-Crofts.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology, 3,* 382-407.

Rescorla, R. A. (1968). Probability of shock in the presence and absence of CS in fear conditioning. *Journal of Comparative and Physiological Psychology, 66,* 1-5.

Rescorla, R. A. & Holland, P. C. (1982). Behavioral studies of associative learning in animals. *Annual Review of Psychology, 33*, 265-308.

Rescorla, R. A. & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* . New York: Appleton-Century-Crofts.

Rosenblatt, F. (1961). *Principles of neurodynamics: Perceptrons and the theory of the brain mechanisms*. Washington, D.C.: Spartan.

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

Rumelhart, D. E. & McClelland, J. L. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1: Foundations*. Cambridge, MA: Bradford Books/MIT Press.

Schank, R. C. (1982). *Dynamic memory*. Cambridge: Cambridge University Press.

Spence, K. W. (1956). *Behavior theory and conditioning*. New Haven, Conn.: Yale University Press.

Stone, G. O. (1986). An analysis of the delta rule and the learning of statistical associations. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1: Foundations*. Cambridge, MA: Bradford Books/MIT Press.

Sutton, R. S. & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review, 88*, 135-170.

Trabasso, T. & Bower, G. H. (1968). *Attention in learning: Theory and research*. New York: John Wiley.

Tversky, A. & Kahneman, D. (1982a). Evidential impact of base rates. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgments under uncertainty: Heuristics and biases*. Cambridge, UK: Cambridge University Press.

Tversky, A. & Kahneman, D. (1982b). Judgments of and by representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgments under uncertainty: Heuristics and biases*. Cambridge, UK: Cambridge University Press.

Wagner, A. R. (1969). Simulus selection and a modified continuity theory. In G. Bower & J. Spence (Eds.), *The psychology of learning and motivation, Volume 3.* New York: Academic Press.

Wagner, A. R. & Rescorla, R. A. (1972). Inhibition in Pavlovian conditioning: applications of a theory. In R. A. Boakes & S. Halliday (Eds.), *Inhibition and Learning* (pp. 301-36). New York: Academic Press.

Widrow, G. & Hoff, M. E. (1960). Adaptive switching circuits. *Institute of Radio Engineers, Wester Electronic Show and Convention, Convention Record, 4,* 96-194 .

SIMULTANEOUS CONFIGURAL CLASSICAL CONDITIONING

Jeffrey C. Schlimmer and Richard H. Granger, Jr.
Department of Information and Computer Science
University of California, Irvine

ABSTRACT

Humans and animals have the ability to learn complicated configurations of environmental cues that are predictive of important events. In classical conditioning, this task is called *configural* conditioning. Psychologists have studied this phenomenon since Pavlov's time, yet several of the contemporary learning models provide only partially satisfactory explanations. Most models provide mechanisms which select among possible predictive stimuli, but they fail to explicitly identify predictive *combinations* of stimuli and are thus restricted to learning only a relatively simple set of possible associations. In this paper we discuss a learning method which accounts for some configural conditioning results. Using an implemented system, we demonstrate the effectiveness of this method by modeling configural conditioning data from a pair of representative experimental studies.

INTRODUCTION

Consider the problem of trying to learn the precise configuration of weather cues that indicate rain. The appropriate description may include falling barometer readings and high humidity or a high degree of atmospheric ionization. The study of the ability of humans and animals to associate *sets* of stimuli with an important event is called *configural* conditioning. For example, in one case dogs were trained to associate a simultaneous presentation of six specific cues with delivery of meat powder (Razran, 1965). The animal then expectantly salivated only when all of the six cues were present and not when any subset of them were. Remarkably enough, a dog can also be trained to expect food only when one of two features occurs *separately* and not when they occur together (Woodbury, 1943). The ability to form associations with a number of Boolean combinations of features (conjunction, disjunction, exclusive-disjunction) is a well studied phenomenon in experiments on human ability (Grings, 1972; Bruner, Goodnow, & Austin, 1965) and in animal classical conditioning (Whitlow & Wagner, 1972; Saavedra, 1975).

Several contemporary animal learning theories, however, have difficulty explaining these results. The Rescorla–Wagner (1972) model, for instance, expresses the strength of a configural association as the sum of the associations for each cue. By adjusting the individual strengths, it *selects* the most predictive stimulus. A configuration consisting of any of a number of cues is easily accommodated. Conjunction and exclusive-disjunction, however, require that the individual cues have a qualitatively different associative strength than their combination. Within a sensory dimension, this difficulty is usually finessed by assuming that the co-occurrence of two stimuli (say a blue light and a red light) results in some new *resonant* property of the stimuli (say a purple light). Feature selection models like this one rely on resonant features in order to learn associations involving conjunctive and exclusive-disjunctive configurations. However, two stimuli may be configured from different modalities, as in the case where a tone and a light are reinforced alone but not together. It is unclear that any additional property of the stimuli is present (though perhaps one might wish to argue that some property of "two-ness" exists). Yet without it, traditional models cannot explain the effectual acquisition of these complex CSs; they assume that the association for a set of cues is simply the sum of associations with each part. A exclusive-disjunctive configuration requires weak learning of the compound to arise out of strong learning of each component. An additional difficulty of assuming the presence of resonant properties is that the number of these features must increase exponentially with the number of cues in the environment.

A number of artificial intelligence learning methods also have difficulties explaining these experimental data. A common assumption in concept attainment work, for instance, is that the identity of an instance can be determined via a conjunctive description of its features (Mitchell, 1982). A simple disjunctive description cannot be learned or represented in many cases, much less the exclusive-disjunction relationship. In this paper we present a model, STAGGER, which has the functional flavor of a feature selection model, but goes beyond this to form new, compound features. We further demonstrate its ability to correctly model the results from two representative experimental studies.

A LEARNING MODEL: STAGGER

The foundation of STAGGER is a distributed representation of association, composed of a set of dually weighted predictive features. During each trial, a cumulative expectation of the US is formed by utilizing the pair of weights associated with each feature. These weights are easily adjusted as learning progresses, and their mathematical interpretation mirrors basic results in learning. A secondary form of learning comes into play after expectation of the US fails; new features are introduced into the representation which are more general, more specific, or inverted Boolean functions of existing features.

At a higher level, associations in STAGGER are initially formed between primary, perceptual features in a similar manner to the processes in feature selection models. As conditioning progresses, compound features are formed internally which do not have an

immediate correlate to individual perceptions. These new features are then part of the selection process, enabling STAGGER to learn complex configurations without relying on potentially nonexistent resonant features. The formation of effective feature compounds, then, is a central part of the process of configural conditioning in our model. In the following sections, we first describe STAGGER's associational representation and its feature selection processes. We then explicate its configural learning mechanisms.

Representation and Expectation

Associations are represented in STAGGER as a set of dually weighted features. The two weights for each feature capture positive and negative implication: one weight represents the sufficiency of the feature for the US, or $(CS \Rightarrow US)$, and the other represents its necessity, or $(\neg CS \Rightarrow \neg US)$. The mathematical measures chosen for these weights mirror the results of contingency experiments on learning in humans (Wasserman, Chatlosh, & Neunaber, 1983) and animals (Rescorla, 1968; Colwill & Rescorla, in press). Specifically, a novel cue comes to be excitatorily associated with an unpleasant stimulus only if the probability of the US in the presence of the CS is greater than its probability in the absence of the CS: $p(US|CS) > p(US|\neg CS)$. In behavioral terms, this means that if either the CS or the US frequently occurs alone, the subject still learns an association between the two stimuli. However, if they both occur alone even a few number of times, learning about their association is severely impaired.

With this in mind, STAGGER uses logical sufficiency (LS), or positive likelihood ratio, as a measure of sufficiency (Duda, Gaschnig, & Hart, 1979). Similarly, logical necessity (LN), or negative likelihood ratio, serves to measure necessity. They are defined as:

$$LS = \frac{p(CS|US)}{p(CS|\neg US)} \qquad LN = \frac{p(\neg CS|US)}{p(\neg CS|\neg US)}$$

LS ranges from zero to positive infinity and is interpreted in terms of odds. (Odds may be easily converted to probability $p = odds/(1 + odds)$.) An LS value less than unity indicates a negative correlation, unity indicates independence, and a value greater than unity indicates a positive relationship. LN also represents odds. However, an LN value near zero indicates a positive correlation, while a value greater than unity indicates negative correlation. For both LS and LN, unity indicates irrelevance. The LS and LN measures adhere to the contingency law, for it can be shown via algebraic manipulations that $LS > 1$ and $LN < 1$ if and only if $p(US|NC) > p(US|\neg NC)$ (Schlimmer, 1986).

In a given trial, all of the individual feature association weights influence US expectation. Following the mechanism used by Duda, Gaschnig, and Hart (1979), expectation of the US is the product of the prior odds of the US, the LS values of all present features, and the LN values of all absent ones.

$$Odds(\text{US}|\text{CSs}) = Odds(\text{US}) \times \prod_{\forall present} LS \times \prod_{\forall absent} LN$$

Table 1: Possible CS–US trial types.

	CS Present	CS Absent
US Present Reinforced	Confirming Positive (C_P)	Infirming Positive (I_P)
US Absent Nonreinforced	Infirming Negative (I_N)	Confirming Negative (C_N)

The resulting number represents the odds in favor of the US occurring. The effect of this multiplicative calculation is that learned associative strengths have a cumulative influence on US prediction. Two features which are highly predictive of the US cause a greater expectation when both are present than when only one of them is. However, as we will see in following sections, STAGGER is not confined to implicit representing configurations (via a summational effect of components) since it formulates new, explicit compound features which develop independent associative strengths.

Learning Mechanisms

In addition to computing a holistic expectation from the dual associational weights, STAGGER incrementally modifies the feature weights and generates new features. These two latter abilities allow STAGGER to adapt its associational description to better reflect the conditioning environment.

Feature selection

The sufficiency and necessity weights associated with each of the features may be easily adjusted. Consider the possible situations that arise during a conditioning trial. Following the terminology used by Bruner, Goodnow, and Austin (1956), a reinforced trial is positive evidence which may either *confirm* the predictiveness of a feature (if it is present in this trial) or *infirm* the feature's predictiveness (if it is absent). Similarly, a nonreinforced trial is negative evidence which either confirms an absent feature or infirms a present one. Table 1 summarizes these possibilities. In terms of these matching events, the contingency law implies that learning occurs in cases involving at most one type of infirming evidence. In situations with even small amounts of both positive and negative infirming evidence, subjects fail to learn an association.

LS and LN may be calculated by keeping counts for each feature of the possible situations listed in Table 1.

$$LS = \frac{C_P(I_N + C_N)}{I_N(C_P + I_P)} \qquad LN = \frac{I_P(I_N + C_N)}{C_N(C_P + I_P)}$$

The prior odds for the US are estimated as $(C_P + I_P)/(I_N + C_N)$. Note that the LN measure will rank features with negative infirming evidence highly, but features with both

Table 2: Feature formation heuristic.

Expectation	Reinforced	Error type	Feature formed
US	¬US	Commission	AND[f1,f2]
¬US	US	Omission	OR[f1,f2]
—	—	Either	NOT[f1]

types of infirming evidence poorly. This reflects learning in partial reinforcement situations (Fitzgerald, 1963) and is consistent with contingency experiments (Rescorla, 1968).

If STAGGER limited its learning to adjustment of the feature weights, the distributed association would be sufficient to accurately describe the class of "linearly separable" concepts (Hampson & Kibler, 1983). In this respect STAGGER is similar to *connectionist* models of learning when those models do not have any "hidden" units. The purpose of the hidden, internal units is to allow the encoding of more complicated associations. Feature formation processes in STAGGER serve an analogous purpose: individual features are combined into more complex Boolean functions.

Feature formation

STAGGER is not limited to acquiring summational associations between immediate, perceptual features. New, internal compound features are introduced by the model, allowing it to encode the potentially complex associations involved in configural conditioning. Because it is able to identify effective featural combinations internally, no assumption regarding additional resonant features is required. STAGGER follows three levels of heuristics in its formation of compound, internal features, and it constructs them using conjunction, disjunction and negation.

The first heuristic suggests a new compound feature when STAGGER makes an expectation error: either expecting the US in a nonreinforced trial or failing to predict the US in a reinforced trial. In the first case, the commission has admitted one too many possible situations as predictive of the US. A compound feature with a restricted application is formed using conjunction. This feature will be true less often than its components and can act to dampen the expectation process. In the second case, STAGGER is failing to include stimuli which do lead to the US, or making an error of omission. A more admitting, or more general, feature compound is formed using disjunction; it will be true more often than its components and thus loosens the class of possible predictors of the US. In either case, STAGGER forms a negated feature compound. Table 2 summarizes this heuristic.

Choosing appropriate features for new formations is accomplished via two additional heuristics. One heuristic *nominates* either present or absent features for combination, and the other narrows the possible features down by *electing* one or two of the most predictive. STAGGER's nomination heuristic specifies whether present or absent features are to be used in forming compound features, depending on the type of feature combination and prediction error. After STAGGER has made an error of commission, features present on

Table 3: Nomination heuristic.

Error type	Function formation	Feature nomination
Commission	AND[f1,f2]	Present, Absent
	OR[f1,f2]	Absent, Absent
	NOT[f1]	Present
Omission	AND[f1,f2]	Present, Present
	OR[f1,f2]	Present, Absent
	NOT[f1]	Absent

this nonreinforced trial may be partially necessary, but are clearly not sufficient for reinforcement. Conjunction nominates two necessary features, and thus a present feature is combined with an absent one. Nominating a present feature is motivated by noticing that some feature was present and suggested that this trial was likely to be reinforced. Disjunction nominates two sufficient features, so two features absent in this nonreinforced trial are chosen; no sufficient features were present. Negation is used to identify safety-signal features (those which, when present, indicate safety from the US) and thus nominates its component from the collection of features which were present. The appropriate nominations following an error of omission (an unpredicted, reinforced trial) are derived by similar reasoning. Table 3 summarizes the nomination heuristic.

The election heuristic further narrows the possible features for combination. Consider a situation leading STAGGER to appropriately form a new conjunction. For example, the familiar concept *father*: a parent and a male. The two features (parent and male) are always reinforced (father) though each is separately nonreinforced (a brother is male). This is negative infirming evidence (Table 1), and therefore LN which tolerates negative infirming evidence is used to elect features for a conjunctive configuration. By a similar argument LS elects features to be used in forming a disjunction. Features are elected equally by both measures for negated formations. Table 4 summarizes this third heuristic.

Table 4: Election heuristic.

Function	Election measure
AND[f1,f2]	$LN(\mathtt{f}i) \ll 1$
OR[f1,f2]	$LS(\mathtt{f}i) \gg 1$
NOT[f]	$LN(\mathtt{f}) \gg 1$ or $LS(\mathtt{f}) \ll 1$

These three heuristics may be used in concert, each one driving the others. For example, in the case of an unexpected, reinforced trial (an error of omission), a disjunction may be formed (see Table 2) to combine the present feature (Table 3) which has the lowest LN value (Table 4) with the absent feature with the lowest LN value. In many cases, these three heuristics cooperate in just this manner. However, there are some situations (as in

Table 5: Configural training.

HL+, H−, L−	Positive patterning	or	AND[H,L]
HL−, H+, L+	Negative patterning	or	XOR[H,L]

Woodbury's (1943) negative patterning discussed below) in which the election heuristic cannot offer any guidance. The process of forming effective internal, compound features then proceeds using the remaining two heuristics.

The feature formation process is limited by pruning ineffective compound features. Specifically, LN is used to assess the validity of a new conjunction. This more restrictive feature is true less often that its components are (it is guaranteed to have the same or less negative infirming evidence), so if it also has less positive infirming evidence (and has a better LN weight), it is deemed an effective feature. LS is used to test a more inclusive compound feature. Inverted features are tested by comparing them to the inverse of their Bayesian measure (e.g., $1/LS$). The role of these pruning measures is similar to that of the test component of a generate-and-test algorithm. The three formation heuristics serve to guide the generation of new feature compounds while the Bayesian measures are used to prune ineffective ones.

STIMULUS PATTERNS

Woodbury (1943) was one of the first American researchers to investigate learning of configural cues. In classical conditioning experiments with dogs, he studied different configurations of a low and a high buzzing sound which served as CSs for the delivery of a food pellet. He investigated two simultaneous configurations: positive patterning (in which only the presence of both buzzing sounds was reinforced), and negative patterning (where only the presence of either of the buzzing sounds was reinforced). In the first situation, a Boolean conjunction of the two buzzing sounds was reinforced; in the second, a exclusive-disjunction was reinforced. Table 5 summarizes these two training conditions.

Figure 1 depicts STAGGER's acquisition of the positive pattern (conjunctive) configuration. The upper heavy line represents conditioned responses (CRs) in trials which contained both of the buzzing sounds. The lower, light, solid line represents CRs to trials which contained only the lower of the two buzzing sounds and the lower, dotted line represents responses to the higher of the two buzzers. Each point represents the percentage of CRs in the last ten trials and is an average over ten separate program executions.

A strict feature selection model could learn to correctly predict reinforcement if there were a resonant feature resulting always and only from the co-occurrence of the low and high buzzers (Rescorla & Wagner, 1972, p. 86, fn. 2). Such a resonant feature is plausible given that both of the stimuli are within the same sensory modality. However, for the purpose of demonstrating the capabilities of the feature formation processes in STAGGER,

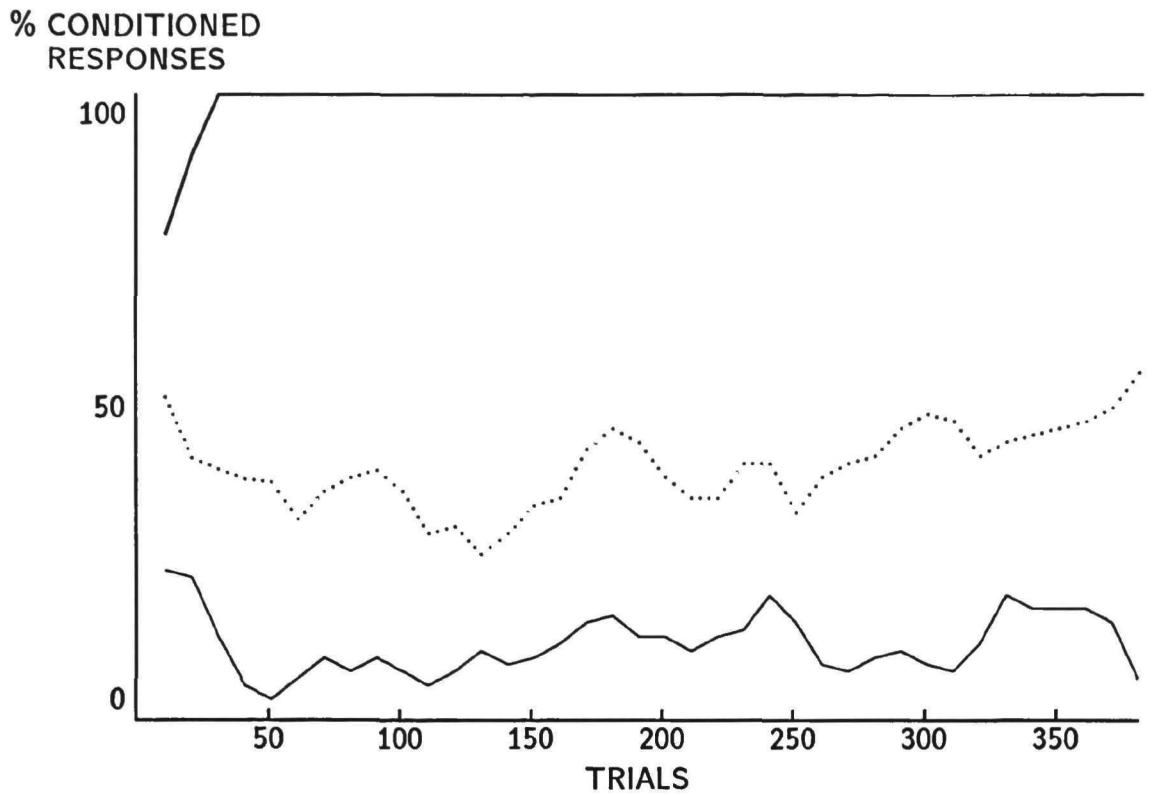


Figure 1: Conditioning to HL+, H−, L−.

this assumption about resonant features was omitted from the trial specification input to the program. Distinction between the combination of the two buzzers and either separately is facilitated by the introduction of the compound feature `AND[low-buzz,high-buzz]` via the heuristics described in section : the *LN* measure ranked low-buzz and high-buzz as the most effective individual predictors in an expected, but nonreinforced trial. A conjunctive feature was constructed using the low-buzz feature which was present and the high-buzz feature which was absent.

Figure 2 depicts STAGGER's acquisition of the negative pattern (XOR) configuration. Again the heavy line represents CRs to a co-occurrence of the two buzzing sounds; the light line, CRs to trials with only the low buzzing sound; the dotted line, CRs to the higher buzzing sound. After approximately 200 trials, STAGGER is effectively distinguishing between reinforced and nonreinforced trials.

If the presence of a resonant buzzing feature was assumed, a strict feature selection model could model this learning, for the resonant low and high buzzing feature could have a strong negative associative strength which would overpower either of the positive strengths of each of the individual buzzing cues.

STAGGER formed two compound features in order to accurately predict reinforcement: `AND[low-buzz,NOT[high-buzz]]` and `AND[NOT[low-buzz],high-buzz]`. A simple disjunction of these compounds captures negative patterning. However, the formation of these configurations was not as straightforward as was the case in Figure 1. As we inti-

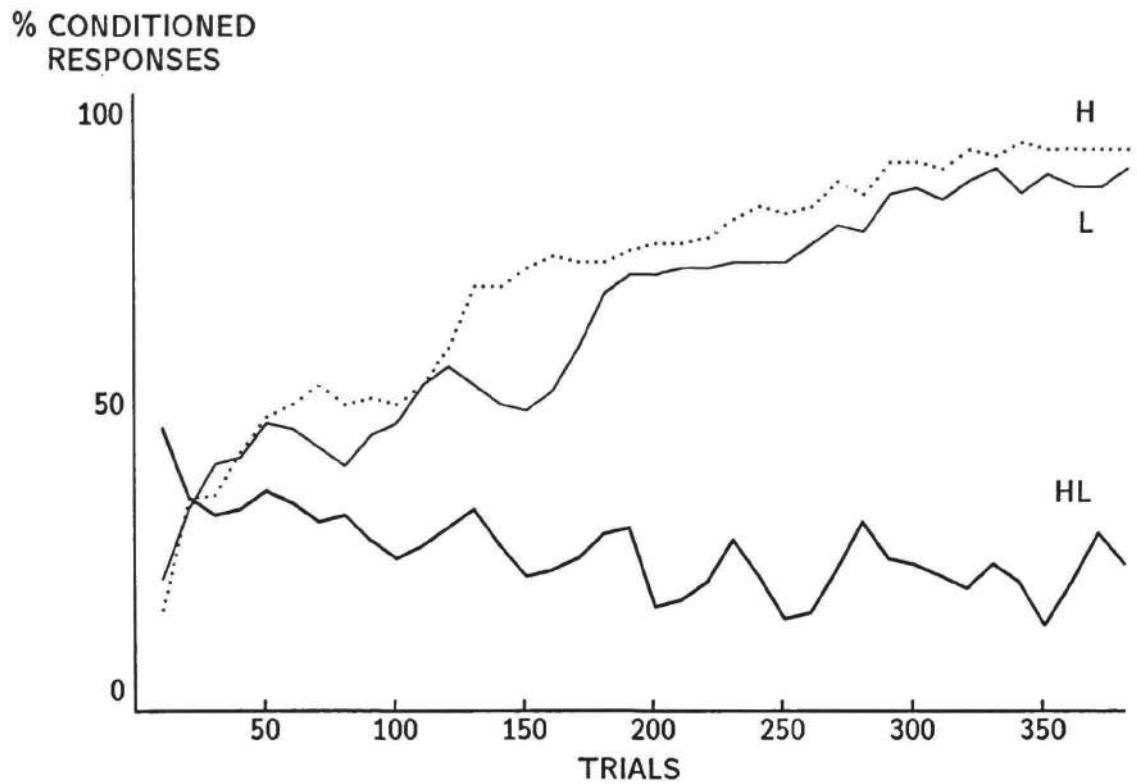


Figure 2: Conditioning to HL−, H+, L+.

mated previously, the *LS* and *LN* measures were unable to elect any effective features, for there was an equal amount of positive and negative infirming evidence for each of the cues. Therefore, exploration for predictive features occurred without Bayesian guidance. On an unexpectedly reinforced trial, for instance, STAGGER added three new compound features: a new disjunctive feature composed of a randomly elected, present feature and a randomly elected, absent feature; a new conjunctive feature made of a pair of randomly elected, present features; and a negation of a randomly elected, absent feature. This potential explosion of exploratory features was still subject to pruning via the Bayesian measures, though, and if each did not outperform the components from which it was composed, it was pruned.

BICONDITIONAL DISCRIMINATION

Saavedra (1975) has also studied simultaneous configural conditioning. However, unlike Woodbury, in each configuration all stimuli were from different sensory modalities. Assuming the presence of resonant features arising from the co-occurrence of features is therefore less reasonable. Instead of only two features, she utilized four, in pairwise configurations such that each feature was present in reinforced as well as unreinforced trials. No property of the features such as "two-ness" would aid in predicting reinforcement. One experimental group was given reinforced presentations of a tone (auditory cue A_1) and flickering light (L_1)

Table 6: Biconditional and component discrimination training.

A_1L_1 +, A_2L_2 +, A_1L_2 −, A_2L_1 −	Biconditional
A_1L_1 +, A_1L_2 +, A_2L_2 −, A_2L_1 −	Component

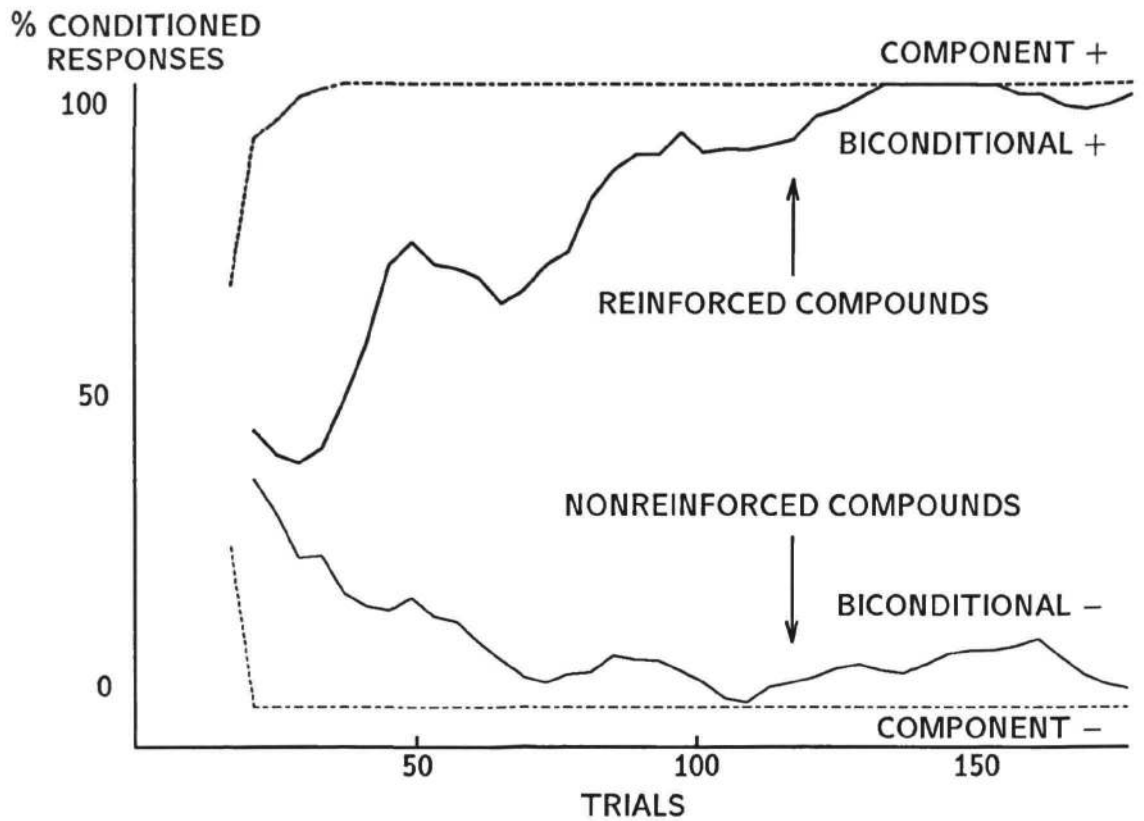


Figure 3: Biconditional versus component discrimination.

or a clicker (A_2) and a steady light (L_2). The alternate combinations were nonreinforced. This training is termed *biconditional discrimination* since reinforcement is conditional on two cues. For comparison, she also tested a simple component discrimination case where the tone was always reinforced. These training schedules are summarized in Table 6.

This experimental manipulation taxes the feasibility of a strict feature selection model since it is unlikely that the necessary resonant features are available; this class of models would predict that such an association would be unlearnable. Animal subjects, however, do learn the biconditional discrimination. Figure 3 overlays STAGGER's performance on both the biconditional and component discrimination cases. Each line represents the average percentage of CRs over ten separate program executions. The upper lines represent responding to the reinforced configurations; the lower lines, the unreinforced configurations. The solid lines represent conditioned responding in the biconditional discrimination case; the dashed lines correspond to component discrimination training.

The component discrimination training proceeds much more rapidly than the bicondi-

tional discrimination because the appropriate stimuli need only be selected in the former case, rather than formed, as in the latter case. In the biconditional discrimination task, STAGGER first forms the compound features AND[A_1,L_1] and AND[A_2,L_2] which are then used in the selection process. Without resonant features arising from the co-occurrence of A_1 and L_1, and A_2 and L_2, a strict feature selection model would be unable to acquire the biconditional discrimination. Associative strengths would have to be high enough for A_1 and L_1 to sum for a positive prediction when they occurred together, but low enough so that when A_1 and L_2 co-occurred, nonreinforcement would be expected. This is clearly impossible.

DISCUSSION

The two representative configural conditioning experiments of Woodbury (1943) and Saavedra (1975) indicate that animals are able to form associations between complex CSs and a US. Two categories of models have been proposed to account for this type of learning: feature selection only, and feature selection plus feature formation.

Feature selection models assume that the association accrued to a stimuli are summed when they co-occur (Rescorla & Wagner, 1972). The associative strength of a configuration of stimuli is simply the sum of the associative strengths of its components. A secondary assumption is that when two stimuli (say A and B) are present that a third *resonant* stimulus is present which has some of the properties of both (represented by AB); this AB stimulus is present always and only when both A and B are. A exclusive-disjunctive configuration like Woodbury's negative patterning is represented by a strong negative associative strength for the resonant feature and weaker positive associations for each of the components. This assumption extends the representational ability of feature selection models to include all possible Boolean functions.

There are two unsatisfactory consequences of assuming the presence of resonant features. First, while it seems plausible to assume that the simultaneous presence of a red light and a blue light adds a feature not present when either are presented separately (a purple light), this assumption seems tenuous when the stimuli are from different sensory modalities. Secondly, the number of stimuli from which the model must select grows exponentially with the number of stimuli that may be configurally associated. For example, if there are three stimuli (A, B, C), there must be four supplementary stimuli (AB, AC, BC, ABC) in order to select any configuration. Razran (1965) reports on experiments where six simultaneous features were conjunctively configured; 57 additional resonant features would be required if a feature selection model were applied. In general the number of resonant stimuli required by these models is $2^n - 1 - n$, where n is the number of perceptual, stimuli. Requiring the model to choose between $O(2^n)$ stimuli may be computationally infeasible.

The alternative we present here is a secondary process which formulates plausibly predictive compound features as they are needed. The number to be examined is therefore limited to those necessary and there is a corresponding reduction in computational load on

the feature selection process. The fact that both approaches rely on the ability to form an association to a combination of cues is not new. However, unlike the strict feature selection model, STAGGER employs a feature formation component which can be used to configure individually perceptible cues into an explicit compound feature usable for learning. Instead of assuming that this process is already performed by the perception system via resonance (and its entailing assumptions of cross-modality resonance and exponential requirements), we prefer the property of necessity-driven feature formation. In this we concur with Razran when he notes:

> What seems more warranted is the view that, inasmuch as configures are formed and deformed through learning, their role is much more a function of the organism's conditioned past than of its sensory present, and, moreover, that their learning reveals the dynamic essence of their "becoming" if not also of their being (Razran, 1965, p. 244, fn. 3).

FUTURE WORK

One phenomenon unexplained by previous work on feature selection is that of sequential configural conditioning. In these types of experiments, effectively predicting the US requires discerning a sequential configuration of the cues in the environment. Woodbury (1943) also trained dogs to expect a food pellet only when the low buzzing sound followed the high buzzing sound. While we have provided some explanation of mechanisms which could give rise to simultaneous configural conditioning, we have yet to address the larger issue of associating sequences with outcomes. We believe that a featural formation approach, where sequences are constructed and their effectiveness evaluated, will prove useful.

ACKNOWLEDGEMENTS

This research was supported in part by the Office of Naval Research under grants N00014-84-K-0391 and N00014-85-K-0854, the National Science Foundation under grants IST-81-20685 and IST-85-12419, the Army Research Institute under grant MDA903-85-C-0324, and by the Naval Ocean Systems Center under contract N66001-83-C-0255. We would like to thank Michal Young for his early involvement with this project and the machine learning group at U.C. Irvine, who collectively have helped clarify a number of interesting issues and have provided a supportive environment for the exploration of new ideas.

REFERENCES

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1965). *A study of thinking.* New York: John Wiley & Sons, Inc.

Colwill, R. M., & Rescorla, R. A. (in press). Associative structures in instrumental learning. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory.* New York: Academic Press.

Duda, R., Gaschnig, J., & Hart, P. (1979). Model design in the Prospector consultant system for mineral exploration. In D. Michie (Ed.), *Expert systems in the micro electronic age.* Edinburgh: Edinburgh University Press.

Fitzgerald, R. D. (1963). Effects of partial reinforcement with acid on the classically conditioned salivary response in dogs. *Journal of Comparative and Physiological Psychology, 56,* 1056–1060.

Granger, R. H., Jr., & Schlimmer, J. C. (1985b). Learning salience among features through contingency in the CEL framework. *Proceedings of the Seventh Annual Conference of the Cognitive Science Society* (pp. 65–79). Irvine, California: Lawrence Erlbaum Associates.

Grings, W. W. (1972). Compound stimulus transfer in human classical conditioning. In A. Black & W. F. Prokasy (Eds.), *Classical conditioning II.* New York: Appleton-Century-Crofts.

Hampson, S., & Kibler, D. (1983). A Boolean complete neural model of adaptive behavior. *Biological Cybernetics, 49,* 9–19.

Mitchell, T. M. (1982). Generalization as search. *Artificial Intelligence, 18,* 203–226.

Razran, G. (1965). Empirical codifications and specific theoretical implications of compound-stimulus conditioning: Perception. In W. F. Prokasy (Ed.), *Classical conditioning.* New York: Appleton-Century-Crofts.

Rescorla, R. A. (1968). Probability of shock in the presence and absence of CS in fear conditioning. *Journal of Comparative and Physiological Psychology, 66,* 1–5.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. Black & W. F. Prokasy (Eds.), *Classical conditioning II.* New York: Appleton-Century-Crofts.

Saavedra, M. A. (1975). Pavlovian compound conditioning in the rabbit. *Learning and Motivation, 6,* 314–326.

Schlimmer, J. C. (1986). *A comparison of feature salience measures* (Technical report #86-99). Irvine, California: The University of California, Department of Information and Computer Science.

Wasserman, E. A., Chatlosh, D. L., & Neunaber, D. J. (1983). Perception of causal relations in humans: factors affecting judgments of response-outcome contingencies under free-operant procedures. *Learning and Motivation, 14,* 406–432.

Whitlow, J. W., Jr., & Wagner, A. R. (1972). Negative patterning in classical conditioning: Summation of response tendencies to isolable and configural components. *Psychonomic Science, 27,* 299–301.

Woodbury, C. B. (1943). The learning of stimulus patterns by dogs. *J. comp. Psychol., 35,* 29–40.

A Layered Network Model for Learning-to-Learn and Configuration in Classical Conditioning*

E. James Kehoe
School of Psychology, University of New South Wales

ABSTRACT

Networks composed of layers of adaptive elements provide a rigorous explanation for complex associative learning phenomena. In particular, a network composed of three adaptive elements can explain previously intractable phenomena, namely the rapid rate of reacquisitions, learning-to-learn, spontaneous configuration, and negative patterning (the exclusive-OR problem). This paper will compare the results of computer simulations to the behavioral results of classical conditioning experiments using the rabbit's nictitating membrane response.

INTRODUCTION

Layered networks of adaptive elements have featured prominently in contemporary theories of biological and machine cognition, particularly with regards to pattern recognition (Barto & Anderson, 1985; Feldman, 1985). Most notably, layered networks provide an elegant means for solving problems of nonlinear representation, for example, the exclusive-OR problem in which the system must learn to respond to each of two inputs but not their conjunction (Rumelhart, Hinton, & Williams, 1985). Rather than being "prewired" to represent particular combinations of inputs, layered adaptive networks of the appropriate type possess the ability to "tune" themselves to significant combinations of inputs (e.g., Barto, 1984; Barto, Anderson, & Sutton, 1982; Rumelhart et al., 1985). A less widely noted feature of layered networks is their ability to explain "learning-to-learn," that is a gain in the flexiblity of the system's output as a consequence of prior training. At a more general level, a capacity for learning-to-learn may provide the

*This research was supported by Grant A28315236 from the Australian Research Grants Committee.

foundation for "insight" and other forms of "understanding" (Harlow, 1949). Accordingly, a major purpose of this paper is to demonstrate that some of the same features of layered networks that permit the recognition of arbitrary patterns also permit learning-to-learn.

This paper will present computer simulations of layered networks that are intended to duplicate the course of associative learning in a biological system, namely classical conditioning of the rabbit's nictitating membrane (NM) response (Gormezano, 1966; Gormezano, Kehoe, & Marshall, 1983). In the NM response preparation, the measured response is an extension of the third eyelid, which is innately elicited as an unconditioned response (UR) by a brief (50 ms) electrical-pulse unconditioned stimulus (US) administered to the surface of the skin posterior to the eye. Learning is produced by sequential presentations of a conditioned stimulus (CS) and the US, and, after a number of CS-US pairings, learning is evidenced by the acquisition of an NM conditioned response (CR) to the CS in advance of the US.

As a biological testbed for layered network models, classical conditioning procedures have several useful features:

(1) Animals can be brought to the learning situation in a relatively naive state, thus approximating the initial state of an untutored network.
(2) Animals do not require any prior verbal instructions, thus learning proceeds as a function of the stimulus inputs and response outputs that occur during the training task.
(3) In classical conditioning procedures, it is possible to pose learning problems in a simplified way that can be duplicated in simple layered networks. For example, the exclusive-OR problem has its behavioral counterpart in Pavlov's (1927) negative patterning task. In that task, the animal is presented a mixture of three types of learning trials: (a) a tone CS that signals the US, (b) a light CS that also signals the US, and (c) a

compound tone + light stimulus that never signals the US. The animal can be said to have solved the negative patterning task when it generates CRs to the tone and the light but not the compound.

(4) In many classical conditioning procedures, the CR appears to the CS in advance of the arrival of the US. By observing these anticipatory CRs, it is possible to trace the course of learning on a trial-by-trial basis. For purposes of testing a network model, the eventual achievement of a solution is perhaps less interesting than observing the intermediate states of the system prior to the solution state. For example, in solving the negative patterning problem, animals initially show considerable CR acquisition to the compound as well as the separate tone and light stimuli, after which responding to the compound gradually declines (e.g., Bellingham, Gillette-Bellingham, & Kehoe, 1985; Whitlow & Wagner, 1972; Woodbury, 1943).

In the remainder of this presentation, I shall describe in three stages a model of classical conditioning based on a layered network scheme. The model originates in those of Barto, Sutton, and their associates, which in turn are based on Hebb's (1949, 1972) theory of synaptic facilitation (e.g., Barto, 1984; Barto et al., 1982; Sutton & Barto, 1981). In brief, each stage of the model will encompass an increasing number of conditioning phenomena. The first-stage model will explain simple CR acquisition to one CS and a primitive form of learning-to-learn evidenced by progressive increases in the rate of successive acquisitions and extinctions conducted with the same CS (Hoehler, Kirschenbaum, & Leonard, 1973; Scavio & Thompson, 1979; Schmaltz & Theios, 1972; Smith & Gormezano, 1965). The second-stage model will encompass a more advanced form of learning-to-learn, namely a facilitation of CR acquisition to a new CS (e.g., light) after prior training with another, highly distinctive CS (e.g., tone) (Holt & Kehoe, 1985; Kehoe & Holt, 1984). Finally, the

third-stage model will explain a variety of simple pattern recognition phenomena, including negative patterning.

STAGE I: SUCCESSIVE ACQUISITIONS AND EXTINCTIONS

Figure 1 shows a schematic diagram of the network. The network contains two "sensory" elements, one for the tone CS (T) and one for the unconditioned stimulus (US). The output from T projects to an intermediate element (X), and the output from X projects to another element (R), which in turn gives rise to the observable behavior (CR/UR). Both nonsensory elements, namely X and R, receive an output from the US element.

Initial CR Acquisition

At the beginning of training with a naive animal, only the outputs from the US to X and R are effective. That is to say, only the US element can trigger an all-or-none firing of X and R. Initially, the T input is unable to trigger the intermediate element, but the T input does render its connection with X eligible for modification by the US input should it occur during a brief eligibility period that follows CS onset (Sutton & Barto, 1981). Thus, as the T-X connection strengthens over successive CS-US pairings, T will begin to trigger X. Then, the X-R connection will become eligible for change by the US's input to R. Observable CRs to the tone will only begin to appear when the intervening connections become strong enough so that T triggers X and then X triggers R. The changes in each of the interior connections, namely T-X and X-R, are governed by the linear operator process commonly used in current models of conditioning (Sutton & Barto, 1981). (See Appendix 1 for a full description of the implementation of the model). The firing of both X and R is all-or-none and is determined by a normally-distributed random threshold variable. Thus, on a given trial, X fires in response to an input from T only if the T-X connection weighting exceeds the threshold value on that trial. Likewise, R fires only if the X-R connection weight exceeds the current threshold.

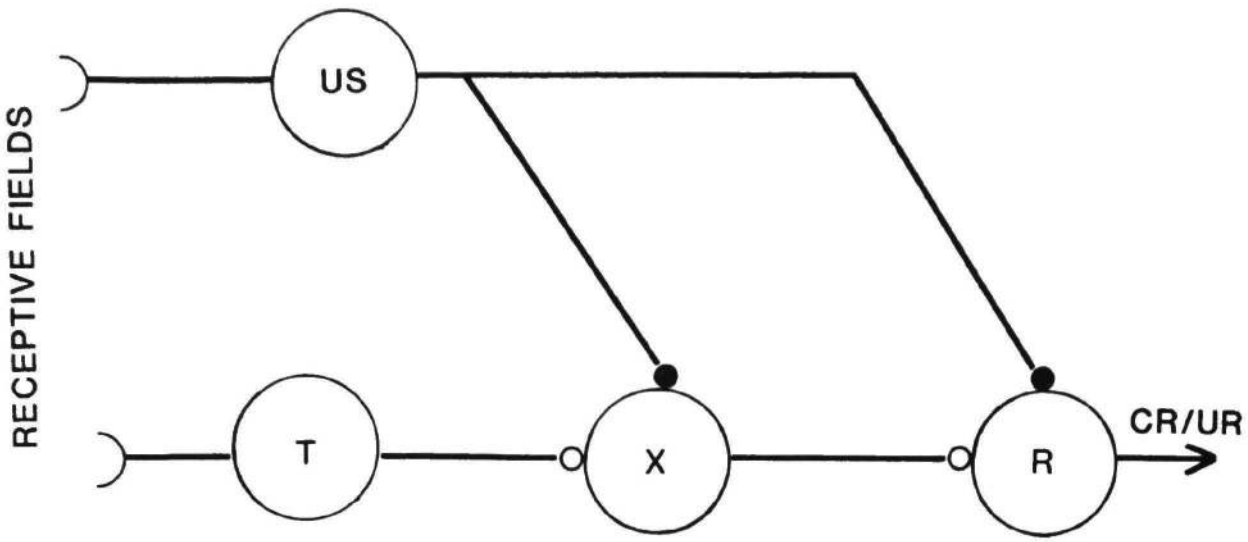


Figure 1. A minimal network of two sensory elements (T, US) and two adaptive elements (X, R) for successive acquisitions and extinctions in classical conditioning.

Figure 2's left-hand column of panels shows the changes across blocks of CS-US trials in (a) the T-X connection, (b) the X-R connection, and (c) the percent CR measure produced by a computer simulation of the network's activities. As can be seen in the bottom panel, it is possible to reproduce a typical acquisition curve. The thresholds and growth rate parameters for both connections were selected so that the simulated curve would approximate the acquisition curve typically obtained in the rabbit NM response preparation under an 800-ms CS-US interval (see Appendix 1). As can be seen in the upper two panels, the T-X connection rises to a high level before the X-R connection shows any substantial change. For example, in the second block of training, the T-X connection was .69 while the X-R connection was only .10.

Subsequent Acquisitions

In its remaining panels, Figure 2 shows the simulated changes for the interior connections and percent CR across an initial extinction, a reacquisition, and a re-extinction. During

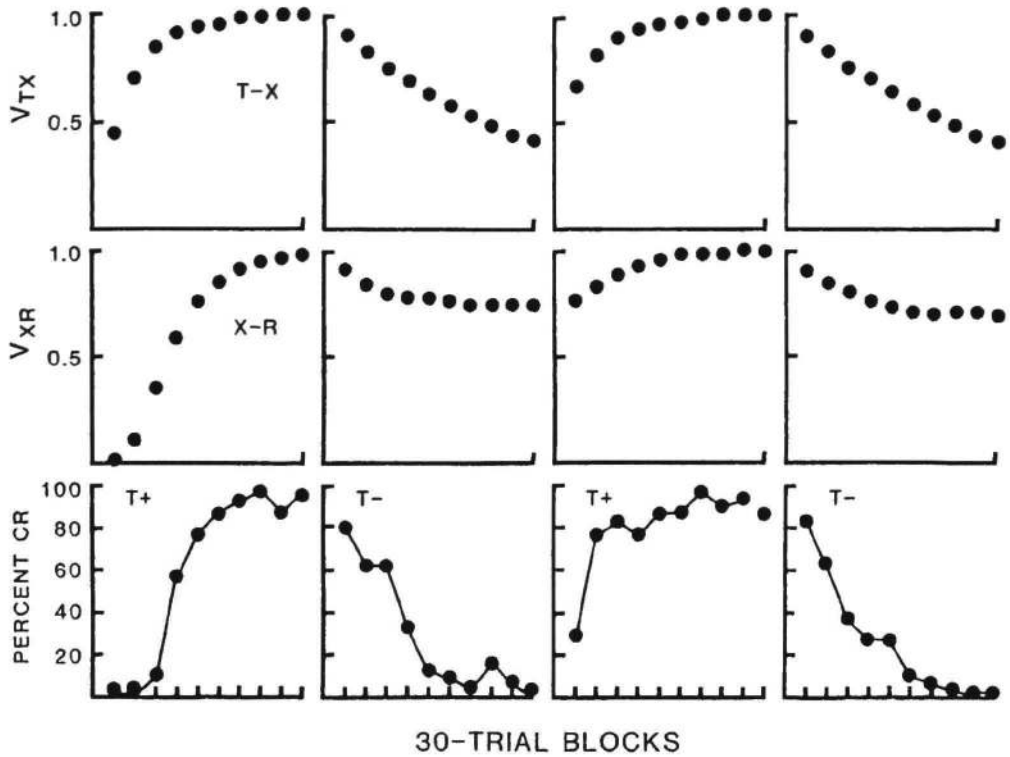


Figure 2. Simulation results for successive acquisitions and extinctions.

---

the initial extinction, the T-X connection declines at a steady rate, while the X-R connection declines to an asymptotic level of .70. As the T-X connection weakens and X's frequency of firing declines, the X-R connection is eligible for modification less and less often. In this way, the X-R connection is largely protected from extinction and thus remains intact. With respect to the simulated percent CR, it can be seen that the CR frequency reaches negligible levels while both the T-X and X-R connections are still appreciable. Consequently, during reacquisition in the third stage, both the T-X and X-R connections need relatively few reinforcements to rise to their asymptotic levels, yielding a relatively rapid rise in CR likelihood.

The second extinction does not appear particularly more rapid than the first extinction. To some extent, this simulated outcome is accurate; the available data suggest that the change in extinction rate is considerably slower than the change in acquisition rate across alternations of the training conditions.

Thus, this version of the model appears to be accurate to at least a first approximation.

STAGE II: LEARNING-TO-LEARN

Figure 3 shows an example of a learning-to-learn effect that has been repeatedly observed in conditioning of the rabbit NM response (Holt & Kehoe, 1985; Kehoe & Holt, 1984; Kehoe, Morrow, & Holt, 1984). In Phase I of this particular experiment, one group of rabbits received CS-US pairings in which the initial CS was an 800-ms tone. Another group served as a rest control. As

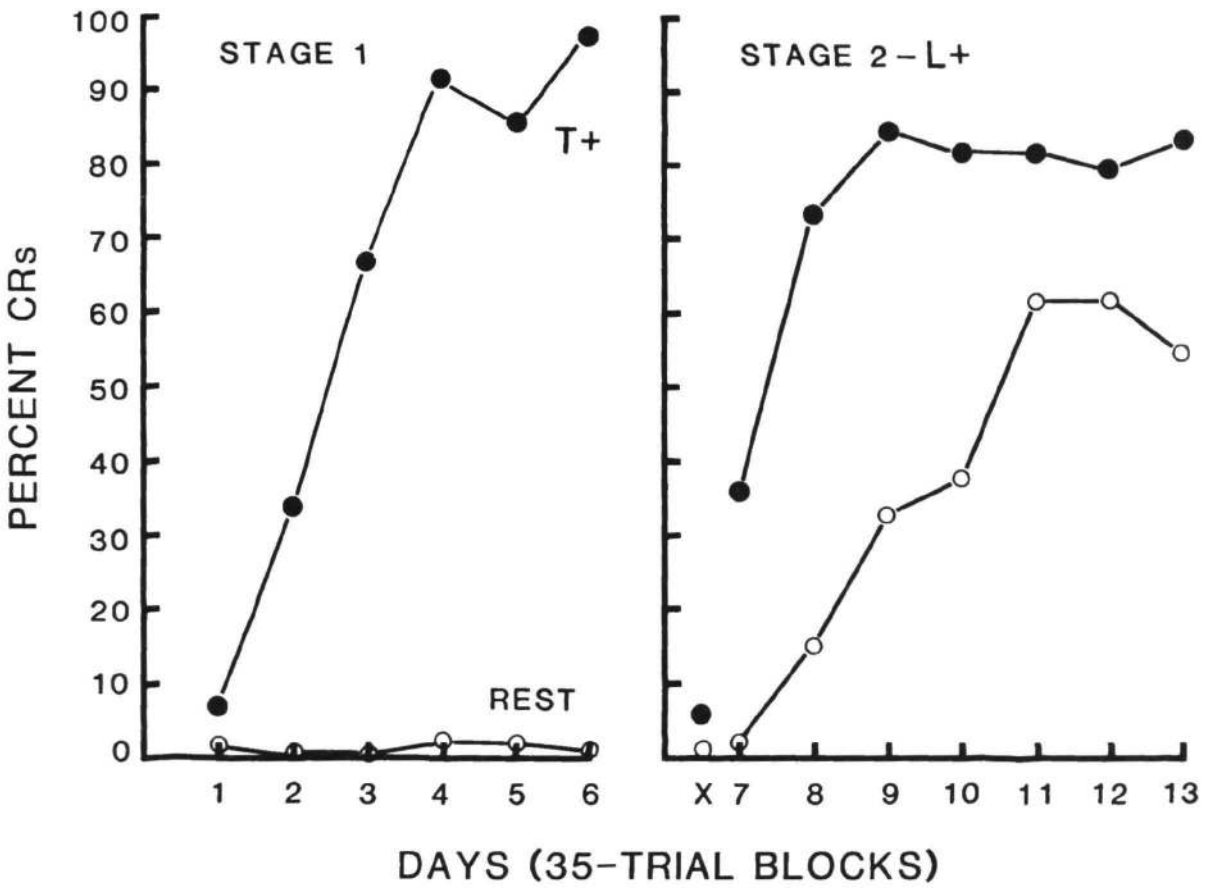


Figure 3. Example of initial CR acquisition to a tone CS (T+) and subsequent transfer to a light CS (L+). The point marked X indicates the initial response levels to the light CS.

expected, the former group showed CR acquisition to the tone, while the rest control group showed negligible responding. At the start of Phase II, both groups received four unreinforced presentations of an 800-ms light to determine the level of immediate cross-modal generalization. In the present experiment as in all our other studies, immediate transfer was not detectable. In Figure 3, the mean response likelihood on the tests are shown above the "X" marker on the abscissa. Only one animal, which happened to be in the pretrained group, responded twice to the light. However, once CS-US training with light was begun, the pretrained group showed extremely rapid CR acquisition to the new CS. For example, the animals in the pretrained group achieved a mean CR likelihood of 36% within the first block of reinforced light trials. By way of comparison, the naive animals in the control group achieved a mean CR likelihood of only 2% within the first block of reinforced light trials. In the present case, the pretrained group showed a higher level of responding to the light than the control group throughout Phase II. However, the asymptotic level of responding in the control group usually converges with that of the pretrained group. In other studies, we have shown that the positive transfer between tone and light is symmetric.

In order to explain the cross-modal learning-to-learn effect, it is only necessary to add an additional sensory element for the light to the network, as can be seen Figure 4. The input from light (L) projects to the intermediate element X just as the input from tone (T) does. Nothing else about the model is changed in any way.

Figure 5 shows the results of computer simulations for the learning-to-learn effect. The simulation of initial CR acquisition with the tone proceeds in the normal way for the model. In particular, observable CRs to the initial CS (tone) wait upon the successive strengthening of the T-X and X-R connections. However, in subsequent reinforced training with the new CS (light), the appearance of CRs requires only the establishment of the L-X connection, because the X-R connection has been already fully strengthened. Thus, as soon as the L-X

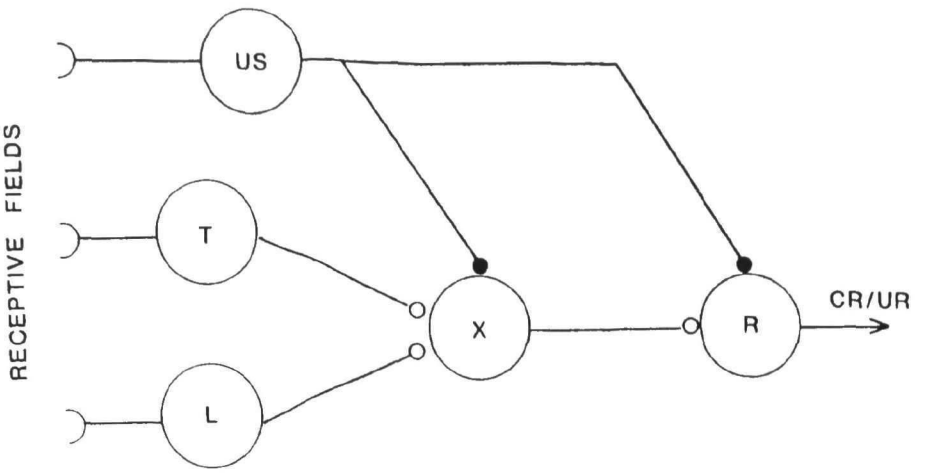


Figure 4. A minimal network of three sensory elements (T, L, US) and two adaptive elements (X, R) for learning-to-learn.

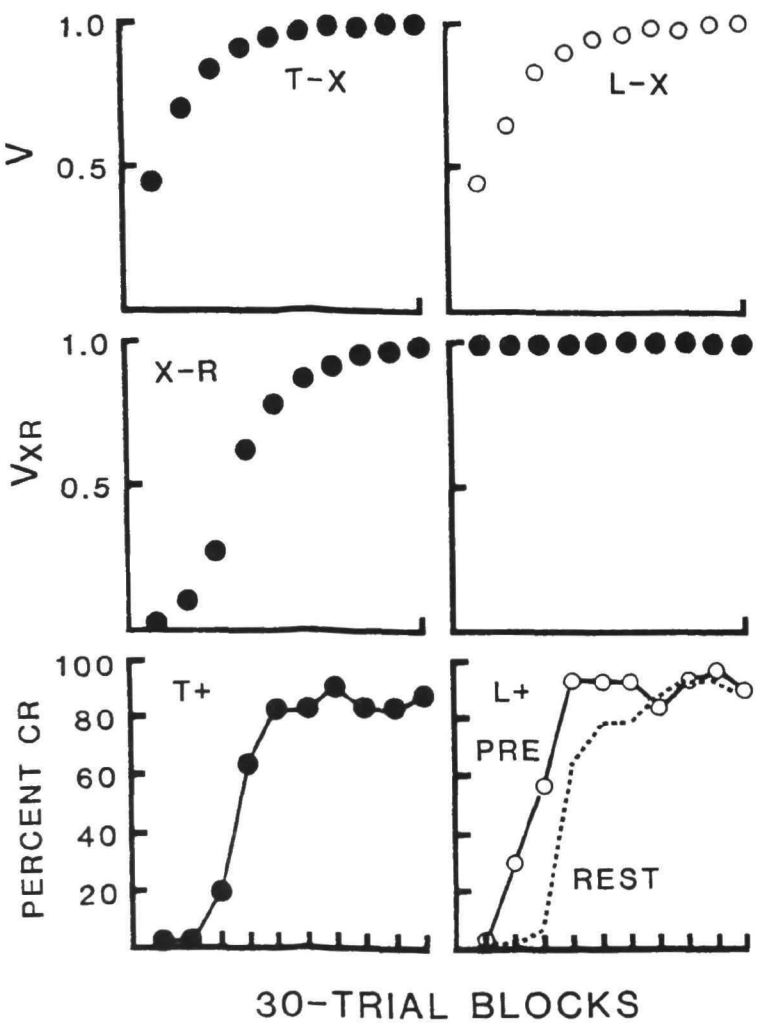


Figure 5. Simulation results for learning-to-learn.

connection becomes strong enough to trigger X, any firings of X triggered by L are immediately translated into observable CRs via the previously-established X-R connection. The rapid CR acquisition to the light is displayed in the learning curve for Phase II labelled as PRE, which denotes pretraining. In the way of contrast, a learning curve for a naive control condition is also displayed, labelled as REST. Thus, by relying on a common connection and the combination of convergent CS inputs, a layered network can explain the learning-to-learn effect.

In addition to demonstrating the learning-to-learn effect, my associates and I have found that it survives extinction of the original conditioned reflex (Kehoe et al., 1984). Figure 6 shows the results of one of our experiments. The key experimental group (4-E) received initial training at a 400-ms CS-US interval with one CS (CS1-US). Between CS1-US training and transfer training with a second CS (CS2-US), the animals in Group 4-E received a CS1-alone extinction procedure. Another experimental group (4-H) remained in their home-cages during the extinction procedure. In addition, two corresponding control groups (28-E and 28-H) initially received exposure to CS1 and the US but at a long 2,800-ms CS-US interval. Examination of the left-hand panel of Figure 6 reveals that Groups 4-E and 4-H showed conventional CR acquisition, while Groups 28-E and 28-H showed negligible levels of responding. The middle-panel shows that the Group 4-E displayed considerable extinction of the CR to CS1, whereas Group 28-E continued to display little responding. Finally, the right-hand panel reveals that, despite the near elimination of the initial conditioned reflex (CS1-CR) in Group 4-E, those animals acquired the new conditioned reflex (CS2-CR) as rapidly as their counterparts in Group 4-H, both of which showed positive transfer relative to their respective control groups.

On the theoretical side, the computer simulations successfully reproduced the ability of the learning-to-learn effect to survive disruption of the initial conditioned reflex. According to the computer simulations, the learning-to-learn effect survives for the same reasons that reacquisition after extinction is more rapid than initial acquisition. Figure 7

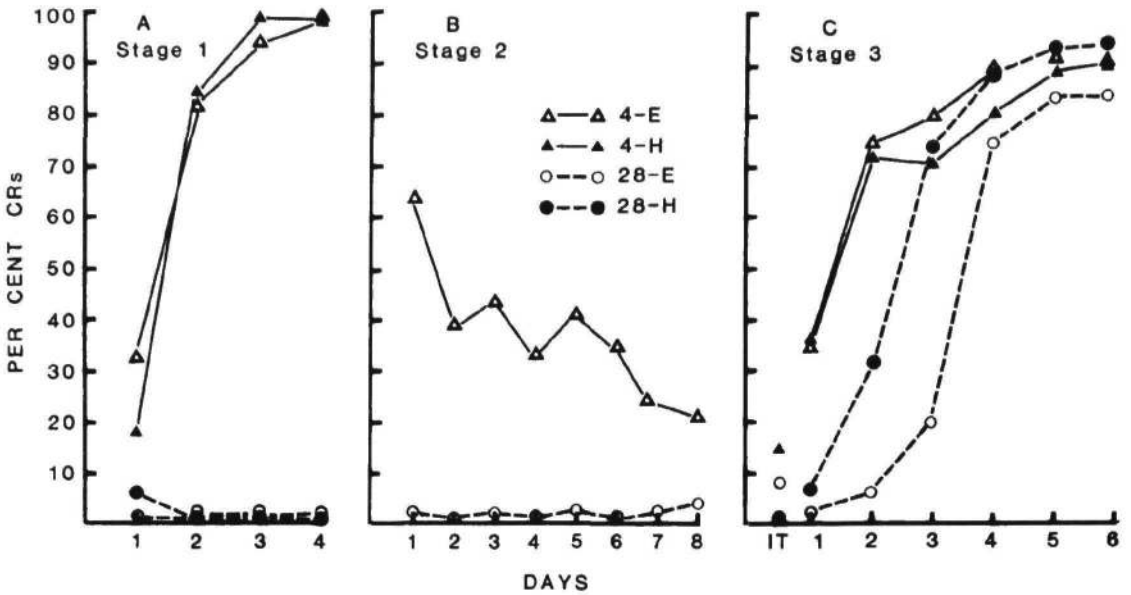


Figure 6. Learning-to-learn in Group 4-E survived extinction of the initial CR (Kehoe, Morrow, & Holt, 1984).

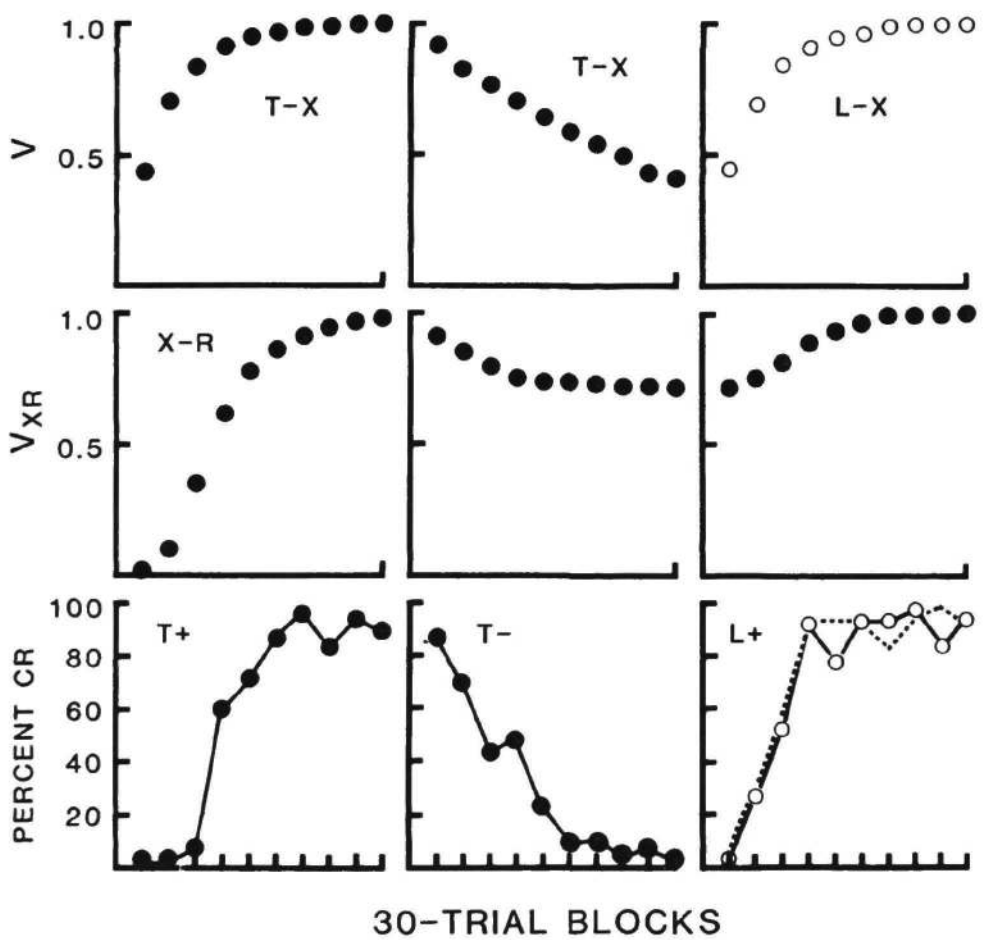


Figure 7. Simulation results for acquisition, extinction, and transfer training.

shows the results of computer simulations for the case in which there is an intervening extinction of the original conditioned reflex. As shown in Figure 7, the X-R connection is largely intact at the end of tone extinction. With the X-R connection still in place, pairings of the alternate CS (L) with the US can take advantage of the X-R connection and rapidly produce CRs as the L-X connection begins to strengthen. The lower right-hand panel of Figure 7 shows two simulated percent CR curves. The solid line represents acquisition to the light in the group that received tone pretraining followed by tone extinction (i.e., Group 4-E). The dotted line represents the simulated acquisition curve from a pretrained group that did not undergo extinction of the original conditioned reflex (i.e., Group 4-H). In agreement with the behavioral data, the two curves overlap perfectly.

STAGE III: CONFIGURAL LEARNING

The rabbit NM response preparation has expressed its sensitivity to patterns of multiple sensory inputs in a variety of ways. Figure 8 shows the course of differentiation between a compound and its components under three different training regimes. The lower panel shows the learning curves obtained under a negative patterning schedule, which corresponds to the exclusive-OR problem. As can be seen, differentiation proceeded slowly; responding to the compound, which was never followed by the US, declined only after extensive training (Bellingham et al., 1985). The upper right-hand panel reveals that differentiation proceeded much more rapidly in a positive patterning procedure, in which reinforced presentations of a tone + light compound (TL+) were intermixed with unreinforced presentations of the separate components (T-, L-) (Bellingham et al., 1985; Kehoe & Schreurs, in press). In logical terms, the positive patterning schedule corresponds to an AND problem. Differentiation of a compound from its components is not confined to procedures entailing explicit discrimination training. As shown in the upper left-hand panel, pairings of a compound with the US can produce spontaneous differentiation of the compound

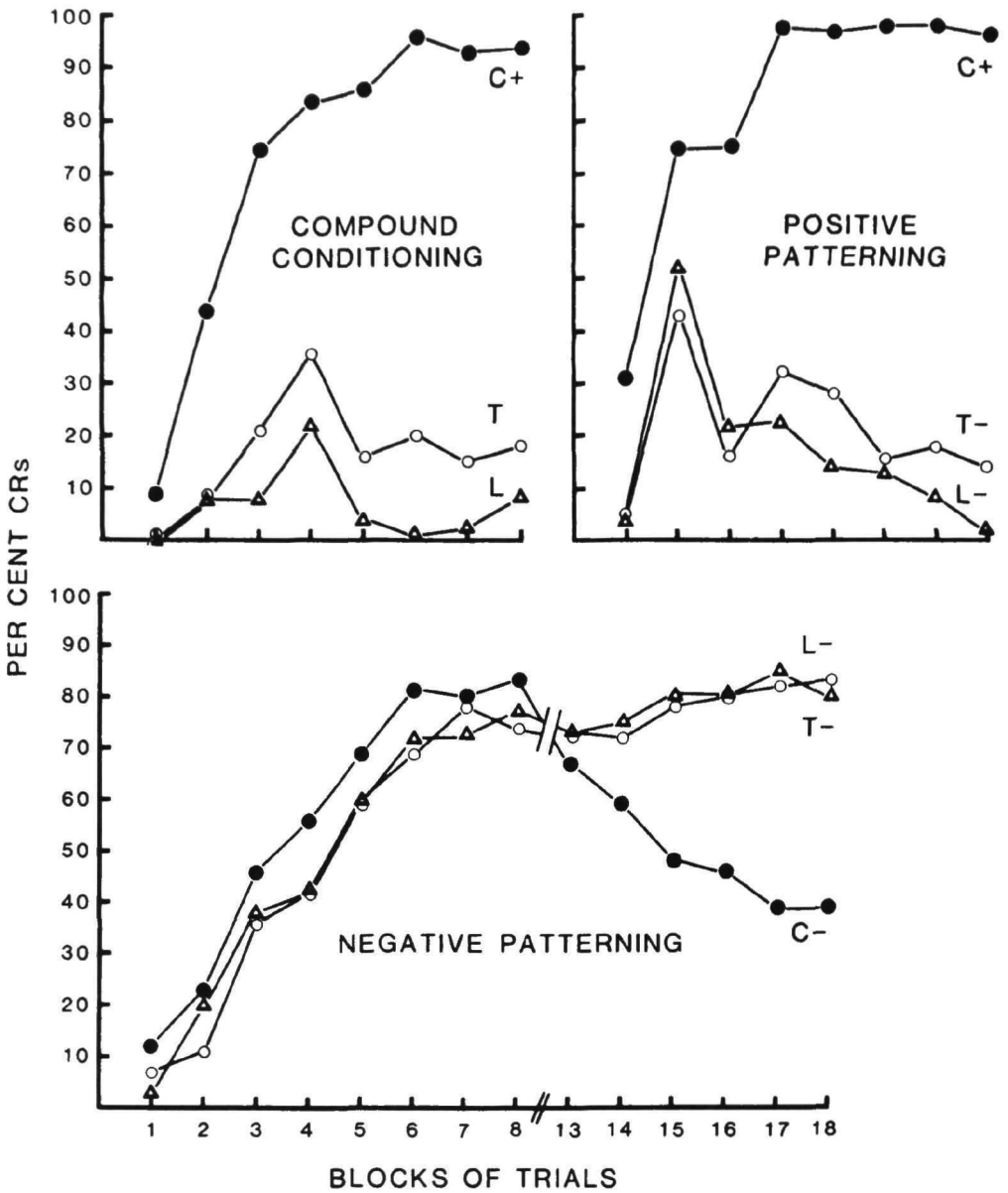


Figure 8. Examples of CR acquisition to a tone, light, and compound (tone + light) in compound conditioning, positive patterning, and negative patterning procedures.

---

from its components (Kehoe, 1986; Kehoe & Schreurs, in press). On the basis of both explicit and implicit differentiation between a compound and its components, numerous theorists have proposed that the nervous system establishes distinctive encodings for the compound and its components, each with its own

excitatory or inhibitory associative strength (e.g., Bellingham et al., 1985; Hull, 1943, 1945; Kehoe & Gormezano, 1980; Razran, 1965, 1971; Rescorla, 1972, 1973; Whitlow & Wagner, 1972). While negative patterning clearly represents a nonlinear combination of the components, positive patterning and spontaneous differentiation may represent cases in which the CR-evoking capacity of the compound results from a linear combination of the CR-evoking capacities of the separate components, tone and light. Nevertheless, a history of reinforced exposure to a compound stimulus engages a combination process, linear or otherwise, that permits the subject to respond to the compound as a functional unit distinct from its components.

Figure 9 shows a schematic diagram of a network that can explain the configural learning phenomena. The network is essentially two parallel instances of the network used in the Stage II model. That is to say, the sensory inputs for tone, light, and the US project to a second intermediate element (Y), which in turn projects to the R element. For purposes of triggering an element by a joint input, it was assumed that the sum of currently eligible connection weights is compared to the element's threshold value. For changing the input weights, the present model followed the lead of Sutton and Barto (1981). In

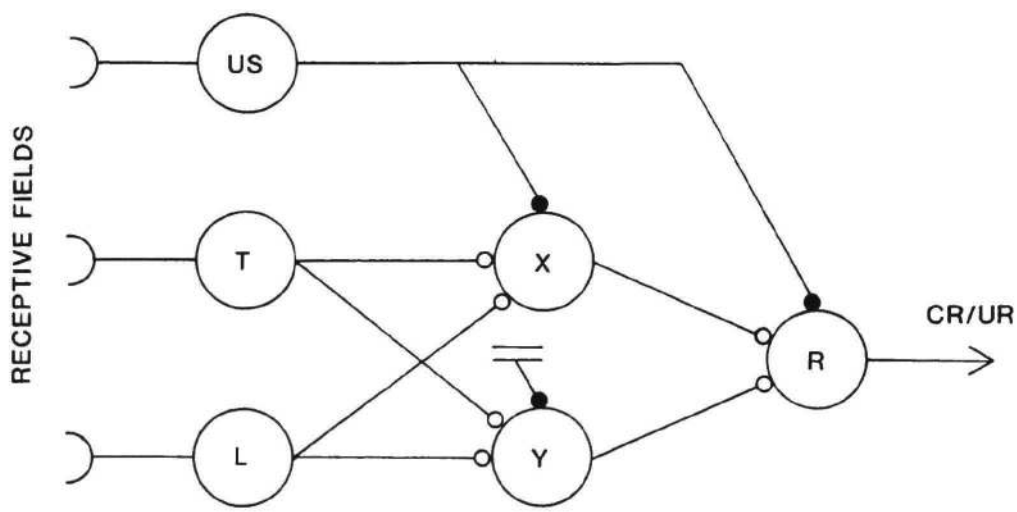


Figure 9. A network of three sensory elements (T, L, US) and three adaptive elements (X, Y, R) for configural learning.

brief, when two inputs to either the X, Y, or R elements were simultaneously eligible for modification, then the inputs competed for the available connection weights supported by the US input. Thus, in training with a single CS, say the tone, the inputs from X and Y to the R element would compete with each other. In compound training with the tone and light, the inputs from T and L to the X element would compete with each other. Likewise, the T and L inputs to the Y element would compete with each other.

This competitive process was originally formulated to account for stimulus selection phenomena, in which increases in the weight of one stimulus input would be either blocked by prior increases in the weight of another concurrent input or overshadowed by more rapid increases in the weight of another concurrent input (e.g., Rescorla & Wagner, 1972; Sutton & Barto, 1981). However, this competitive process can also cause elements to become tuned to the combined T and L inputs. Specifically, competition between the T and L inputs would ensure that neither input by itself would gain sufficient connective weight to be able to reliably trigger the next element.

In order to discover a set of parameters that would accurately simulate configural learning, I manipulated two groups of parameters, namely the mean threshold value of each element (Tj) and the learning rate parameter for each element (aj). Figure 10 shows the results of using the Stage III network to simulate the results of the compound stimulus experiments. The curves were obtained when (1) the X element had a higher learning rate than that of the Y element (ax = .100, ay = .001) and (2) the X element had a higher mean threshold than the Y element (Tx = .65, Ty = .25).

Inspection of Figure 10 reveals that the proposed model was able to simulate (a) the virtually complete differentiation between the compound and its components in compound conditioning, and (b) the relatively slow acquisition of negative patterning characterized by an initial rise in responding to the unreinforced compound stimulus followed by a gradual decline. The only detail missing in the simulations was the initial rise

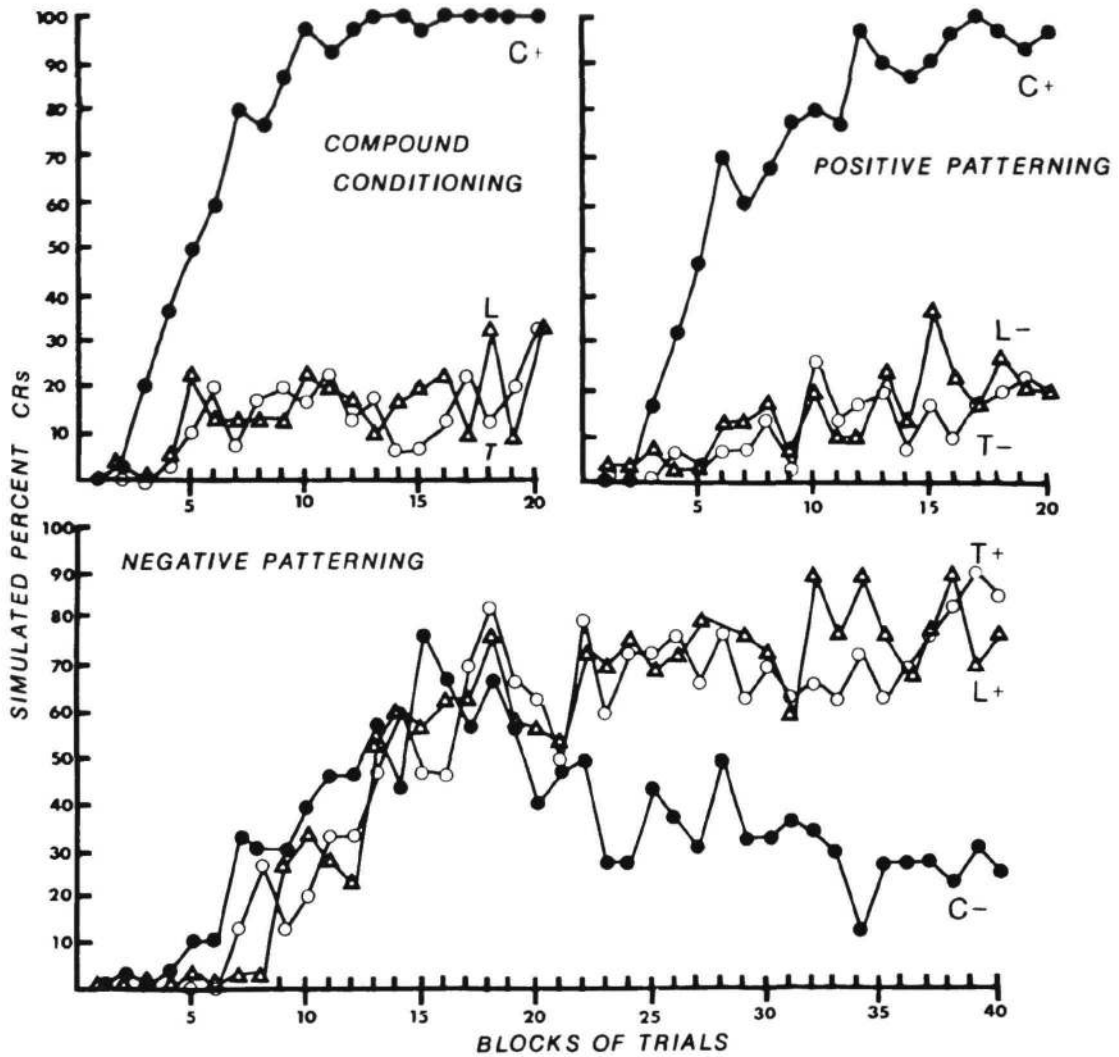


Figure 10. Simulations of compound conditioning, positive patterning, and negative patterning.

in responding to the unreinforced components in the positive patterning schedule. However, in the rabbit NM response preparation, such a rise does not appear to be a universal feature of positive patterning (Bellingham et al., 1985). While the Stage III network was constructed to simulate the quantitative outcomes of the compound conditioning and patterning schedules, further simulation runs using the same parameter values have indicated that the Stage III network retains the basic characteristics and limitations of the competitive models

(Rescorla & Wagner, 1972; Sutton & Barto, 1981). In particular, the Stage III network reproduces blocking and conditioned inhibition, while being unable to generate latent inhibition.

DISCUSSION

The present simulations reveal that layered network models have considerable scope for explaining a variety of distinctive learning phenomena that have previously proved intractable to rigorous, elegant explanation. In particular, the network model not only reached the same endpoints as the behavioral phenomena but followed much the same course of acquisition. Although the particular network model used in this presentation was intended to be as a nonspecialized as possible, it is nevertheless only an example of a larger class of models. For purposes of explaining rapid reacquisition and learning-to-learn phenomena, a large variety of layered network structures would be suitable. In an earlier version of the Stage II model, there were direct connections from each sensory input to the R element as well as to the intermediate element X. Simulations of that model revealed that it too could generate rapid reacquisition and learning-to-learn. These same phenomena should also appear under a huge range of rules for the adaptive elements, provided that the interior connections (e.g., X-R) do not change considerably faster than the connections from the sensory elements to the interior elements (e.g., T-X). However, successful simulation of the configural learning phenomena may occur only under a narrow range of network structures and adaptive rules, because these phenomena require a more delicate balancing of acquisition rates and threshold values to yield the appropriate connection weights.

REFERENCES

Barto, A. G. (Ed.) (1984). Simulation experiments with goal-seeking adaptive elements. (AFWAL-TR-84-1022). Wright-Patterson AFB, OH: Avionics Laboratory, Air Force Wright Aeronautical Laboratories.

Barto, A. G., & Anderson, C. W. (1985). Structural learning in connectionist systems. Paper presented to the Seventh Cognitive Science Conference.

Barto, A. G., Anderson, C. W., & Sutton, R. S. (1982). Synthesis of nonlinear control surfaces by a layered associative search network. Biological Cybernetics, 43, 175-185.

Bellingham, W. P., Gillette-Bellingham, K., & Kehoe, E. J. (1985). Summation and configuration in patterning schedules with the rat and rabbit. Animal Learning and Behavior, 13, 152-164.

Feldman, J. A. (Ed.) (1985). Special issue on connectionist models and their applications. Cognitive Science, 9.

Gormezano, I. (1966). Classical conditioning. In. J. B. Sidowski (Ed.), Experimental methods and instrumentation in psychology (pp. 385-420). New York: McGraw-Hill.

Gormezano, I., Kehoe, E. J., & Marshall, B. S. (1983). Twenty years of classical conditioning research with the rabbit. In J. M. Sprague and A. N. Epstein (Eds.), Progress in psychobiology and physiological psychology: Vol. 10 (pp. 197-275). New York: Academic Press.

Harlow, H. F. (1949). The formation of learning sets. Psychological Review, 56, 51-65.

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

Hebb, D. O. (1972). Textbook of psychology (3rd Edition). Philadelphia: W. B. Saunders.

Hoehler, F. K., Kirschenbaum, D. S., & Leonard, D. W. (1973). The effects of overtraining and successive extinctions upon nictitating membrane conditioning in the rabbit. Learning and Motivation, 4, 91-101.

Holt, P. E., & Kehoe, E. J. (1985). Cross-modal transfer as a function of similarities between training tasks in classical conditioning of the rabbit. Animal Learning and Behavior, 13, 51-59.

Hull, C. L. (1943). Principles of behavior. New York: Appleton-Century-Crofts.

Hull, C. L. (1945). The discrimination of stimulus configurations and the hypothesis of neural afferent interaction. Psychological Review, 52, 133-139.

Kehoe, E. J. (1986). Summation and configuration in conditioning of the rabbit's nictitating membrane response to compound stimuli. Journal of Experimental Psychology: Animal Behavior Processes, 12, 186-195.

Kehoe, E. J., & Gormezano, I. (1980). Configuration and combination laws in conditioning with compound stimuli. Psychological Bulletin, 87, 351-378.

Kehoe, E. J., & Holt, P. E. (1984). Transfer across CS-US intervals and sensory modalities in classical conditioning of the rabbit. Animal Learning and Behavior, 12, 122-128.

Kehoe, E. J., Morrow, L. D., & Holt, P. E. (1984). General transfer across sensory modalities survives reductions in the original conditioned reflex in the rabbit. Animal Learning and Behavior, 12, 129-136.

Kehoe, E. J., & Schreurs, B. G. (in press). Compound-component differentiation as a function of CS-US interval and CS duration in the rabbit's conditioned nictitating membrane response. Animal Learning and Behavior.

Pavlov, I. P. (1927). Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex (G. V. Anrep, trans.). London: Oxford University Press.

Razran, G. (1965). Empirical codifications and specific theoretical implications of compound-stimulus conditioning: Perception. In W. F. Prokasy (Ed.), Classical conditioning (pp. 226-248). New York: Appleton-Century-Crofts.

Razran, G. (1971). Mind in evolution. New York: Appleton-Century-Crofts.

Rescorla, R. A. (1972). "Configural" conditioning in discrete-trial bar pressing. Journal of Comparative and Physiological Psychology, 79, 307-317.

Rescorla, R. A. (1973). Evidence for the "unique stimulus" account of configural conditioning. Journal of Comparative and Physiological Psychology, 85, 331-338.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), Classical conditioning II (pp. 64-99). New York: Appleton-Century-Crofts.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). Learning internal representations by error propogation (ICS Tech. Rep. No. 8506). San Diego: University of California. Institute for Cognitive Science.

Scavio, M. J., Jr., & Thompson, R. F. (1979). Extinction and reacquisition performance alternations of the conditioned nictitating membrane response. Bulletin of the Psychonomic Society, 13, 57-60.

Schmaltz, L. W., & Theios, J. (1972). Acquisition and extinction of the classically conditioned response in hippocamectomized rabbits (Oryctolagus cuniculus). Journal of Comparative and Physiological Psychology, 79, 328-333.

Smith, M., & Gormezano, I. (1965). Effects of alternating classical conditioning and extinction sessions on the conditioned nicititating membrane response of the rabbit. Psychonomic Science, 3, 91-92.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. Psychological Review, 88, 135-171.

Whitlow, J. W., & Wagner, A. R. (1972). Negative patterning in classical conditioning: Summation of response tendencies to isolable and configural components. Psychonomic Science, 27, 299-301.

Woodbury, C. B. (1943). The learning of stimulus patterns by dogs. Journal of Comparative Psychology, 35, 29-40.

APPENDIX 1
SIMULATION OF THE LAYERED NETWORK MODELS

The computer simulation of the Stage I, II, and III networks was implemented in the following fashion:

1. The output of each element (FIREj) was either 1 or 0. The outputs of sensory elements T, L, and US were specified in the program on a trial by trial basis, while the outputs of X, Y, and R were determined by learning and output rules.

2. Each point of connection between successive elements had a connection weight designated as Vij. The connection weights were designated as Vtx, Vty, Vlx, Vly, Vxr, and Vyr, where the first letter in the subscript refers to the origin of the output and the second letter refers to the element receiving the output. These connection weights started at zero and were unbounded in both the positive and negative directions. The connection weights from the US output to the X, Y, and R elements were fixed at 1.00.

3. Each trial was divided into two time steps, the CS period and the US period.

4. During the CS period, the following events occurred:

a. The output of the T and L elements was determined by the programmed trial sequence, and the appropriate connections with X and Y became eligible for change. For example, on a compound trial, there were outputs from T and L. Accordingly, the T-X, T-Y, L-X, and L-Y connections all became eligible for change.

b. A separate threshold (Tj) was independently determined for X, Y, and R by generating a random number between 0.00 and 0.99. Across trials, the distribution of thresholds was approximately normal. To alter the threshold, a constant was added or subtracted from the random number.

c. The output of X, Y, and R was:

FIREj = 1 if sum of eligible input weights > Tj

FIREj = 0 otherwise.

For example, on a compound trial, the output of X was determined by comparing Vtx + Vlx against Tx. At the same time, the output of Y was determined by comparing Vty + Vly against Ty.

d. After the outputs of X and Y were determined, then the output of R was determined by same process. For example, if only X fired on a particular compound trial, then the output of R (FIREr) was determined by comparing Vxr against Tr.

5. During the US period, the eligible connection weights throughout the network were modified according to the

Rescorla-Wagner model, where the change in a connection weight (dVij) followed the formula:

$$dVij = aj\ (Lj - \sum Vij),$$

where

aj was the rate of change parameter for the target element of the connection. ($0 < aj < 1$). On non-reinforced trials, aj was modified by parameter B0 ($0 < B0 < 1$) (See Point 7 below).

Lj was the total connection strength of eligible connections on a target element that could be supported by the US input on any given trial (Lj = 1.0 on reinforced trials. Lj = 0.0 on nonreinforced trials.)

$\sum$ Vij was the net sum of the associative strengths of all concurrent eligible inputs to the jth element.

6. In order to compute the CR likelihood for the tone CS, light CS, and compound CS for each block of training trials, a series of 30 "phantom CS periods" was conducted at the end of each block. Thus, the CS period for each type of trial was repeatedly conducted and the likelihood of a CR was determined. However, the change formula applicable during the US period of training trials was not used. Effectively, these phantom CS periods served as repeated test trials but without the extinctive effect that a prolonged series of test trials would have had in a behavioral experiment.

7. The simulations of successive acquisitions, extinctions, and learning-to-learn (Figures 2, 5, 7) used the following parameter values: ax = ar = 0.02, mean Tx = 0.75, mean Tr = 0.50, B0 = 0.15. The simulations for compound conditioning, positive patterning, and negative patterning (Figure 10) used the following parameter values: ax = 0.100, ay = 0.001, ar = 0.004, mean Tx = 0.65, mean Ty = 0.25, mean Tr = 0.50, B0 = 0.33.

SIMULATION OF THE CLASSICALLY CONDITIONED NICTITATING MEMBRANE RESPONSE BY A NEURON-LIKE ADAPTIVE ELEMENT: A REAL-TIME VARIANT OF THE SUTTON-BARTO MODEL

Diana E.J. Blazis, John E. Desmond, John W. Moore, and Nell E. Berthier

Department of Psychology
University of Massachusetts
Amherst, Massachusetts 01003

ABSTRACT

The Sutton-Barto (SB) model of learning is based on a neuron-like adaptive element. The model has computational features suitable for describing a variety of classical conditioning phenomena, including blocking, conditioned inhibition, and higher-order conditioning. However, it presently does not describe within-trial phenomena related to conditioned response (CR) topography. We here describe in detail an extension of the SB element, referred to as the Sutton-Barto-Desmond (SBD) model, which is capable of simulating topography of the conditioned nictitating membrane response (NMR) of the rabbit. The SBD model places certain constraints on the SB model's parameters and makes some additonal assumptions about the form of inputs to the element. The model describes (1) the gradually increasing amplitude of the CR within a trial with the peak amplitude at the temporal locus of the US, (2) the decrease in CR onset latency over training, and (3) appropriate interstimulus interval (ISI) functions, with optimal learning occurring with an ISI of .25 seconds. In addition, the model lends itself to descriptions of neuronal firing related to the CR. We believe the SBD model may have implications for neurobiological studies of learning and memory.

INTRODUCTION

Several extensions of the original Sutton and Barto (SB) model of connectionistic learning (Sutton & Barto,1981; Barto & Sutton, 1982) have been introduced in recent years. These connectionistic models have proven capable of supporting a wide variety of complicated learning control problems, providing plausible architectures for distributed processing in adaptive networks.

The success of the SB model and its extensions is apparent not only in the operation of adaptive networks, but also at the level of single elements. The original SB model was described in terms of single-neuron-like adaptive element operating under a learning rule similiar to that proposed by Rescorla and Wagner (1972), and was shown to be capable of simulating many features of classical conditioning, particularly of the rabbit nictitating membrane response (NMR). Among these features were the anticipatory nature of the conditioned response (CR), acquisition and extinction, blocking,higher-order conditioning, and interstimulus interval (ISI) effects upon trace conditioning of the rabbit NMR.

Fig.1 illustrates the original SB neuron-like adaptive element. Note that the US in the original model is signalled by a pathway of a fixed efficacy, denoted λ. Inputs for each conditioned stimulus (CS) vary in transmission efficacy according to the strength of a learned connection for each CS, the synaptic weight V_{CS}. Two memory processes contribute to changes in synaptic weight: $\bar{x}$, which

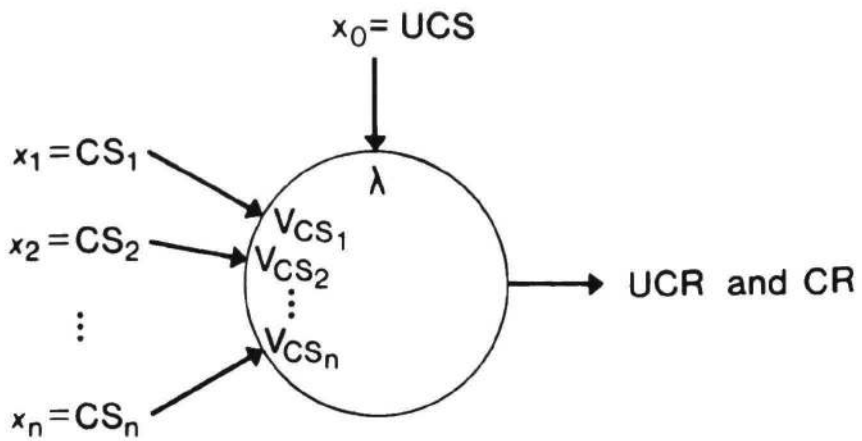


Figure 1: The original Sutton and Barto neuron-like adaptive element. Input pathway x_i has transmission efficacy V_{CS} corresponding to the associative strength of CS_i. The US (labeled UCS) is signalled via a pathway of fixed efficacy λ. Element output contributes to the UR (labeled UCR) prior to conditioning and to both the CR and the UR after conditioning. (from Barto and Sutton,1982).

determines the extent to which a given synapse is eligible for modification, and $\bar{x}$, a trace which represents the memory of the element's output in the preceding timestep. The single output of the element is computed as the weighted sum of all inputs. In the sense that the element receives several inputs, but transmits only one output, the SB element can be conceptualized as a "classic" neuron.

One of the research objectives of this laboratory is to examine the validity of single-neuron schemas of classical conditioning of the rabbit NMR. To that end we have taken a variey of approaches: electrophysiological, anatomical and behavioral among them. Our interest in the SB neuron-like adaptive element is the generation and development of a variant which simulates the topography of the NM CR as it appears in real rabbits in real time. The NMR is the sweeping of the nictitating membrane over the surface of the eyeball; it is a protective response that is a passive consequence of the retraction of the eyeball into the orbit by the retractor bulbi and extraocular muscles. This response has been studied within-trials and over trials in both single and multiple-CS protocols, thus offering an extensive database for the assessment of any model (Gormezano, Kehoe & Marshall, 1983). Recently, we demonstrated that the algorithm subserving the original SB element, though silent with respect to the generation of CR topographies, can be extended to the simulation of the neurophysiological and behavioral features of the rabbit NMR, predicting the ontogeny of the conditioned response (CR) over trials as well as within trials (Moore,Desmond,Berthier,Blazis,Sutton & Barto,1985; Moore, Desmond, Berthier, Blazis,Sutton & Barto,in press).

The CR topography characteristics that we sought to model included (1) gradually-increasing neuronal firing within a trial with attainment of peak CR amplitude at the temporal locus of the US; (2) a decrease in CR onset latency over training; and (3) an appropriate ISI function, with optimal conditioned strength accruing at an ISI of 250 milliseconds (ms). In order to produce these real-time phenomena, additional assumptions and constraints upon the variables of the original

SB model became necessary. We modified the SB element to yield appropriate topography in a single-CS forward-delay training paradigm in real-time. The question then became whether the model could retain the success of the original model's predictions in the domain of multiple-CS conditioning without further modification of the parameters.

This report will present the equations and assumptions of the constrained SB element, referred to as the Sutton-Barto-Desmond (SBD) model. We will show that the SBD element extends the SB model into real time while retaining the predictions of the SB element in multiple-CS domains.

THE MODEL: ASSUMPTIONS, EQUATIONS AND CONSTRAINTS

We begin our discussion of the SBD model by addressing the assumptions we have made about conditioning, stating the equations which make up the model, and justifying the constraints we placed upon the rules of the original SB model.

An assumption fundamental to the SBD model is that computations affecting synaptic weights occur before, during, and after the US. The original SB element assumed a 300 ms US, a duration sufficient to complete all computations within a trial before US offset. In rabbit NM conditioning, a more realistic US duration is 50 ms; under the conditions of the original model, a 50 ms US does not allow sufficient time for $\bar{x}$, the eligibility trace, to return to zero. Implementing post-US computations into the SB model required constraining the decay of various traces and including a mechanism whereby the effectiveness of the US changes over trials. These changes in the model are discussed more fully below.

Another assumption of the model concerned the form of the CS input to the adaptive element. Extending the SB element to model CR topography was accomplished in part by treating the CS input, designated as x in the model, as a continuous function. In the original model x was a step function equal to 1 at CS onset and 0 at CS offset. Although a step function defines the onset and offset of the external stimulus, it does not allow the SB model to predict rabbit NMR topography. For the purposes of the adaptive element, we assumed that the functions describing x mimic the behavioral CR. Thus, while the CS is on, x is shaped by a function which begins to rise gradually in an S-shaped fashion soon after CS onset, eventually maximizing at the temporal locus of the unconditioned stimulus (US). Assuming that t is a timestep representing 10 ms of real time:

$$x_i(t) = [arctan(mt + b) + 90]/(180 * h)) \tag{1}$$

for time $t = 1$ to CS offset. We selected the arctangent function because it is sigmoidal in shape and because it is a convenient one to implement in FORTRAN programming. The x-shaping parameters m, b, and h are specified as input variables with current default values equal to 0.35,-12.5 and 1.0, respectively. Simulation experiments involving systematic variations of the x-shaping parameters revealed that the model is very sensitive to their values; the default values represent those which yield the best topography for the 250 ms CS.

The second function mimics the decay of the CR in post-US timebins. It is implemented at CS offset, decays geometrically and is computed by multiplying successive values of x by a scalar:

$$\begin{aligned} x_i(t+1) &= k(x_i(t)) \\ k &= 0.85. \end{aligned} \tag{2}$$

Equations 1 and 2 specify, then, that the value of the input trace to the adaptive element begins to rise 70 ms after CS onset at the periphery,a latency justified by reports indicating that the minimal conditionable ISI for rabbits is 70 ms (Salafia, Lambert, Host, Chiaia & Ramirez, 1980). The input trace asymptotically reaches the value of one and decays to zero following CS offset.

The output of the SBD element $s(t)$, is defined as:

$$\begin{aligned} s(t) &= \sum_{i=1}^{n} V_i(t)x_i(t) + \lambda'(t) \\ 0.0 &\leq s \leq 1.0 \end{aligned} \tag{3}$$

where $\lambda'(t) = 0$ prior to the occurrence of the US. During US presentation, $\lambda'(t)$ is calculated as follows:

$$\lambda'(t) = \begin{cases} \lambda - V_i(t), & \text{if } 0 \leq V_i(t) \leq \lambda; \\ 0, & \text{if } V_i(t) > \lambda; \\ \lambda, & \text{if } V_i(t) \leq 0. \end{cases} \tag{4}$$

where V_{max} = the largest starting weight for all CSs present on a given trial, and λ = a constant that reflects US intensity.

Following the US, λ' is decremented :

$$\lambda'(t+1) = 0.9 * \lambda'(t) \tag{5}$$

The computation of $\lambda'(t)$ in equation (4) assumes that the contribution of the US to the element's output decreases as CS synaptic strength increases. In addition, λ' allows the element to predict diminution of the UR as conditioning progresses, a phenomenon previously observed in rabbits by Donegan (1981). Equation (5) reflects the assumption that the US input to the element is not equal to zero at US offset but instead, decays progressively, thereby influencing post-US decrements in V.

In the present implementation, CR topography is defined by the sliding arithmetic mean of the values of the current output s and that of the two preceding timesteps. This was done to smooth the transition from CR to UR. We further assume that the output of the element is bounded between 0.1 and 1.0. A negative s was deemed inappropriate for modelling NMR topography since a negative output implies NM retraction and exopthalmus, CR-opposing responses which are generally not observed in the rabbit NMR. The value of 0.1 reflects a threshold between the SBD element's output and the motoneurons which generate the peripherally observed response. The upper limit of 1.0 is derived from evidence indicating an upper limit for the amplitude of the behavioral NMR. Bounding s in this way also facilitates the computations relating s to neuronal firing rates of 10 to 100 Hertz needed for peristimulus– time histograms which cumulate neuronal firing over trials in the simulations.

Thus, equation (3) indicates that the output of the element at any given time t is equal to the weighted sum of its inputs. In a single–CS forward–delay paradigm, the output prior to the US is identical in form to the input, x, and the magnitude of the output is proportional to the synaptic weight.

The equation which dictates changes in synaptic weight (associative strength) is retained from the original SB model. Synaptic weights can be positive, negative or zero, which we interpret as

corresponding to excitatory, inhibitory or neutral inputs, respectively. $V_i(t)$,the synaptic weight of the ith conditioned stimulus, is modified as follows:

$$V_i(t+1) = V_i(t) + c[s(t) - \bar{s}(t)]\bar{x}_i(t) \tag{6}$$

where c is a learning rate parameter equal to 0.15, s is the element's output, and $\bar{s}$ is the element's prediction of its output based on its prior activity:

$$\begin{aligned} \bar{s}(t+1) &= \beta[\bar{s}(t) + (1-\beta)s(t)] \\ \beta &= 0.6 \end{aligned} \tag{7}$$

Note that the values of the learning rate parameter c, and of β in Equation 5 exert important influences upon the ability of the SBD model to accurately portray rabbit NM conditioning. For example, c can decrease or hasten the rate of learning and affect CR topography, although asymptotic synaptic weights are not affected.

Variable $\bar{x}$ is a CS duration-dependent stimulus trace which defines the extent to which a given synapse or connection is eligible for modification. In SBD, unlike SB, the eligibility trace mirrors and lags 30 ms behind the input trace.

$$\bar{x}(t+1) = x_i(t-2) \tag{8}$$

during the time that the CS is on. When the CS is off,

$$\bar{x}_i(t+1) = \delta * \bar{x}_i(t) \tag{9}$$

where $\delta = e^{-3/d}$, d = CS duration in timesteps; $d \geq 25$. These computations define a period of eligibility which begins some time after CS onset and persists beyond CS offset.

The goal of simulating appropriate ISI functions was accomplished in part through our specification of the decay of $\bar{x}$. Note that δ becomes larger for relatively longer CSs, thus slowing the decay of $\bar{x}$ and allowing greater opportunity for decrements in V following CS offset. Thus asymptotic synaptic weights are lower for CSs of relatively long duration.

As mentioned above, one goal in the development of the SBD model was to extend the SB model to encompass the simulation of neuronal firing. The current implementation of the SBD model lends itself readily to the computation of neuronal firing on single trials and accumulation of spikes over trials to form peristimulus-time histograms (PSTHs). Space limitations preclude a complete description of this portion of the model; see Moore et al (in press) for details. An example of neuronal firing is shown in Fig. 2, which shows a PSTH of neuronal firing to a 250 ms CS followed immediately by 30 ms US after 50 trials.

PREDICTIONS OF THE SBD MODEL

We have examined the behavior of the model at many levels: within- trials and over trials in single and multiple–CS conditioning. What follows is a brief account of our findings, beginning with topography for single–CS conditioning and interstimulus- interval functions, and progressing to multiple–CS procedures.

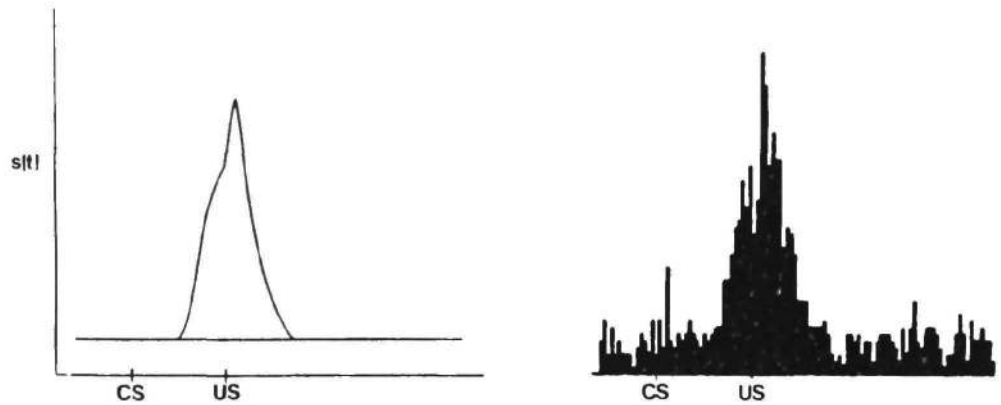


Figure 2: Simulated CR/UR complex and a PSTH for single–CS forward– delay training with a 250–ms CS terminating simultaneously with the onset of a 30–ms US after 50 trials. $\lambda = 0.9$, $c = 0.15$.

NMR Topography

The default values for the parameters shaping the input trace x (m= 0.35, b= -12.5 and h =1.0) allow the model to appropriately simulate many topographical features characteristic of the rabbit NMR. One, mentioned earlier, is a decrease in the latency of the onset of the CR as conditioning proceeds. The amplitude of the CR increases over training, with maximal amplitude occurring at or near the temporal locus of the US. In addition, the amplitude of the UR decreases over training, a phenomenon shown in real rabbits by Donegan (1981). During simulated extinction trials, the model produces the increase in the latency of CR onset and the diminution of the CR amplitude seen in the laboratory.

Though the current version of the SBD credibly simulates many aspects of NMR topography at optimal ISIs, there are other features of rabbit nicitating membrane conditioning which the model does not portray successfully. For example, in protocols involving very long CS durations, rabbit NM CRs begin to rise about halfway through the ISI (Smith, 1968). However, the current specification of the input trace allows the NMR to begin far earlier than that. What the model needs, then, is a mechanism specifying what Pavlov referred to as "inhibition of delay".

Preliminary simulation experiments involving systematic variation of parameters which shape x indicate inhibition of delay can be built into the model by allowing m, the parameter which determines the rise time of x, to increase over conditioning in an s-shaped function which is determined partly by the ISI and partly by the number of trials. m is allowed to maximize at .35, the default value of the current implementation, and a prerequisite for proper asymptotic behavior and topography. Since m is recomputed at every trial, x is also computed at every trial. Note that x in the current version of the SBD model is computed only once. The idea is that the input to the element is not a static entity, but rather one which varies in its efficiency according to the optimality of the ISI and the amount of exposure to the CS that the element has received.

Changing the quality of the input trace over the course of training provides a theoretical framework for generating appropriate topography during trace conditioning. Like its parent model, the

SBD model predicts trace conditioning. The model extends SB by predicting that trace conditioning yields less conditioning than forward-delay for the same ISI. These predictions are borne out in real NMR conditioning. However, the CRs of real rabbits occur in the trace interval preceding the US, while the SBD model predicts that CRs occur during the CS.

Producing CRs in the trace interval rather than during the CS might be achieved by altering assumptions regarding the onset and timecourse of x. First, we can assume that the offset of x is simply extended beyond the offset of the nominal CS, and that the decay rates of x and $\bar{x}$ are prolonged for trace conditioning. However, simply extending x is insufficient because CRs would still rise rather early during the nominal CS, and short CSs would probably not yield conditioning even when the ISI is optimal by empirical standards. Another approach is to base the computation of x not upon the nominal CS duration, but upon the ISI, a tactic not taken in the present FORTRAN implementation of the SBD model. Such an approach is supported by an early study in rabbit NMR trace conditioning, wherein a 50 ms CS was presented at ISIs ranging from 0 to 4 seconds (Smith, 1968); the present version of SBD would not even begin to show conditioning until the nominal CS duration exceeds 70 ms.

Testing of an experimental program basing the computation of a dynamic x upon the true ISI (CS onset to US onset) is now underway. Preliminary results indicate that it may indeed be possible to simulate CRs which occur during the trace interval. However, defining x based upon ISI naturally yields the same topographies and asymptotic weights for both forward-delay and trace protocols, contrary to the laboratory evidence. Furthermore, this implementation does not account for evidence indicating that the *offset* of the nominal CS and the onset of the US set up a temporal CS for the animal (Liu and Moore, 1969). This finding suggests that x ought to be generated during the trace interval. What kind of x would a trace interval produce? Can we assume that the trace interval x is of a different character from that generated by a "real" CS? Lastly, how and when during conditioning would the element come to regard the trace interval as the true CS? Future simulation experiments will address these questions.

As the preceding comments indicate, there are areas where the ability of the model to provide an accurate description of CR topography is incomplete. However, we are encouraged not only by how easily changes in x can account for these problem areas, but also by how these changes in x suggest to us what the actual "CS" can become for an animal.

Interstimulus Interval (ISI) Functions

As we have shown, the present implementation of the SBD model can, with a few exceptions, generate reasonable descriptions of within-trial events. Given this ability, how does the model fare in its description of events which occur over trials in a number of paradigms? The simplest starting place is the examination of the relative synaptic weights at asymptote for CSs of a variety of durations, the ISI function. The rabbit NMR literature describes the ISI function as an inverted U-shaped function with peak conditioning occuring to CSs of 250 ms duration (Gormezano, Kehoe and Marshall,1983). As Fig. 3 shows, the default values of the SBD model produce ISI functions for acquisition of forward-delay and trace conditioning which are generally consistent with the literature, with the exception of the prediction of negative weights for CSs of 100 ms duration. Extinction (not shown) also proceeds in appropriate fashion, with more optimal CSs extinguishing more slowly.

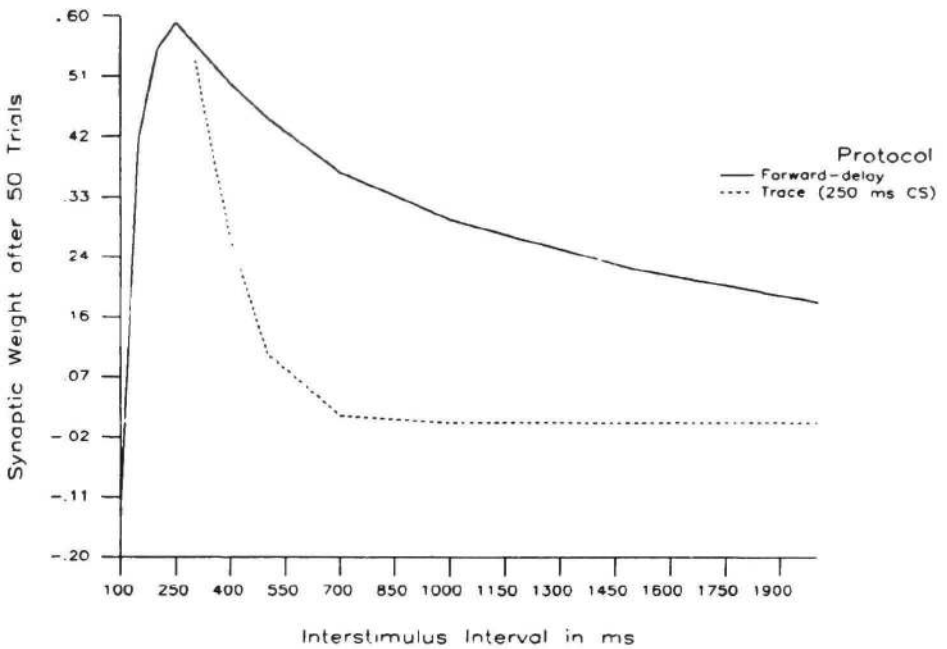


Figure 3: Synaptic weights after 50 trials for CS–US intervals ranging from 100 to 2000 ms for forward-delay conditioning and trace–conditioning protocols.

Conditioned Inhibition

The SBD model satisfactorily simulates Pavlovian conditioned inhibition (CI), a fairly difficult multiple-CS discrimination paradigm. A simulation of the CI paradigm is presented in Fig. 4. In CI, two trial types are presented in a pseudorandom sequence: the first trial type consists of CS_1 presented with a US; and the second consists of a compound of $CS_1$and CS_2 presented without the US. Fig. 4 shows that the synaptic weight for CS_1 becomes positive, while that for CS_2 becomes negative. Negative weights are interpreted as an indication of conditioned inhibition. The SBD model not only predicts CI, it also generates salient features of the paradigm including the extended number of trials required for asymptotic performance and the initial slightly excitatory character of the unreinforced compound. An unusual and untested prediction of the SB model is that the conditioned inhibitor will become excitatory if it precedes the conditioned excitor in the non-reinforced compound presentations.

As we mentioned earlier, the original Sutton-Barto learning rule is similiar to the Rescorla-Wagner model of associative learning. One of the weaknesses of both the SB and RW models is the prediction that a conditioned inhibitor will extinguish if presented alone, a prediction that has not been verified empirically. The SBD model predicts that a conditioned inhibitor presented alone will retain its inhibitory character and in this respect the model emulates the empirical evidence. Inhibition does not extinguish in the SBD model because the output of the element, s, is never less than zero. Consequently, the term $(s - \bar{s})$ in Equation 1 is also equal to zero and no changes in V for that inhibitory CS can occur in the absence of a US.

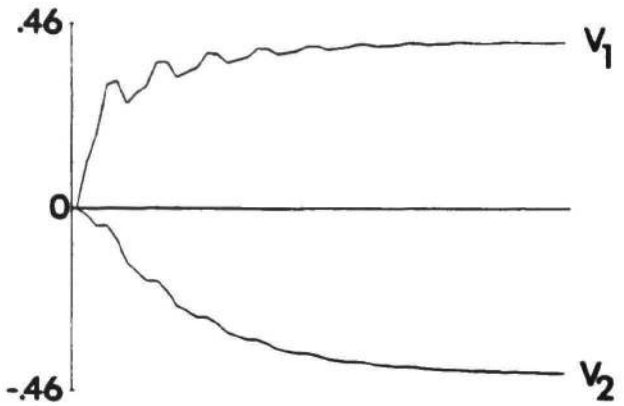


Figure 4: Synaptic weights (V_1 and V_2) for CS_1 and CS_2 for a simulated conditioned inhibition paradigm as a function of trials.See text for discussion.

Kamin Blocking

The model successfully predicts blocking of conditioning by a fully pre-trained CS to a novel added CS. Blocking is predicted by both the RW model and the original SB model. In Stage 1 of a blocking paradigm, CSB, the CS which will become the blocker, is presented with a US until an asymptotic level of responding is produced. In Stage 2, CSB is paired with an added, novel CS, denoted here as CSA. The compound of CSs A and B is reinforced. The added CS will fail to condition normally, a phenomenon demonstrated in the NMR preparation by Marchant and Moore (1973). The SBD model predicts that maximal blocking to CSA occurs when CSB is fully trained and when both CSs are presented simultaneously in Stage 2. Simulations of blocking with incompletely-trained blockers or with serial presentations of CSs A and B yield a variety of effects too numerous to describe here. However, one particularly interesting and novel prediction of both SB and SBD is that the blocker will become inhibitory in Stage 2 of a blocking paradigm if the added CS precedes and overlaps the blocking stimulus.

Higher-order conditioning

The SBD model recapitulates the success of the original model in its treatment of higher-order conditioning. In Stage 1 of a higher-order conditioning task, CS_1 is paired with a US and trained to an asymptotic level of responding. In the second stage, another CS, CS_2,is presented, followed by the fully-trained CS_1. Even though the usual US is not presented, the pairing of CSs 1 and 2 results in increased responding to CS_2. If refresher presentations of CS_1+US are not included, responding to CS_2 and eventually CS_1 falls off.

Other multiple-CS phenomena which are encompassed by the model include several untested predictions regarding within-trial timing of serial compounds. Discussion of these topics will be the focus of a future report.

CLOSING COMMENTS

Sutton and Barto (1981) recognized that their model places a heavy computational burden on a single neuron. However, they identified several possible cell-physiological mechanisms for components of the model, including the eligibility trace and the prediction of reinforcement. The SBD model places an additional burden on the neuron, in particular the computation of $\lambda'(t)$, which reduces the effectiveness of the US as associative strength increases. The justification for $\lambda'(t)$ arose from the assumptions that computations affecting synaptic weight occur not only during the CS, but after the US as well. Without the $\lambda'(t)$ rule, synaptic weights after extensive training are unreasonably low. Furthermore, the $\lambda'(t)$ rule enhances the model's performance regarding ISI functions and response topography.

There are features of classical conditioning of the rabbit NMR which the current implementation does not address. Among these are intertrial interval phenomena, stimulus salience effects, and attentional effects. Some of these phenomena can be easily encompassed within the framework of the SB and SBD models. Others, for instance, attentional phenomena, could be included, but perhaps at some cost in terms of our intuitions about the sheer number of computations that a single cell could perform.

Detailed descriptions of the behavior of the components of the SBD model and of the model's performance in a variety of simulations involving within-trial timing of stimulus events are now underway. Indeed, the strength of the model lies in its ability to generate predictions regarding within-trial events in rabbit NMR conditioning. To the neurophysiologist, such predictions provide hypotheses about the timing functions of the nervous system components involved in NM conditioning. For those working in adaptive architectures, the model's successes and failures in multiple–CS domains like compound conditioning can suggest the types of computation a single element can sustain when its inputs are assumed to model desired output. For animal learning theorists, empirical verification of the predictions of the SBD model can extend the sparse literature pertaining to within–trial events in rabbit NMR conditioning. For now, we are encouraged by the preliminary success of the model and suggest that its structures and constraints may have implications for the understanding of the physiology of learning and memory.

Acknowledgements

This research was supported by grants AFOSR 830125 and NSF BNS 8317920.

REFERENCES

Barto, A.G. & Sutton, R.S. (1982). Simulation of anticipatory responses in classical conditioning by a neuron-like adaptive element. *Behavioral Brain Research*, *4*, 221–235.

Donegan, N.H. (1981) Priming-produced facilitation or diminution of responding to a Pavlovian unconditioned stimulus, *Journal of Experimental Psychology: Animal Behavior and Processes*, *7*, 295–312.

Gormezano, I., Kehoe, E.J. & Marshall,B.S. (1983). Twenty years of classical conditioning with the rabbit. In J.M. Sprague and A.N. Epstein (Eds.)*Progress in Psychobiology and Physiological Psychology, 10*,197–275.

Liu, S.S. & Moore, J.W. (1969). Auditory differential conditioning of the rabbit nictitating membrane response: IV. Training based on stimulus offset and the effect of an intertrial tone. *Psychonomic Science, 15*, 128–129.

Marchant, H.R. III & Moore, J.W. (1973). Blocking of the rabbit's nictitating membrane response in Kamin's two-stage paradigm. *Journal of Experimental Psychology, 101*, 155-158.

Moore,J.W., Desmond,J.E., Berthier, N.E., Blazis,D.E.J., Sutton,R.S. & Barto, A.G. (1985). Connectionistic learning in real time: Sutton-Barto adaptive element and classical conditioning of the rabbit nictitating membrane response. *Proceedings of the Seventh Annual Conference of the Cognitive Science Society*, 15-17 August 1985, Irvine, California.

Moore, J.W., Desmond, J.E., Berthier, N.E., Blazis, D.E.J., Sutton, R.S. & Barto, A.G. (in press). Simulation of the classically- conditioned nictitating membrane response by a neuron-like adaptive element: Response topography, neuronal firing, and interstimulus intervals. *Behavioral Brain Research.*

Rescorla, R.A.& Wagner, A.R. (1972) A theory of Pavlovian conditioning: variations in effectiveness of reinforcement and nonreinforcement. In A.H. Black and W.F. Prokasy (Eds.), *Classical Conditioning II: Current Research and Theory*, Appleton-Century-Crofts, New York.

Salafia, W.R., Lambert, R.W., Host, K.C., Chiaia,N.L., & Ramierz, J. (1980). Rabbit nictitating membrane conditioning: Lower limit of effective interstimulus interval.*Animal Learning and Behavior,1*, 85–91.

Smith, M.C. CS-US interval and us intensity in classical conditioning of the rabbit's nictitating membrane response. *Journal of Comparative and Physiological Psychology,66*, 679-687.

Sutton, R.S. & Barto, A.G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review, 88*, 135–170.

Introspection and Reasoning about the Beliefs of other Agents[1]

Anthony S. Maida

Department of Computer Science
The Pennsylvania State University

ABSTRACT

A cognitive agent uses representations to reason about the world. An "introspective" cognitive agent has the ability to manipulate representations (meta representations) of its own representations. If such an agent were to embody its beliefs in its representations, then the agent could reason about its own beliefs by manipulating its meta representations.

A "belief reasoner" can reason about the beliefs of other agents. There has been considerable research in the construction of belief reasoners. This paper observes that the construction of such systems can, in large part, be reduced to the task of constructing introspective systems.

We illustrate how an introspective agent can use analogy-based reasoning to construct an architecture for belief reasoning on the basis of examining its own architecture.

1. Introduction.

This paper views the task of reasoning about the beliefs of other agents in terms of more basic processes of introspection in which an agent reasons about its own beliefs. We begin with the assumption that agents think in a proposition-like "mentalese." The intuition underlying our line of argument is as follows: If agents think in a proposition-like mentalese, then to describe their belief states, we can use a metalanguage capable of describing arrangements of propositions. If the language is capable of describing any possible arrangement of propositions, then the language is capable of describing any possible belief state of the agent. If an agent already uses such a language to introspectively describe and reason about its own beliefs, it should be possible for this agent to adapt this language, by some analogy-based process, to describe the beliefs of other agents.

2. Cognitive Science: The Representational Theory of Mind

The notion that an agent's beliefs determine its behavior is compatible with the metaphor of a "knowledge-based system" as used in artificial intelligence or the "representational theory and computational theory of mind" as used by philosophers of cognitive science (e.g., Fodor, 1981). In this view, a cognitive agent represents the world by the use of some internal language, a "mentalese," such as a propositional language. This view has led to the "knowledge representation

[1] This research was supported by ONR contract N00014-85-K-0521. Thanks to Ross Canfield, Minqui Deng, Minkoo Kim, Drew McDermott, Joe Niederberger, and Bonnie Webber for help with various incarnations of this manuscript.

hypothesis," upon which much of artificial intelligence and cognitive science is based (cf, Fodor, 1980; Pylyshyn, 1984; Smith, 1982). Thus, the primary assumption of this paper is compatible with much of mainstream cognitive science.

2.1. The Semantics of Belief Sentences.

To say that an agent has a belief is to say that the agent has constructed a representation in its mental language and that the agent takes this representation as accurately describing something. For instance, say that an agent named "Pat" believes that a dog named "Fido" is ferocious. This means that in the mental language which Pat uses to represent the world, there is an expression that resides in his data base, which represents the proposition *Fido is ferocious*. It also means that Pat bears some relation to this expression indicating that it is believed by Pat.

If we assume that Pat thinks with propositions, then the sentence "Pat believes Fido is ferocious" can be taken to be true exactly if the proposition (1):

(1) (is-ferocious Fido)

resides in Pat's data base and the proposition is somehow marked as true (possibly implicitly, by virtue of it simply residing in the knowledge base). We have sketched for the above belief sentence a semantics in terms of the propositional content of a knowledge base and we will call this a *knowledge based semantics*.[2]

2.2. Belief Spaces as Nested Mental Models.

A perspicuous notation for depicting the beliefs of cognitive agents is by the use of belief spaces (cf., Fauconnier, 1985). A belief space can be construed as a mental model of another agent's representation of the world.[3] This naturally leads to an architecture which is a tree of nested mental models or belief spaces (cf., Fauconnier, 1985; Maida, 1984). Figure 1 depicts such a tree.

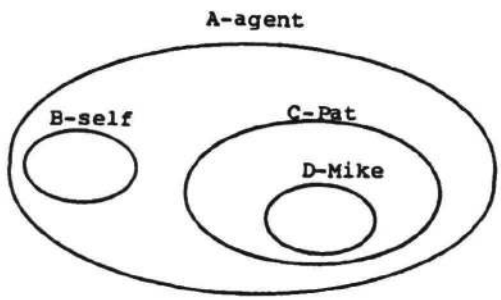


Figure 1. A tree of belief spaces. Nested ellipses indicate subtrees.

Each ellipse in this figure is a belief space; the nestings indicate nestings of the belief spaces. The ellipse labeled "A" indicates the agent's knowledge (beliefs). The ellipse labeled "B" indicates the agent's knowledge about its own knowledge; that is, objects in this space represent objects in the parent space. Ellipse "C" indicates the agent's knowledge about Pat's knowledge; that is, objects in this space represent objects in Pat's data base (not shown). Finally, the system has knowledge that Pat has knowledge of Mike's knowledge (indicated with ellipse D); that is, objects in this space represent objects in Pat's space that represent objects in Mike's space.

---

[2] This contrasts with the possible-world semantics of Hintikka (1962). See Halpern & Moses (1985) for a guide to modal logics of knowledge and belief. The possible-worlds approach attempts to define knowledge without reference to the internal structure of the agent who has the knowledge. The approach described in section 2.1 is sometimes called the "syntactic approach."

[3] Belief spaces or analogs have been used by Moore (1973), Cohen (1978), Martins & Shapiro, 1983; Rapaport & Shapiro, 1984; and Kobsa (1985). In linguistics, Fauconnier (1985) has made heavy use of belief spaces.

Simulative Reasoning. *Simulative reasoning*[4] is the process of one agent reasoning about the beliefs of a second agent as if the second agent were reasoning with his own beliefs. The belief-space architecture is highly suggestive of simulative reasoning. In principle, one could have inference processes in each belief space. In this paper, we will assume there is one actual inference engine in the root space. To conduct simulative reasoning in a child space, the inference engine will simulate a virtual inference engine in the child space.

2.3. Introspection as Representations of Representations.

Suppose we have a cognitive agent who not only maintains representations which describe things external to itself, but also maintains representations of its representations. The agent could then have beliefs about its representations and we could say that the agent is introspective.

3. Reducing Belief Reasoning to Introspection.

To some, the belief-space architecture may seem unparsimonious. Would it be plausible to assume that a cognitive agent would just happen to have an architecture consisting of a tree of data bases, simply to reason about others' beliefs? Fortunately, we can reduce the belief-space architecture to more basic principles. The belief-space architecture seems to be a consequence of any knowledge-based system that has a sufficiently rich self-model and capacity to reason by analogy.

The basic idea is to have the system replicate a theory of its own inference ability in a model of another agent, so that the copy is suitably modified to appropriately describe that other agent. We will call such an operation *projective analogy.*[5]

3.1. Replication of One's Self-Model and Inference Machinery.

Suppose an agent has a partial description of its own structure, including a descriptive sketch of the operations of its own analogy-based reasoning ability. This is illustrated in Figure 2a below. The large circle indicates the set of propositions that the agent believes. The inner circle, labeled "self-model," consists of the set of propositions that describe the agent's beliefs. Note that this "self-model" is the agent's self image. For reasoning about its beliefs, the agent has access only to the information delimited by the inner circle (i.e., its self-model), and not the outer circle. If the agent were to make an analogy between its own structure and the structure of some other entity, the analogy would have to be based on its self-model.

If the system learned that another agent, say Pat, has structure similar to it, then the system could create a description of Pat, shown in Figure 2b, by making a copy of its own self-model and modifying it appropriately to apply to Pat. (For instance, this might be done by replacing all occurrences of the symbol "self" in the description with the symbol "Pat.")

There is one more step needed in order to construct a tree of belief spaces. Suppose the agent's self-model contains a description of the analogy process. If so, then a copy of this description will have been duplicated in the description for Pat. If the agent then reasons about what would happen to Pat's data base if he (reciprocally) thought about the system, it will conclude that Pat will construct, within his own data base, a model of the system. We will call this *reciprocal* projective analogy. That is, *the original agent can apprehend that Pat can use his own reasoning to reason about it.* Hence, we have sketched how the construction of a tree of nested belief spaces might be automated. The next section will describe an example.

---

[4] Creary (1979) seems to have been the first to use the term. Dinsmore (1985) discusses simulative reasoning from a linguistic perspective.

[5] We call this "projective analogy" because, first, it is a kind of analogy, and second, the process of attributing one's own characteristics to another is called "projection" by some psychologists.

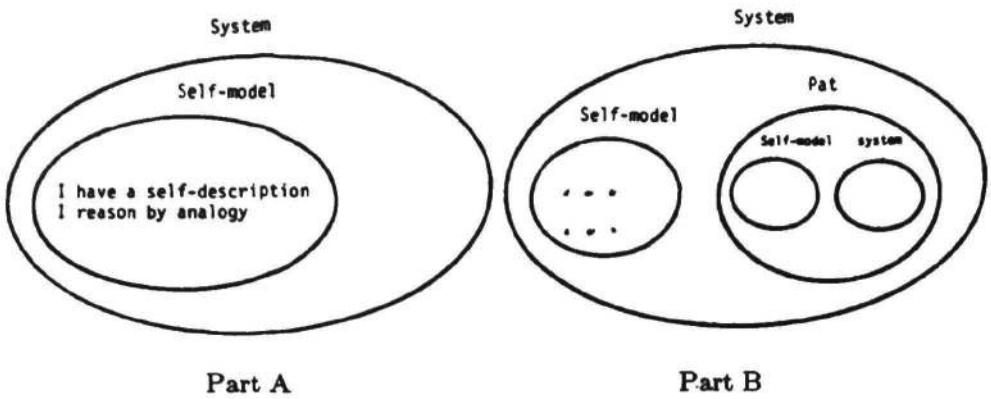


Figure 2. If a system with a self-model (A) learns of an entity, say Pat, who has similar structure to itself, then the system can create a model of Pat by projective analogy (B).

4. A Detailed Example.

In this section, we give a detailed example of the projective analogy process. It is the simplest example we can think of. We are going to get an agent to realize that another agent (Pat) can reason by *modus ponens*. However, we will fail to get the agent to realize that Pat realizes that it (the original agent) reasons by *modus ponens*. The reason for the failure will be that the original agent does not have an explicit model of the analogy process; the agent does not realize that it reasons by analogy and thus cannot attribute this characteristic to another Pat.

For our example we will break down the process of evolving from an introspective agent into a belief reasoner into two steps. They are:

> We must give the agent a theory of its own inference ability. We will call this an *auto-rational theory.*[6]
>
> We also need a capacity to replicate this theory and modify the copy to describe the inference ability of another agent; that is, a capacity for *projective analogy.*

The auto-rational theory may only partially describe the agent's inference ability. In our case, the auto-rational theory will not describe the capacity for projective analogy. With a partial theory, the replicated-and-modified theory constitutes a partial theory of the other agent.

4.1. A Partial Auto-Rational Theory.

We will postulate an agent that has three rules of inference wired into its mental structure. These are: 1) a procedural version of *modus ponens;* 2) an ability to do *propositional introspection;* and, 3) an ability for *projective analogy.* These three abilities characterize our agent's capacity to do inference. If the agent can represent explicitly that it has these abilities, then it has a theory of its own rationality--an auto-rational theory.

Modus ponens and *propositional introspection* are defined below. The letters p and q range over propositional formulas and the symbol *bt* stands for "believes that."

> **Modus Ponens.** If the propositions p and $p->q$ reside in the agent's data base, then the proposition q will reside in the agent's data base.
>
> **Propositional Introspection.** If the proposition p resides in the agent's data base, then the proposition *(bt self "p")* resides in the agent's data base.[7]

---

[6] Moore (1985) used the term *auto-epistemic* to refer to a theory of one's own knowledge. We are concerned with a particular aspect of one's own knowledge, namely one's knowledge of his rationality. An auto-rational theory is a kind of auto-epistemic theory.

[7] We will assume that a proposition is marked as true if it resides in the data base.

Suppose further that the axiom schemas (a)-(d) below reside in the agent's data base.

(a) (bt self "(p & p->q) -> q)")
(b) (bt self "(bt self "p") -> (bt self "(bt self "p")")")
(c) ((bt self "p") & (bt self "p->q")) -> (bt self "q")
(d) (bt self "p") -> (bt self "(bt self "p")")

Axioms (a)-(d) present a partial auto-rational theory of the agent's reasoning. Expressions (a) and (b) are *descriptive* but they are not *causal*. Axiom (a) is true exactly if the agent reasons by *modus ponens* as the axiom describes. However, the agent's believing the axiom does not cause the agent to reason by *modus ponens*. Axiom (b) is true exactly if the agent can do *propositional introspection*.

Expressions (c) and (d) are causal because they can be interpreted directly by the inference engine. They enable the system to do simulative reasoning about its beliefs in its self-model. We will call expression (a) the assumption of *awareness of rationality;* (b) the assumption of *awareness of propositional introspection;* (c) the assumption of *auto-syllogistic interpretation;* and (d) the assumption of *introspective interpretation*.

Auto-syllogistic interpretation allows the agent to simulate reasoning by *modus ponens* in its self-model. *Introspective interpretation* allows an agent to simulate *propositional introspection* in its self-model.

4.2. Genesis of Attributions of Rationality in Others.

Figure 3a, using the belief-space notation, depicts an agent with the above auto-rational theory. Expressions (a) and (b) are in the agent's self-model. Expressions (c) and (d) are in the root space. Figure 3b depicts the agent after the theory has been replicated and modified by projective analogy to describe another agent, Pat. The replicated expressions are indicated by apostrophes.



Figure 3. Replication of a partial auto-rational theory.

We will now describe the projective analogy process. It involves two components: 1) universal generalization;[8] and 2) universal quantifier elimination. *Universal generalization* transforms a copy of expression (a) into (q) by replacing a constant (which must be of type "agent") with a universally quantified (actually quantified over agents) variable as shown below:

(a) (bt self "((p & p->q) -> q)")

⇓ Universal Generalization on Agents

(q) (forall (x) (bt x "((p & p->q) -> q))")

[8] Universal generalization is a form of inductive reasoning. In this paper it will be the means by which we make a generalization about agents.

Universal quantifier elimination transforms an expression such as (q) into (a') below by instantiating the variable with a constant.

(a') (bt Pat "((p & p->q) -> q)")

Upon applying the composition of universal quantifier elimination and universal generalization to each of the expressions (a), (b), (c), and (d) we get expressions (a'), (b'), (c'), and (d') below.

(a') (bt Pat "((p & p->q) -> q)")
(b') (bt Pat "(bt Pat "p") -> (bt Pat "(bt Pat "p")")")
(c') ((bt Pat "p") & (bt Pat "p->q")) -> (bt Pat "q")
(d') (bt Pat "p") -> (bt Pat "(bt Pat "p")")

The expressions (a')-(d') describe the replicated and modified structure of Figure 3b. We started with an agent who had the mental representation of Figure 3a, and by projective analogy, the agent arrived at the structure of Figure 3b. Expressions (a') and (b') comprise the agent's model of Pat. Expressions (c') and (d') allow the agent to do simulative reasoning in Pat's belief space.

The agent can attribute to Pat only what it realizes about itself. Notice that the agent does not attribute an ability to do projective analogy to Pat. This is because the agent does not explicitly that it itself has this ability. Notice also that the agent has a model of itself, but does not view Pat has having a model of himself.

4.3. Awareness that One has a Self-Model.

Without a self-model, the agent would not realize that it thinks. However, to realize that it has a self-model, the agent must have a model of its self-model. It is for this reason that the projective analogy process did not attribute a self-model to Pat. Based on analogy with itself, the system did not realize that it had a self-model. However, the system has the ability to infer a model of its initial model. It can do simulative reasoning of itself in its self-model.

Self-Replication of One's Self-Model. As the agent thinks with its self-model, this can cause it to realize that it has a self-model. The agent can actually construct a copy of its self-model in its self-model. This is done as follows.

Since the agent can reason by *modus ponens*, it can manipulate expressions (a) and (d) to get expression (aa) below. Similarly, the agent can manipulate expressions (b) and (d) to get (bb) below. Since the agent can reason by *propositional introspection*, it can apply this to expression (c) to get expression (cc) below. The agent can also apply *propositional introspection* to get expression (dd) below.

(aa) (bt self "(bt self "(p & p->q) -> q)")")
(bb) (bt self "(bt self "(bt self "p") -> (bt self "(bt self "p")")")")
(cc) (bt self "((bt self "p") & (bt self "p->q")) -> (bt self "q")")
(dd) (bt self "(bt self "p") -> (bt self "(bt self "p")")")

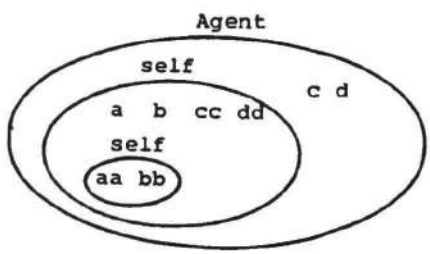


Figure 4. Using simulative reasoning to create a model of one's self-model.

4.4. Simulative Reasoning.

With what we have so far, the agent can do simulative reasoning in Pat's subspace by virtue of expression (d'). Since the agent can reason by *modus ponens*, it can manipulate expressions (a') and (d') to get expression (aa') below. Similarly, the agent can manipulate expressions (b') and (d') to get (bb') below.

(aa') (bt Pat "(bt Pat "(p & p->q) -> q")")

(bb') (bt Pat "(bt Pat "(bt Pat "p") -> (bt Pat "(bt Pat "p")")")")

Expressions (aa') and (bb') can be construed as Pat's self-model within the system's model of Pat, as shown in Figure 5.

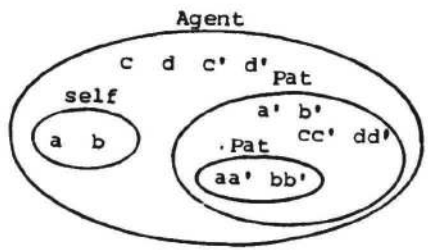


Figure 5. Using simulative reasoning to create a model of another's self-model.

4.5. Reciprocal Projective Analogy.

In section 4.4 we saw a replicated-and-modified model of Pat (i.e., expressions (a') and (b')) partially undergo another replicate-modify cycle producing expressions (aa') and (bb'). However, this new model is not very functional. For instance, there are no formulas analogous to (dd') which would enable simulative reasoning in that space. Additionally, this second-generation model cannot even partially replicate.

Realizing that Others Believe that You are Rational. For the agent to believe that Pat could realize that the agent itself is rational, we would need to create a model of the agent within Pat's subspace. We have called this *reciprocal projective analogy*. This would require simulative reasoning in Pat's subspace, using the rule of projective analogy. This cannot be done because the agent does not model Pat as having the ability to do projective analogy.

5. Summary and Conclusions.

So far we have achieved the following. We have shown how it might be possible for one agent to infer that a second agent has beliefs and that this second agent realizes it itself has beliefs. We have argued theoretically that it should also be possible to infer that this second agent can infer that other agents have beliefs, and that those other agents can make similar inferences. In essence this amounts to an architecture of belief spaces. The importance of this is that the intuitive notation of belief spaces can be interpreted as a kind of cognitive architecture with possible psychological reality. The architecture would have a natural explanation.

Limitations and Further Work. The language we have been using to express autorational theories has been, for the most part, *propositional* as opposed to predicate based. That is, it views propositions as atomic and cannot describe their subparts. This means that the language is inherently unable to describe various interesting phenomena. In particular, *projective analogy* cannot be described in the language because that would involve the reference to subparts of propositions. Consequently, we cannot have a fully introspective agent who reasons by analogy if his declarative mental language is only propositional. It would never be able to represent to itself the process of *projective analogy*. The reason we could not get our agent to realize that Pat realized it was rational was traced back to the fact that our original agent did not have an explicit model of the projective analogy process. It appears that this cannot be remedied in a propositional framework.

The robustness of the projective analogy process we described should also be scrutinized. Our characterization is really a kludge to demonstrate the feasability of the idea. It is unlikely to be robust. The domain of projecting properties from oneself to others should simply be another domain to study analogical reasoning.

In summary, topics for future research include the following: 1) we need more detailed introspective models, particularly beyond the propositional level; and, 2) we need to look at more cases of an agent reasoning by analogy from his own structure to make inferences about another agent.

References

[1] Cohen, P. R. On knowing what to way: Planning speech acts. Ph.D. Thesis, Technical Report No. 118, Department of Computer Science, University of Toronto, January, 1978.

[2] Creary, L. G. Propositional attitudes: Fregean representation and simulative reasoning. *Proc. IJCAI, 1979, 6, 176-181.*

[3] Dinsmore, J. Mental Spaces from a Functional Perspective. Manuscript, Department of Computer Science, Southern Illinois University at Carbondale, 1985.

[4] Doyle, J. A Model for Deliberation, Action, and Introspection. Doctoral dissertation submitted to the Massachusetts Institute of Technology; also M.I.T. Artificial Intelligence Lab. Memo AIM-TR-581, 1980.

[5] Fauconnier, G. *Mental spaces: Aspects of meaning construction in natural language.* Cambridge: MIT Press, 1985.

[6] Fodor, G. Methodological solipsism considered as a research strategy in cognitive psychology. In John Haugeland (Ed.), *Mind Design* Cambridge: MIT Press, 1981.

[7] Halpern J. Y. & Moses Y. A Guide To The Modal Logics Of Knowledge And Belief: Preliminary Draft. *Proceedings of the Ninth International Joint Conference on Artificial Intelligence,* Los Angeles, California, August 18-23, 1985, Vol. 1, pp. 480-490.

[8] Hintikka, J. *Knowledge and Belief.* Cornell University Press, Ithaca, New York, 1962.

[9] Kobsa, A. Using situation descriptions and Russellian attitudes for representing beliefs and wants. *Proceedings of the Ninth International Joint Conference on Artificial Intelligence,* Los Angeles, California, August 18-23, 1985, Vol. 1, pp. 513-515.

[10] Maida, A. S. Selecting a humanly understandable representation for reasoning about knowledge. *International Journal of Man-Machine Studies,* 22, 1985, 151-161.

[11] Maida, A. S. & Kim, M. Belief, Equality, and Quantification in the Belief Space Engine. Unpublished manuscript, 1985a.

[12] Maida, A. S. & Kim, M. The Belief Space Engine user's manual (draft). Department of Computer Science, Penn State University, University Park, PA, 16802, August, 1985b.

[13] Martins, J. P. & Shapiro, S. C. Reasoning in multiple belief spaces. *Proceedings of the Eighth International Joint Conference on Artificial Intelligence,* Karlsruhe, West Germany, August 8-12, 1983, pp. 371-373.

[14] Moore R. C. D-SCRIPT: A computational theory of definite descriptions. *Advance papers from the Third International Joint Conference on Artificial Intelligence,* Stanford, California, August, 1973, pp. 223-229.

[15] Moore, R. C. Reasoning about knowledge and action. *Proceedings of the Fifth International Joint Conference on Artificial Intelligence,* Cambridge, Massachusetts, August, 1977, pp. 223-227.

[16] Moore, R. C. Semantical considerations in nonmonotonic logic. *Artificial Intelligence,* 25(1), 1985, 75-94.

[17] Pylyshyn, Z. W. *Computation and cognition: toward a foundation for cognitive science.* Cambridge: MIT Press, 1984.

[18] Rapaport, W. J. & Shapiro, S. C. Quasi-indexical reference in propositional semantic network. *Proceedings of the Tenth International Conference on Computational Linguistics,* Stanford, California, July 2-6, 1984, pp. 65-70.

[19] Smith, B. C. Reflection and Semantics in a Procedural Language. Ph. D. Thesis and Tech. Report MIT/LCS/TR-272, MIT, Cambridge, MA, 1982.

Causal reasoning about quantities

Kenneth D. Forbus
Qualitative Reasoning Group
University of Illinois

Dedre Gentner
Psychology Department
University of Illinois

Abstract

Causality plays an important role in human thinking. Yet we are far from having a complete account of causal reasoning. This paper presents an analysis of causal reasoning about changes in quantities. We abstract from AI theories of qualitative physics three dimensions along which causal reasoning about quantities may be decomposed. We then use this framework to make some psychological predictions.

1. Introduction

People have a deep intuition that causality is a central and cohesive aspect of human mental life. Consequently, the problem of causality has long occupied philosophers and scientists. But the search for a unified theory that can explain human causal reasoning, much as theories of grammar explain syntactic processing, has so far been unsuccessful. These failures have led Hayes (1985) and others to conclude there is no deep theory of causality. Instead, causal reasoning may be simply a family of inferences whose properties will vary according to the content of the argument. This paper analyzes one kind of causal argument, causal reasoning about changes in quantities, to provide an account of the relevant issues and draw some implications for psychology.

In this paper we abstract from the AI qualitative physics literature three factors involved in causal reasoning: (1) whether there is an *explicit mechanism* or not; (2) whether the connective relations between quantities have a *direction of causation* built in, and (3) which type of *measurement scenario* is involved. We begin by laying out these factors as they apply in current work in qualitative physics. We analyze the relationships among these factors and discuss the current literature in light of these distinctions. Then we discuss their implications for psychology. We advance some conjectures and consider suggestive evidence from protocols.

2. Models of causality in changing quantities

One of the central concerns of qualitative physics (e.g., Bobrow, 1985) is describing how continuous physical properties change over time. Informally, it seems that people treat many of these deductions as causal:

"Pouring more milk into the glass will cause the level to go up."

"Turning on the stove increases the temperature of the burner, which causes heat to flow, which eventually causes the water on the stove to boil."

Workers in qualitative physics have made a number of proposals about how to model these conclusions. In order to compare these proposals, we isolate three important factors. Any theory of causation involving quantities must make some choice within each of these dimensions. Thus we have a basis for comparing theories and organizing psychological predictions. These three factors are:

1. *Explicit/Implicit Mechanisms:* Does the theory include an explicit notion of mechanism, in addition to objects, that is the root cause of all changes?
2. *Directed/Nondirected Connectives:* Are the relationships between quantities expressed by functional dependencies whose directionality is taken to express the direction of causation, or by nondirected constraint equations?
3. *Measurement Scenarios:* What sense of change is being discussed? Is it one change in a sequence, the difference between initial and final states, the difference between alternate possible worlds, or something which is occurring continuously?

We examine each aspect in turn, noting the issues involved and how the current systems of qualitative physics deal with them.

2.1. Mechanism: Explicit versus Implicit

The issue of mechanism in physics arises in a subtle way. Traditional physics expresses many ideas informally since it can draw on our commonsense view of the world. Qualitative physics provides ways to formalize some of these ideas. One such aspect is expressing when different equations are valid. For example, the equations that describe the relationship between volume and temperature are different for a piece of ice, some water in a glass, and steam in a pressure cooker, even though the substance is the same in each case. There are several ways to formalize this knowledge, each involving different levels of ontological commitment.

Bare logical implication is one extreme alternative. Given the right predicates one can correctly specify when equations are applicable, but this alternative provides no organizational structure for physical knowledge. Thus most systems of qualitative physics provide some organizing mechanism, and we shall not consider this alternative further.[1]

The other extreme is to add mechanism, i.e. a special ontological class (or classes) for things that are to be the "agencies of causation". All changes are then stipulated to be directly or indirectly caused by some member of this class of mechanisms. We call these *explicit mechanism* accounts. The current explicit mechanism theories in qualitative physics use continuous or discrete *processes* as the class of mechanisms (Forbus, 1981, 1984; Simmons 1983; Weld 1984). Examples of processes include liquid flow, heat flow, and boiling. These processes operate on objects in various ways, causing changes to occur in them.

A middle position is that there is some mechanistic connection between parameters but there is no separate agency apart from the objects themselves. We call these *implicit mechanism* accounts. Current implicit mechanism accounts organize their laws around *devices* (de Kleer & Brown, 1984; Williams, 1984). Examples of devices include resistors, capacitors, and transistors.

---

[1] An apparent exception is the system of Kuipers (1984), but his goal is to produce a new qualitative mathematics compatible with any scheme for structuring the equations. Based on personal communication, we attribute to him the explicit mechanism position.

A system in the world is modeled by connecting together collections of device models into a network, and changes arise as a consequence of components interacting with other parts of the network.

At first glance explicit-mechanism theories might seem more complex since they posit extra entities. However, they can in fact simplify reasoning. Explicit-mechanism accounts facilitate making and using closed world assumptions (Collins et. al. 1975), which are necessary for beings with finite knowledge and computational resources. An example of a closed-world assumption in causal reasoning is "If nothing in the class of mechanisms I know about is causing a change, then the change cannot occur." If this assumption is violated, an explicit-mechanism account provides a possible way out, namely to postulate a new member of the class of mechanisms. Having a theory of mechanisms limits the search space when faced with contradictions.

2.2. Connectives: Directed or nondirected

The second aspect of causal reasoning about quantities concerns the relationship between the form of qualitative laws and their role in causal reasoning. The formal language of traditional physics is mathematics, typically differential equations. Clearly a qualitative physics must include some qualitative rendering of differential equations, and all of them do. But there are two choices for how equations are used to express causality:

1. *Directed connectives:* The qualitative equations are written as functional dependencies, where the direction of dependence is identified with the direction of causality, i.e.

   $$Q_0 = F(Q_1, Q_2 \ldots Q_n)$$

   will sanction the inference that a change in one of $Q_1, \ldots, Q_n$ can cause a change in Q_0, but not vice-versa.[2] For example, we might write Newton's second law as

   $$a = F / m$$

   to express that we can cause the acceleration to change by changing the force we apply or the object's mass, but not the other way around.

2. *Nondirected connectives:* Qualitative equations are written as constraint equations, and if there are n terms and n-1 of them are known, then the nth term can be calculated. Furthermore, no matter which quantities are involved, logical dependence can be interpreted as causation. For example, in electricity Ohm's law,

   $$V = I * R$$

   states that the voltage across a resistor equals the product of the current and the resistance. We can change the current to cause the voltage to change, and change the voltage to cause the current to change.

The difference between the two positions may be difficult to see at first, since functional dependency implies logical dependency and any constraint equation may be written as a function. The critical fact is that any constraint equation can be written as n different functions, where n is the number of variables in the equation. The directed connective position is that, while each of these different functions may be used in reasoning, only one of them will be distinguished as causal. With directed connectives the role of a qualitative law in causal reasoning is determined by its form, with nondirected connectives the role is determined by how it is used.

---

[2] This choice is not identical to the classical functional view of causation introduced by Mach (as described in Bunge (1979)), because not all functions are identified as causal.

In the current systems of qualitative physics, the choice of connectives has been more or less identified with the choice of explicit versus implicit mechanism. Explicit mechanism theories tend to use directed connectives, and implicit mechanism theories tend to use nondirected connectives. The exceptions are Kuipers (1984) and Williams (1984), who use both. Table 1 shows the 2 X 2 set of possibilities generated by crossing the Mechanism and Connectives dimensions.

We think the reason for this correlation is that explicit mechanism theories provide a notion of "independent" parameters, those directly affected by some mechanism. Effects then propagate outward from these distinguished parameters, like the level of water in a cup changing in response to pouring more water into it. The mechanism thus imposes the direction of causality on the system. Implicit mechanism theories do not identify independent parameters in advance (but see below), so it is hard to pre-judge the way a law will be used causally. Clearly there is no logical barrier to theories which inhabit any and all cells of this table, and later we describe combinations we believe may play signficant roles in human causal reasoning.

Each choice has problems. With nondirected connectives, causal reasoning requires an initial perturbation (such as increasing an input voltage to a circuit). However, people also have causal intuitions about situations even when they do not see the initial perturbation of the objects involved. For example, people are willing to say that the steam they see coming out of a kettle is caused by the boiling occurring inside it, even when they did not see the stove being turned on.

Furthermore, even if an initial perturbation is provided, it seems that some annotations about an equation's causal role are still necessary to avoid inappropriate causal inferences.

Table 1 – Possibilities for Mechanisms and Connectives
The choices along the mechanism and connectives dimensions have not been independent in systems of qualitative physics. The "*" indicates a system which predominately lies in that cell, but allows the other kind of connective as well.

Connectives \ Mechanism	Explicit	Implicit
Directed	Forbus, Kuipers*, Simmons, Weld	
Nondirected		de Kleer & Brown, Williams*

Certain ways of using equations do not correspond to our intuitions about causation.[3] Returning to Ohm's law, we do not assume that increasing the current causes the resistance to change. Even theories which use nondirected connectives must, it seems, rely on some sort of annotation about direction. For example, in de Kleer and Brown's ENVISION program, we find in the description of their confluence heuristic (ibid, page 73),

> "In the specific case of the valve, the converse (where the area is changed) is impossible as the area is an input–only variable of the valve."

This concept of an "input–only" variable constitutes an annotation that violates the nondirectedness of their equations. Unfortunately, there is no theoretical guidance in their account as to which parameters should be so marked.

Directed connectives also have problems. They require the model builder to explicitly state which way causality works in all circumstances. Unfortunately, certain laws can be used causally in more than one direction. Ohm's law, once again, is a good example. If we are reasoning about a voltage source we want to make V be the independent variable, i.e., a change in voltage causes a change in current. If we are reasoning about a current source we want to make I be the independent variable, i.e., a change in current causes a change in voltage. It appears that in principle one can create models that use multiple directed connectives to capture these different causal interpretations, by explicitly specifying a context for each direction. However, writing nondirected constraint equations appears much simpler for these cases.

Some systems, notably those of Williams and Kuipers, attempt to circumvent these difficulties by allowing a mixture of directed and nondirected connectives. The obvious advantage is that the modeler is then free to choose whatever connective seems appropriate. Such freedom, however, can be dangerous. So far no mixed system has provided theoretical constraints on the choice of connective type, which means the choice must be made on an ad hoc basis. We believe such theoretical constraints probably exist, and could be generated by extending an explicit mechanism account.

2.3. Measurement Scenarios

The third aspect of qualitative causal reasoning concerns the sense in which a quantity is said to be changing. Consider again a kettle half–filled with water sitting on a stove. Suppose at some time T_0 we turn the stove on. At some later time T_1 the water begins to boil, and at time T_2 the water has completely boiled away. Most people would agree that from T_0 to T_1 there is a heat flow that causes the heat of the water and its temperature to rise, and that from T_1 to T_2 the boiling is causing the amount of water to decrease and the amount of steam in the room to increase. However, there are four different ways that one might discuss changes in quantities, even within this simple scenario:

1. *Incremental measurements.* We can think about what happens at T_0 when the flow begins. We could say,

   > "The temperature difference between the water and the burner causes heat to flow between them. This will then cause the heat of the water to rise, which will then cause the temperature of the water to rise"

---

[3] See Forbus (1984) for a detailed discussion.

The incremental scenario takes a sequential view of the property changes, demanding that one change occurs before another. In essence, this scenario extends the kind of causality we use for macroscopic discrete events (such as a row of dominos falling in succession after one is pushed over) into the realm of continuous changes. A prototypical example in the continuous realm is following a "piece of liquid" through a hydraulic system.

2. *Discrete measurements:* We can think about the difference between the world at T_0 and at T_2, without considering what happened in between. For example, we might note that there is now no water in the kettle, and the room we are in is more humid than when we started.

3. *Differential measurements:* We can think about what would happen if some property of the situation were different. For example, we might conclude that increasing the temperature of the stove would cause the steam generation rate to increase, and thus the water would boil away sooner. Essentially, we are comparing two possible worlds, related to each other by some change in property or occurrence.

4. *Continuous measurements:* We can think about what is happening during some particular kind of activity. For example, we can say during the interval between T_0 and T_1 that the increase in the heat of the water is causing its temperature to increase, even though both changes are occurring at the same time.

Each of these measurement scenarios has been used in qualitative physics. The incremental scenario was first introduced by de Kleer (1979), and is also used by Williams (1984). The discrete scenario has been used by Simmons (1983) and Weld (1984). The only specific proposal involving the differential scenario is *differential qualitative analysis* (see Forbus, 1984), but it is still relatively unexplored. The continuous scenario is used by most current systems of qualitative physics, including (de Kleer & Brown, 1984; Forbus, 1984; Kuipers, 1984; Williams, 1984).

The discrete, continuous, and differential views each have their distinctive role to play in reasoning about quantities. As argued in (Simmons, 1983; Weld, 1984), often the details of how some change occurs are unclear or irrelevant. A lower-precision discrete model which represents only end-states may best match the available information[4]. Conversely, the continuous view becomes essential when we are concerned with what is happening during a particular activity. The differential view provides information about how things would turn out differently if some change were made, and thus is useful in debugging.

The incremental scenario has considerable intuitive appeal. Unfortunately, so far this scenario has been problematic as a formal model. It requires a distinct notion of time, called *mythical time*. Unlike standard theories of time, mythical time is only partially ordered, and no real time passes between instants of mythical time. While some attempts to clarify the nature of mythical time and its relationship to normal time have been made (de Kleer & Brown, 1984; Williams, 1984) there is still no full formal account. Nevertheless, the incremental scenario is very important for psychological accounts of causality.

3. Psychological Implications

Table 2 summarizes the set of distinctions we have made. It can be seen that there are 16 theoretically possible cells, of which 5 contain AI qualitative physics theories. The "unified causal

[4] Domains for which the best models are discrete are outside the scope of this paper.

Table 2 – The space of causal theories about quantities

Measurement Scenario	Mechanism: Explict — Connectives: Directed	Mechanism: Explict — Connectives: Nondirected	Mechanism: Implict — Connectives: Directed	Mechanism: Implict — Connectives: Nondirected
Incremental				de Kleer & Brown, Williams*
Discrete	Simmons, Weld			
Differential	Forbus			
Continuous	Forbus, Kuipers*			de Kleer & Brown, Williams*

theory" approach to human causal reasoning would be to ask which cell is the one humans use. The questions raised by the "inference family" view of causation are more complex. The questions become:

1. Which cells do people typically use?

2. Are there characteristic patterns of use, such as novice–expert differences?

The remarks which follow are speculative, because these factors have not previously been fully isolated and subjected to systematic psychological investigation. Therefore what follows is a set of conjectures made in the hope of getting the empirical ball rolling. We begin with the first two factors, *explicit vs. implicit mechanisms* and *directed vs. nondirected connectives.*

These dimensions are highly correlated in theories of qualitative reasoning. Table 2 shows the concentration of AI theories in the two outer columns: explicit mechanisms with directed connectives, or implicit mechanisms with nondirected connectives. Here we ask how each of the four possible combinations of explict/implicit mechanisms and directed/non-directed connectives (i.e., the columns of Table 2) might be manifested in human reasoning about quantities. We suspect that (a) examples of all four combinations can be found in human reasoning; (b) the outer two columns, heavily explored in qualitative physics, do in fact represent common human causal arguments; but (c) the implicit mechanism/directed connective combination, unexplored by qualitative physics, also represents an important class of human causal reasoning.

We will illustrate the four classes with the familiar domain of car engines.

1. *Explicit mechanisms with directed connectives:* e.g.

   "Opening the throttle increases the flow rate of gas to the engine, which causes the engine to work faster."

   Here the reasoner uses a set of processes to make causal inferences. This is the kind of reasoning modeled by Qualitative Process theory.

2. *Implicit mechanisms with directed connectives:* e.g.,

   "Driving faster causes fuel consumption to increase."

   Instances of this class are sometimes instances of diSessa's *phenomenal primitives* or of what we call the Causal Corpus (diSessa, 1983; Forbus & Gentner, 1986).

3. *Explicit mechanisms with non-directed connectives:* e.g.,

   "Inside the engine, during the compression stroke the decrease in volume inside the cylinder causes the pressure to increase. During the expansion stroke, the increased pressure due to combustion pushes the cylinder down, causing the volume to increase."

4. *Implicit mechanisms with non-directed connectives:* e.g.,

   "The increased voltage at the input causes the current through resistor R_1 to rise. Since resistor R_2 is connected to resistor R_1, this increased current will cause the voltage across R_2 to rise as well."

Since the surface structure of causal arguments is almost always directed, it can be hard to distinguish between directed versus non-directed connectives. We take as evidence for non-directed connectives statements of the form "A causes B" and "B causes A" by a subject about the same situation where no signficant state change has occurred.

Conjecture 1: Of these four, the class most prototypical in human causal reasoning is explicit mechanisms with directed connectives.

For instance, subject OB was asked "If air temperature goes down, what happens to the air pressure (assuming a closed room)?" He replied:

"As the air temperature goes down, the particles move less quickly, so it lowers air pressure."

Here he reasons that the drop in air temperature means a decrease in the speed of the air molecules, which causes a drop in pressure[5].

Conjecture 2: Two exceptions to Conjecture 1 may be experts and young children. While we suspect explicit mechanims and directed primitives are typical for human causal reasoning, we think there are two clear exceptions:

Experts: Expert models in certain domains, such as electronics, appear to be non-directed. Further, experts know how to use constraint equations and conservation laws, and therefore can reason non-mechanistically. We believe experts still use directed connectives when appropriate (such as arguments about force transmission), but also have other options.

Young Children: According to Piaget (1960), very young children lack notions of directed mechanisms; not until about 8 years old do children show fully mechanistic reasoning. Piaget's

[5] See Collins & Gentner (1983, 1986) for a more detailed treatment of mental models of evaporation.

interviews with 4–5 year-old children led him to conclude that they have synthetic, holistic understandings of causality. For example, when asked why a river flows, an adult or older child would answer in terms of the slope, or difference in height between the source and the destination. But five- and six-year-old children give very different answers: "Because people make oars. They push." or "Because there are big fish that swim." or "To make the fountains flow."

Piaget's interviews clearly show a difference between the way young children and older subjects talk about causality, However, recent research has cast doubt on his strong claim that young children lack mechanistic causality. When children are given tasks in familiar, concrete domains, and are asked to make predictions based on causal relations rather than verbally explaining them, even preschoolers show evidence of directed causal mechanisms (Baillargeon & Gelman, 1980; Bullock & Gelman, 1979; Bullock, Gelman & Baillargeon, 1982). Thus, young children are capable of directed mechanistic causality in familar domains. However, in unfamiliar domains such as evaporation or heat-flow, young children may reason non-mechanistically simply because they don't have enough domain knowledge to postulate mechanisms.

Thus, for both explicit mechanism and directed causality, there may be a U-shaped curve. We may find that both extreme novices and advanced experts show non-mechanistic reasoning about quantities, for opposite reasons. But aside from these two extreme groups, we believe commonsense causal reasoning is built around mechanisms. Indeed, we conjecture that even when people don't know the mechanism behind a change they postulate one, as in the Causal Corpus (Forbus and Gentner, 1986).

Now we turn to the choice of measurement scenarios.

Conjecture 3: The incremental scenario is the most basic of the measurement scenarios. In the most natural form of an incremental scenario, events occur in a causal chain, each event causing the next. Along with explicit mechanisms and directed connectives, this kind of sequential causality has considerable introspective appeal as a causal argument. The popularity of Rube Goldberg's elaborately tortuous causal chains is one indication of this idea's appeal. It is a robust way of reasoning about mechanistic causality. There is evidence that it is learned very early, at least for familiar devices (Bullock, Gelman & Baillargeon, 1982).

Examples of the use of incremental scenario occur in people's reasoning about flow systems such as electricity (Gentner & Gentner, 1983). Here subject DDD is answering the question: "Why do electrical plugs have two prongs?"

> "...little negative electrons get forced into that one prong — like suddenly there is this new space for them to go into and they have been lying in wait in my wall, waiting for this prong to come in, and they go into that one prong, and through my light bulb...And then it makes the light...it zips right down the other side of the big loop and you have a current going, and it makes that little light which is why we tricked it into my plugs. That must be why there are two plugs [prongs]. It's a differential and you need two things for there to be a difference."

Notice that the subject followed the electrons on their path from the wall to the light and back. A simpler answer is that current flows because of a voltage difference across the two prongs. But although the subject alludes to these quantities at the end of the passage, her natural approach to the question is to reason incrementally.

As another example, Subject OB is told about a pot of water sitting in a closed room and is asked "As the water temperature goes up, what do you think would be the effect on evaporation rate?" He states:

> "The water temperature going up - that means that the particles are moving faster, so they're more likely to escape, and so therefore we get a higher evaporation rate."

Conjecture 4 (Corollary to Conjecture 3): Incremental causality is so psychologically natural that people often rely on it even inapplicable. In some situations the incremental scenario will lead to incorrect conclusions. In this protocol, for example, subject CL is asked to describe what happens in the simple system shown in Figure 1.

> "All right, as the water emerges from the pump it flows at constant velocity through pipes of equivalent volume... As the pipe constricts, the flow of the water becomes slower... As it emerges from that constriction... there's a surge as the pipe expands; there's a surge in the velocity of the water, and velocity becomes slightly greater than the initial velocity

This response illustrates the difficulty novice reasoners typically have with steady-state systems. Like most novices, CL does not understand Bernoulli's principle: his expectation is that the water will slow down in the narrow pipes, whereas in fact the opposite will occur. In his reasoning, the water starts at the pump and heads into the system, encountering obstacles along the way. It is like turning a piece of stuff loose at the start of a toboggan run and watching its progress. This incremental argument leads to problems because it leads one to believe that the pieces of stuff "pile up" against each other when they reach a constriction, and thus slow down.

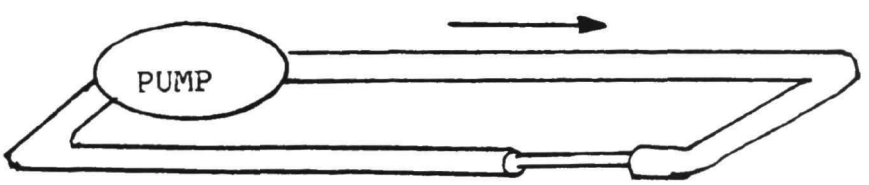


Figure 1 – A simple fluid system
The fluid system shown below consists of a pump and a constriction. Subjects are asked to reason about what happens to various quantities at different parts of the circuit.

Abandoning the incremental model for a steady state model reduces the chance of this plausible error.

The "naive incrementalism" illustrated in this protocol seems to apply to other domains as well. In electricity, for example, novices typically maintain that any of several quantities - voltage, current, power, force, energy, or velocity of electrons - is large at the start of the circuit, just after the battery, and small at the end of the circuit, just before the battery (Gentner & Gentner, 1983).

Conjecture 5: Differential Scenarios are also common in reasoning about quantities. As an example, Subject OB, asked whether an increase in water temperature will affect air pressure, says:

> "OK. If the water temperature goes up, we're going to increase the evaporation rate. If we increase the evaporation rate...we're increasing the amount of water in the air and therefore the air pressure will go up at an increasing rate. Again, it was increasing anyway, so now it's increasing a little faster."

4. Conclusions

We suggest that there are several distinct notions of causality that have psychological force. This multiplicity does not, however, render causal reasoning an inappropriate subject of study. Instead, it means that the form of our analysis must change. Studies of causal reasoning must focus on particular classes of arguments, not causal reasoning in general. While the principles obtained in this way will probably not be as general as those which are the goal of a more general analysis, we can hope that in fact we will have better success. This paper presents an example of such an analysis, examining causal arguments involving changes in physical quantities.

We have abstracted from AI research on qualitative physics three aspects of causal reasoning about quantities, *explicit versus implicit mechanisms*, *directed versus nondirected connectives*, and *type of measurement scenario*. We have shown where the current systems of qualitative physics lie, and shown some new directions such research can take. We have used this analysis to draw some implications for psychology, presenting five conjectures for empirical test.

5. Acknowledgements

Several of these ideas were originally developed in collaboration with Allan Collins and Lance Rips. We also thank Renee Baillargeon and Jerry DeJong for their helpful comments. This research is supported by Office of Naval Research, Contract No. N00014-85-K-0559.

6. Bibliography

Baillargeon, R., & Gelman, R. (1980), Young children's understanding of simple causal sequences: Predictions and explanations. Presented at the meetings of the American Psychology Association, Montreal.

Bullock, M., Gelman, R. (1979), Children's assumptions about cause and effect: Temporal ordering. *Child Development, 50,* 89-96.

Bullock, M., Gelman, R. & Baillargeon, R. (1982), The development of causal reasoning. **The developmental psychology of time.** New York: Academic Press.

Bobrow, D. (Ed.), (1985), *Qualitative Reasoning about Physical Systems*. Cambridge, Mass.: MIT Press.

Bunge, M. (1979), *Causality and modern science*. New York: Dover Publications, Inc.

Collins, A.M., Warnock, E., Aiello, N., and Miller, M. (1975), Reasoning from incomplete knowledge. In D. Bobrow and A. Collins (Eds.), *Representation and understanding*, New York: Academic Press.

Collins, A.M., and Gentner, D., (1986), How people construct mental models. In D. Holland and N. Quinn (Eds.), *Cultural models in language and thought*. Cambridge, England: Cambridge University.

de Kleer, J. and Brown, J. (1984), A qualitative physics based on confluences. *Artificial Intelligence*, **24**.

DiSessa, A. (1983), Phenomenology and the evolution of intuition. In D. Gentner and A. Stevens (Eds.), *Mental Models*, Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Forbus, K. (1984), Qualitative process theory. *Artificial Intelligence*, **24**, pp 85–168.

Forbus, K. and Gentner, D. (1986), Learning physical domains: Towards a theoretical framework. In R. Michalski, J. Carbonell, and T. Mitchell (Eds.), *Machine Learning, Volume II*, Los Altos, California: Morgan Kaufmann Publishers, Inc.

Gentner, D. and Gentner, D. R. (1983), Flowing waters or teeming crowds: Mental models of electricity. In D. Gentner and A. Stevens (Eds.), *Mental models*, Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Hayes, P. (1979), The naive physics manifesto. In D. Michie (Ed.), *Expert systems in the micro-electronic age*, Edinburgh: Edinburgh University Press.

Piaget, J. (1960), *The child's conception of physical causality*. Paterson, New Jersey: Littlefield, Adams & Co.

Simmons, R. (1983, December), Representing and reasoning about change in geologic interpretation. MIT Artificial Intelligence Lab TR–749.

Williams, B. (1984), Qualitative analysis of MOS circuits. *Artificial Intelligence*, **24**.

Weld, D. (1984), Switching between discrete and continuous process models to predict genetic activity. MIT Artificial Intelligence Lab TR–793, October.

NEITHER PICTURES NOR PROPOSITIONS:

THE INTENSIONALITY OF MENTAL IMAGERY

DANIEL REISBERG

REED COLLEGE

DEBORAH CHAMBERS

NEW SCHOOL FOR SOCIAL RESEARCH

We explore several implications of the claim that mental images are mental representations, i.e. that imaging creates an intensional context. The experiments show, first, that images are entirely unambiguous in what they represent, and, second, that how an image is understood places limitations on what the image is likely to call to mind. We argue that images have two different kinds of properties: some by virtue of being mental representations, and some by virtue of being a particular kind of representation, i.e. embodying meanings in a certain (quasi-perceptual) way. We discuss both the relationship between images and pictures, and between images and other forms of representation.

We begin with a simple claim: Mental images are conscious mental representations, and this has important consequences for how images function. In this paper, we spell out just part of what this innocent-sounding claim implies. While our primary agenda is to show that this view is correct, we also intend to show that this view is genuinely different from how others in the field conceive of images, and, most important, to show how much is at

---

We wish to thank Ed Casey and Friderike Heuer for their comments. In addition, both this work and our thinking in general have been deeply influenced by Julian Hochberg and Irvin Rock, and we gratefully acknowledge that debt.

stake in the argument.

One way to think about mental images is to think of them as being just like pictures, so that imagery is in many ways just like perception. There is much to recommend this view, including both common sense introspection and the vast amount of data from imagery research. Those data, for example, indicate that how we inspect an image (how we scan across it, or how we zoom in to see detail) seems to match how we inspect pictures. Likewise, we seem to be influenced in comparable ways by imaged and actual stimuli. The imagery medium itself reveals many picture-like qualities: Images, like pictures, respect spatial relations in a certain way. Aspects that are salient in a picture are also prominent in an image, etc.

However, there is one regard in which images are not like pictures. A picture is a physical thing, with an existence independent of the perceiver. If we want to know what a picture represents, we must inspect it in order to form some interpretation. And that is the critical part: Pictures must somehow be interpreted. Among other consequences, this creates the possibility for misinterpretation, and, since many interpretations might be possible, a picture can be ambiguous, as for example the Necker cube is.

One might claim that all of this is true for images. On this conception, imagery begins with some raw material, an array of lines or points. To figure out what the image is an image of, we interpret it, through a process related to perceiving. Kosslyn (1980, 1983) offers one version of this view; similar arguments can be found in Finke (1980).

There is, however, an alternate account of how mental images are comprehended. (Cf. Casey, 1976; Chambers and Reisberg, 1985; Fodor, 1981; Kolers and Smythe, 1984. Reed, 1974, and Hinton, 1979, have also offered similar conceptions.) At the core of this argument is the claim that images, as mental representations, only exist through our understanding of them; hence the image and the understanding are inseparable. There is no free-standing icon in imagery in need of interpretation, and there is no interpretive process.

To put this differently, images are embodiments of thoughts. Consider for a moment what it means to have a thought, for example, a thought about tigers. It would be absurd to say, "Right now, I believe I'm thinking about tigers, but this belief might be false." It might be true that I do not know much about tigers, making the thought impoverished. It may turn out that some of the things I know about tigers are false, so that my thought will be counterfactual. Nonetheless, if I understand the thought as being about tigers, then it is about tigers, because it is only through my understanding of it that the thought has any definition at all. There is certainly no need to interpret the thought to learn what it is about; there is no way to be mistaken about what a thought refers to, and no possibility for ambiguity in what a thought is about.

Images are one of the forms thoughts can take. Thus, images are thoughts, or, in the customary terms, are mental representations. When an

image comes into being, it comes into being understood in a particular way, as an image of some particular thing. Just as with thoughts in any other form, there is no interpretive process needed to learn what the image depicts, and there is no way to be confused or mistaken about an image's contents.

Our research began with an examination of an assertion central to this view: the claim that images, like thoughts in general, cannot be ambiguous. To test this claim, however, we must deal with a complication: Whether or not one can reinterpret an image, one can certainly replace an image. Therefore, one could image the Necker cube (Figure 1), for example, by first imaging Cube A, then B, then A again, and so by a succession of replacements seem to be reinterpreting a single image. This possibility is easily removed: To replace an image, one needs to know with what to replace it. Thus the critical question is not simply whether one can reparse an image by imposing an already familiar scheme; the question instead is whether one can discover an unanticipated, uncued shape in an image.

Subjects in our first experiment were shown pictures of several ambiguous figures, to make certain they understood figural reversal. Subjects were next shown the test stimulus (the duck/rabbit, Figure 2A) and were asked to form a "mental picture" of this figure, so that they would be able to draw it later on. We took a variety of steps to ensure that subjects were unfamiliar with this figure, among them excluding psychology students from the experiment.

The test stimulus was shown only for five seconds. Pilot data indicated this was sufficient for encoding the figure, but not enough time for naive subjects to find both construals. This is critical, given our concern about excluding image replacement. Hence we want to ensure that the first reversal does not happen while viewing the picture, so that any reversals from the image (if they occur) will reflect bonafide discoveries.

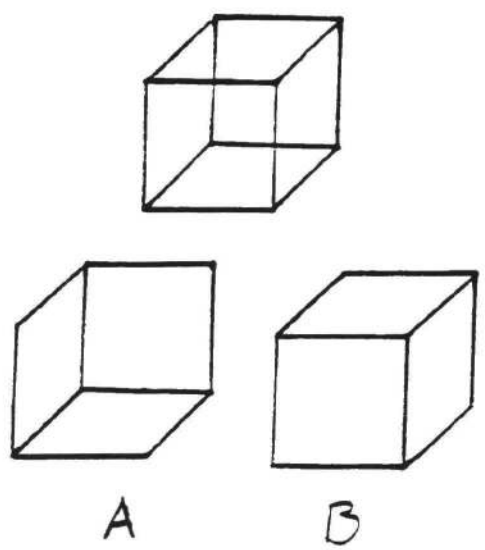


FIGURE 1: THE TWO ALTERNATE VIEWS OF THE NECKER CUBE.

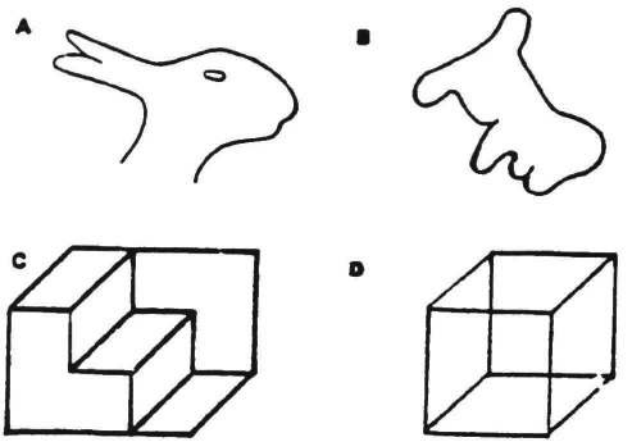


FIGURE 2: THE AMBIGUOUS FORMS USED IN THE STUDIES (SEE TEXT FOR DETAILS).

Next, subjects were shown the chef/dog (Figure 2B). We hoped this would make it totally clear to subjects what their task was. Like the duck/rabbit, this form is a simple line drawing; both figures reverse in a similar fashion. Subjects were told that shifting their gaze from the lower left corner of the chef/dog to the lower right might help in finding the alternate construal.

Subjects were then asked to recall their image of the duck/rabbit. All had either seen the duck or the rabbit initially; they were now asked to identify the alternate view. We gave the subjects a standardized series of prompts and hints, asking if they had a "clear mental picture" and whether they could find an alternative "in the same way that (they) did for the chef/dog." Subjects were urged to look at the "east corner" of the imaged figure, then the "west," and so forth. (These cues are effective in eliciting reinterpretations if there is an actual stimulus available, rather than an image.)

Finally, subjects were given a piece of paper and asked to draw a picture of the imaged stimulus, then to inspect their own drawing until they found the alternate construal. If needed, subjects were given the same cues that had been used with the image.

Exactly 100% of the subjects failed in the imagery task. That is, subjects never discovered the duck in a rabbit image or the rabbit in a duck image. In complete contrast, 100% of the subjects succeeded, a moment later, in reinterpreting their own drawing. Thus, the subjects did have an adequate memory of the figure, and did understand our task.

We have reproduced this result several times, with minor procedural changes. We have tried changing the instructions, emphasizing to subjects that they should encode a literal, unbiased copy of the figure. (To this

end, we give subjects a brief lesson about distortion effects in memory, and urge them not to fall prey to these.) We have also tried giving more initial practice with figural reversal, and we have examined test figures other than the duck/rabbit (i.e. the Schroder staircase and the Necker cube, Figures 2C and 2D). The results are unchanged by any of these (Table 1).

These are obviously strong results, and warrant caution in their interpretation. In an early report of these data (Chambers and Reisberg, 1985), we consider many of the ways one might try to explain away these findings, and the results survive all of the attacks. One of those considerations, however, is worth a brief mention. Subjects' success in reinterpreting their own drawings provides an important support for our claims, indicating that subjects did encode an accurate copy of the stimulus. To ask whether these drawings really were ambiguous, we showed the drawings to a new group of subjects, subjects who had never seen the original figure. In a procedure similar to that already described, these new subjects were able to discover both interpretations of the drawings, strengthening the claim that the earlier subjects were in fact creating an ambiguous drawing from an unambiguous image.

We have also obtained parallel results with auditory imagery (Reisberg, Smith and Sonenshine, 1986). As our ambiguous stimulus, we employ rapid repetitions of the word "stress." When these repetitions are aloud, the resulting soundstream is ambiguous about the locations of the word boundaries, creating a stimulus which could be stress, dress, rest or tress. Our procedure with this stimulus is modeled after the one already described: we acquaint subjects with ambiguous auditory stimuli (using repetitions of life, with the soundstream ambiguous between life and fly). We then ask subjects to imagine the repetitions of stress, and ask them to find more than one interpretation of it, just as they did with life. In analyzing these results, we need to take steps to rule out both guessing strategies and subvocalization, but once this is done, the results echo those we have been discussing: Just as with visual imagery, subjects reliably fail to reinterpret the auditory image.

TABLE 1: SUMMARY OF RESULTS WITH VISUAL AMBIGUOUS FIGURES

Experiment and test stimulus	N	Number of reversals: From image	Number of reversals: From own drawing
1 Duck/rabbit	15	0	15
2 Duck/rabbit	10	0	10
4 Duck/rabbit	10	0	10
Necker cube	10	0	10
Schroder staircase	10	0	6

(After Chambers and Reisberg, 1985)

Unfortunately, though, there are many observations which do not seem to fit with this conception of imagery. In many mental rotation studies, for example, subjects must decide if a stimulus is a letter or a mirror-reversal of one. The data strongly indicate that, in making this decision, subjects first rotate the image; only then can they decide what the image depicts. Thus the understanding of this image seems to be arriving rather late: Until the rotation is complete, the subject does not know what the image is an image of, in seeming contradiction to our view.

Our account of this (and related cases) appeals to a notion already mentioned, namely, image replacement. It is clear that images can serve as memory cues, evoking other images or other thoughts. In the mental rotation case, one first has in mind an image of a particular form. That image may or may not evoke the thought of some letter. If it does, one replaces the shape image with a letter image, with the replacement, as it turns out, isomorphic with its predecessor.

While images can call forward other ideas in this way, it is critical to keep in mind that this evokation begins with an unambiguous representation. The image is understood in a certain way, and this gives the image a specific phenomenal appearance. It is this phenomenal appearance which governs what the image resembles, and, finally, it is the resemblance pattern which guides what an image will call to mind. We stress that this cascade of implications begins with how the image is understood, and so, in this way, the understanding places limits on what the image will evoke, and on what we can learn from or about the image. To put all this differently, the notion of image replacement seems initially troublesome for our view, since it appears to render the view immune to falsification. As we will see, though, this notion is not at all a source of trouble; far from it, image replacement provides a means of showing the importance of how an image is understood.

What exactly does it mean to understand an image in a certain way? This is in several regards an empirical issue; our experiments explore the obvious suggestion that the understanding which characterizes <u>imagery</u> is the same as that which characterizes <u>perception</u>; the results clearly favor this view.

In unpacking this argument, it may be simplest to begin with the data. We tell subjects in the next experiment that we are studying memory for abstract forms. Subjects are shown a succession of shapes, each for 5 seconds. After the shape is removed, subjects are asked to form an image of it, and then to rotate the image by a certain amount, sometimes 90 degrees and sometimes 180 degrees, and so on. Finally, subjects draw a picture of the form in its rotated position.

This sequence is designed to set the subject for seeing abstract, unidentifiable forms, so that we can now smuggle in our test figure: The tenth shape in the series, presented with no special notice, was an outline drawing of Texas, rotated so that its eastern edge was at the top of the picture (Figure 3). Subjects encoded this shape, then imaged it, and finally rotated the image 90 degrees, so that they were now contemplating



FIGURE 3: TEST STIMULUS FOR THE ORIENTATION STUDIES.

an image of a correctly righted map of the Lone Star State. Then, instead of asking the subject to draw a picture, as they had been doing, we tell them that this shape "resembles a familiar geographic form," and we ask them to identify that form.

This experiment rests on two premises, both extrapolations from the data of perception. First, phenomenal shape depends heavily on an assignment of orientation. That is why, for example, diamonds look differently from squares, as can be demonstrated in a variety of ways. (These include resemblance patterns, subjects' ability to judge whether the figure's corners are right angles, with much less sensitivity to this in diamonds, and so on.) Second, once the perceptual system has assigned an orientation, there seems to be resistance to changing this assignment. As Mach noted many years ago, if one rotates a square, it does not become a diamond; it becomes a tilted square. (This presumably plays a role in how we recognize objects despite changes in retinal orientation.)

Subjects will presumably understand our test stimulus as being oriented such that the perceived top is the side topmost in the drawing. This specification about how the form is to be understood will be part of the image, and will influence the subjective appearance of the image. Rotating the image will not change this specification, if imagery is like perception in this regard. Given all this, we claim that, by virtue of having an orientation different from that of Texas, the image is a different shape. Geometrically, the image and Texas are isomorphic, but, psychologically, they represent different forms. Hence, the image does not resemble Texas, and will not call Texas to mind.

The data are quite clear-cut. Of the 15 subjects, 100% failed to identify Texas in the image, even though the image has been rotated into, so to speak, the Texas position, and despite the cue that the image resembles a geographic form. We know in addition that the subjects did adequately

encode the shape: After the subjects had tried unsuccessfully for 60 seconds to identify the shape, we asked them to draw a picture of their image. Eight of the 15 subjects identified Texas in their drawings.

Is orientation really critical in the failure to recognize Texas? If so, we should be able to change the outcome of this procedure if we change how subjects understand the image's orientation. We know from many perceptual studies that rotation by itself does not change how a form is understood; we relied on this in the study just described. But we also know that deliberate intention can step in and take over in reassigning orientation, e.g. if subjects are told directly to think about a form as having a different top (cf. Rock and Leaman, 1963; Attneave and Reid, 1968). Our next study exploits this fact, using the same procedure as the previous study, but with a slight change in instructions: Instead of asking subjects to rotate the image of each figure, we simply told them to think about each as having a new top, making the left side the figure's top, or the right side, and so on.

As before, subjects were shown a series of shapes, each for 5 seconds. Subjects imaged each, then were asked to reassign the form's top. Sometimes we told them to think of the left edge as the top, sometimes the right, all matched to the previous experiment. Subjects then drew each form, with its new top at the top of the drawing. For the test figure, again Texas, subjects were asked to reassign the left side as the top, setting the map upright. Once again, we told them that the shape resembled a familiar geographic form, and asked them what the form was.

In sharp contrast to the first study, 7 of the 15 subjects, almost half, identified Texas in their image (see Table 2). Thus, orientation does seem to be the key. If we change how the subject understands the image's orientation, we change the results.

We have some other data which corroborate the Texas results: We show subjects the shape in Figure 4, with the same kind of cover story, so that they encode it as merely an abstract form. If we tell them to rotate the image, they never discover anything. If we tell them to think of the bottom edge as being the top, they discover the old man, but not the map of the U.S. If we instead tell them to think of the left edge as being the top,

TABLE 2: SUMMARY OF RESULTS FOR THE ORIENTATION STUDIES

Instruction	N	Number of subjects who recognize Texas	
		From image	From own drawing
"Rotate"	15	0	8
"Reassign top"	15	7	11

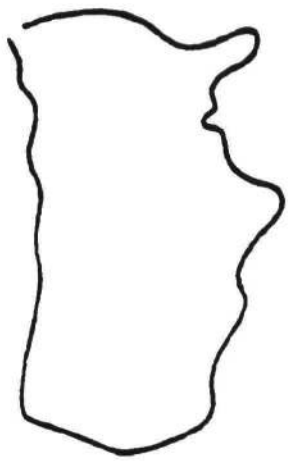


FIGURE 4: THE "OLD MAN / UNITED STATES" FIGURE

they discover the U.S., but not the old man. All of this obviously mimics the Texas results already described.

What is going on in these data? We have already mentioned the core of the argument: Whether an image will be recognized as familiar, or what the image will evoke from memory, depends on how the image is understood. If the form is isomorphic with some previously viewed shape, but understood differently, then it will fail to remind the subject of the previously viewed shape. But we need to connect this to our earlier comments about mental representation, and we also need to clarify what it means to understand a form in a particular way. Let us return to the example we used earlier: Imagine thinking about a tiger. We argued before that this necessarily means that you know that you are thinking about a tiger. But this does not mean merely that you are associating the label "tiger" with your thought. Instead, by knowing that your thought is a tiger thought, you know that this is a real beast, that "tiger" is the name others call it, that this is a concept with which you are familiar, about which you have further knowledge, etc., etc. In short, you are attaching meaning to the notion, tiger.

For contrast, we could find someone who has never heard of tigers, and we could tell that person: Imagine a fictional beast with a cat-like shape, stripes, etc. Then we could tell this person to call this beast "tiger," a label which, for this person, is a made-up word. Would that person have a "thought without meaning," without knowing what a tiger is? Clearly not. That person would know that a tiger was a fictional beast, created by the imagination with these instructions, and that would be the meaning of tiger for this person. Moreover, this person would be having a different thought than you or I do, when we think about tigers.

It is this background of associations and knowledge which is involved in saying that mental representations are meaningful, that they are trans-

parent to the understanding. We can describe this as a phenomenologist might: Thoughts are embedded in a context of understanding, against, so to speak, a particular mental landscape, what Husserl called a set of horizons, or with what William James called a particular "fringe of consciousness," defining that consciousness. Or we can describe this by noting that mental representations are referentially opaque. Hence we cannot describe the representation merely by pointing to the thing in the world which is represented; we instead need to ask how the represented thing is thought about, what the representer knows about it. To take the tiger example, the person thinking about the fictional tiger, on the one side, and you or I, on the other side, are representing what turns out to be the same beast; nonetheless, we understand it in different ways, and we are having different thoughts.

How to apply this to images? When one thinks about a form, one thinks about it in ways that differentiate that form from other forms. As in the tiger case, this does not amount merely to placing a label on the form; instead, the form is embedded in a context of understanding that defines what the thing being represented is. If the form is thought about in a new way, then it is a new form. As a minimal implication, a form understood differently should look different, should have a different phenomenal shape. Look again at the Texas figure, shown in Figure 5. The left shape does, for most viewers, look like a different form from the right one, just as a diamond looks different from a square. To be sure, one can force oneself to think about the right shape as being sideways -- i.e. impose a different understanding. In this case, one sees immediately that the shapes are isomorphic. But if one does not resist the orientation given in the figure, the two outlines look like two different shapes. We do not need to rely, though, on phenomenal impressions, since these changes in perceived form can have strong impact on similarity relations. Figure 6, borrowing a demonstration from Goldmeier, makes the point clearly.

While we have focused here on the specification of orientation, the data of perception indicate that other aspects also contribute to the perception of form. And it is indeed the data of perception which are relevant, as a quick example will illustrate: One might expect that symmetry will be an important property of percepts, immediately noticed, playing a key role in resemblance and so on. But, as Figure 7 shows, symmetry in general does not have these

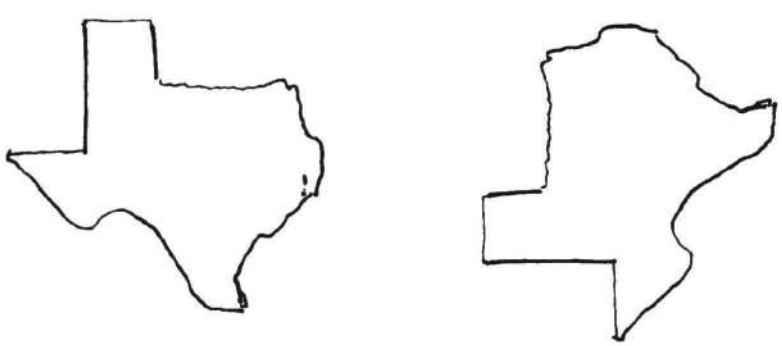


FIGURE 5: A CHANGE IN ORIENTATION ALSO CHANGES PHENOMENAL SHAPE.

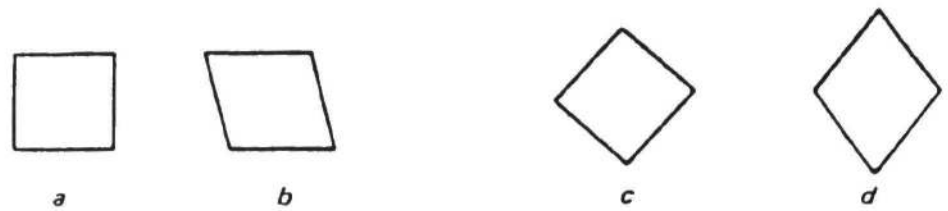


FIGURE 6: A AND C ARE IDENTICAL EXCEPT FOR ORIENTATION; LIKEWISE FOR B AND D. THE CHANGE IN ORIENTATION CAUSES A CHANGE IN PHENOMENAL SHAPE, AND A CONCOMITANT CHANGE IN SIMILARITY PATTERNS.

properties; it is only symmetry about the vertical axis which seems to characterize a form. Symmetry around other axes is considerably more difficult to detect, and does not seem to influence resemblance patterns.

Happily, we do not need to go out and collect the relevant data, since much of it is already available, thanks largely to Gestalt psychology and its descendants. The full set of appearance specifications (and we count a half-dozen or so) includes unit formation, figure/ground assignment, orientation, assignment of relative depth, and so on. In all these ways, perception goes beyond the information given, imposing meaning on a form. These specifications literally shape how the form looks, and changing these will change appearance. This in turn will change resemblance patterns, and so will change what a form is likely to call to mind.

Just as the specifications shape percepts, they also shape images. Hence changing how an image is understood will create a new phenomenal form. As Casey (1976) puts it, "to imagine something differently is to imagine something different." As a consequence, while the present experiments exploit orientation, similar imagery demonstrations should be possible with other specifications. One of our current studies is exploring this point, but there are

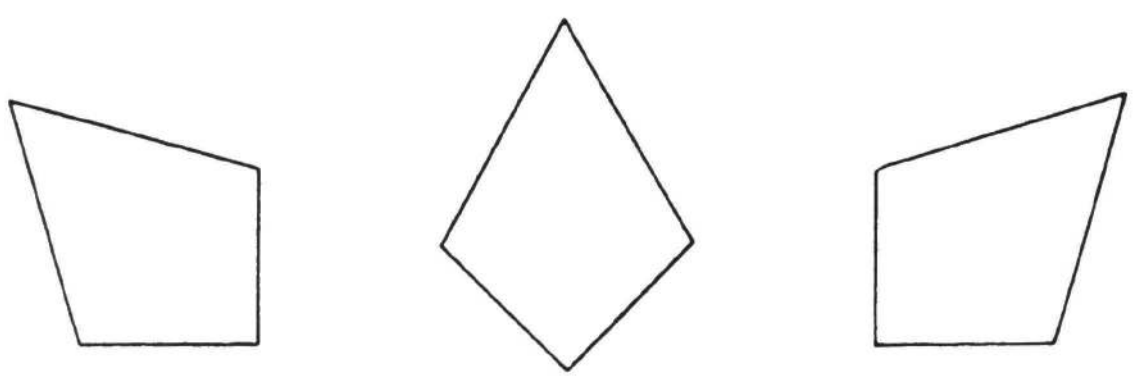


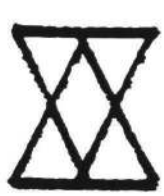
FIGURE 7: IT IS SYMMETRY AROUND THE VERTICAL AXIS WHICH INFLUENCES FORM.

already some relevant data in the literature (e.g. Hinton, 1979; Reed, 1974; Slee, 1980). In Slee's task, for example, subjects are first shown a series of shapes, including a Roman numeral twenty (Figure 8). Later, subjects are shown probe figures (such as the parallelogram) and asked whether these were contained in any of the earlier forms. Subjects are explicitly urged to imagine various constructions around the probe shapes, in order to discover if they can "build" their way back to any of the shapes in the memory set.

When the probes and memory figures differ in their organization, this task is very difficult, and failure rates are high. This obviously parallels our own findings: By virtue of the different organization, the probe does not resemble the memory figure, and so will not call it to mind. (Interestingly, the Reed and Slee data differ from ours in one respect: Our data yield 0% success rates; they report occasional successes. Neither Reed nor Slee, however, have any check in their procedures on how subjects initially encoded the figure. It seems entirely possible that the successes come from subjects who encoded the form in a manner consistent with the probe.)

In these various ways, then, mental images behave as mental representations must behave, with their function bounded by their particular meaning. In fact, by being a property of mental representations in general, rather than just of images, we should be able to find parallel demonstrations in other domains. We have been discussing the parallels between imagery and perception, but we can also find comparable effects in more distant regions. For example, consider Tulving's encoding specificity demonstrations. In these studies, subjects are led to understand a verbal stimulus one way during the learning process and a different way during the test. Because of this, subjects fail to recognize a word seen just minutes before, just as our subjects failed to recognize the familiar shape of Texas. And this parallel should exist, inasmuch as certain properties are entailed by being a mental representation, whether the representation is an image or in some other form, whether the thing represented is a shape or a word.

Subjects are shown a series of figures, including this "Roman-numeral 20".



Subjects are later asked, "Was this part of one of the previous figures?"



FIGURE 8: SUMMARY OF SLEE PROCEDURE.

We do not, however, wish to suggest that these various forms of mental representation are interchangable. Images, as one mode of representation, are a particular means of representing, of embodying an idea, and they employ a particular medium. The nature of imagery demands that some aspects of the idea be spelled out, while other aspects can be omitted. For example, to abuse the tiger example one last time, if one thinks about the tiger without imagery, one may or may not include in the thought whether the tiger is standing or sitting, whether it is facing to the left or to the right, and so on. If, however, one images a tiger, one must commit oneself on these aspects, but give no thought to whether the tiger is hungry, for example.

In other words, there are properties that images must have because they are mental representations -- they cannot be ambiguous, they cannot be indeterminate in certain regards. At the same time, there are properties that images have because they are images -- they must spell out position in a way that other representations may not, they spell out a viewing perspective, etc. It is this second set of properties, the ones that make images images, which is tapped by current imagery paradigms. And these are important properties: Since the image represents the tiger in a certain shape, one might be reminded of things sharing that shape. Since the image might leave out the fact that tigers are predators, one might not be reminded of other predators, and so on. Thus, if the same idea were embodied in a different representational guise, one is likely to fill in a somewhat different set of aspects, creating a different pattern of resemblances, and so a different pattern of evokations.

This obviously invites a flood of questions: What are these modes of representing? How many of them are there? How are they different or how are they alike? You will recognize these as just the issues at stake in the Kosslyn / Pylyshyn exchanges, and we believe they are precisely the right issues to be pursuing. Our view of imagery draws elements from both of the participants in these exchanges: With Pylyshyn, we are arguing that images are not only cognitively penetrable, they are cognitively penetrated. With Kosslyn, we are arguing that there are special properties which identify imagery. More, we share Kosslyn's view that the representational medium underlieing imagery is also the medium underlieing perception. That obviously fits both with our data and the many commonalities between images and percepts.

We have been focusing on the parallels between images and percepts with regard to resemblance and evokation, so it is worth circling back to the earlier discussion of ambiguity. We argued at length that images are totally unambiguous; we want to stress that the same is true for percepts. The percepts associated with the Necker cube are not indeterminate about depth, they specify one configuration or another. Likewise for the Rubin vase/face: If one perceives the vase, the perceived stimulus has a particular depth arrangement, the ground is "completed" so that it continues behind the figure, and so on. Hence the perceived vase has different properties from the face; it also evokes different memories (including a different label). Most important, subjects will deny ever having seen the face before if, on some prior occasion, the figure was organized as the vase. By any of these indices, the vase and face are simply two different percepts, each by itself unambiguous.

There clearly is ambiguity in the vase/face, but it is in the stimulus, not in the percepts derived from it. This in fact allows us to deal with an earlier loose end: When subjects in our early experiments imaged the duck/rabbit, they reliably failed to discover both interpretations of the form. But when subjects are perceiving the duck/rabbit, they routinely reinterpret the shape. This contrast, we believe, comes directly from the presence of a stimulus in perception, i.e. an input which is independent of our understanding of it. Hence it is possible to set our understanding aside and return to this "raw material," potentially allowing us to arrive at a different percept. Imagery lacks this equipotential stimulus, making images (in this regard) more rigid, more inflexible than percepts.

This draws us to an important concluding point. While we have emphasized the commonalities between images and percepts, we also need to note the important differences between imagery and perception. Perception begins with a stimulus, and is largely occupied with the understanding or identification of that stimulus. At the end of this process is the percept, a mental representation, and hence all that is entailed by being a mental representation. It is thus the products of imagery and perception which are similar, but the processes leading to each are clearly different. The image, quite unlike the percept, is not the result of some interpretive process, because, as we have argued throughout, there is nothing to be interpreted. Thus, images and percepts are parallel representations with, we believe, virtually non-overlapping histories.

REFERENCES

Attneave, F. and Reid, K. (1968) Voluntary control of frame of reference and shape equivalence under head rotation. Journal of Experimental Psychology, 78, 153-159.

Casey, E. (1976) Imagining: A phenomenological study. Bloomington, Ind.: Indiana University Press.

Chambers, D. and Reisberg, L. (1985) Can mental images be ambiguous? Journal of Experimental Psychology: Human Perception and Performance, 11, 317-328.

Finke, R. (1980) Levels of equivalence in imagery and perception. Psychological Review, 87, 113-132.

Fodor, J. (1981) Imagistic representation. In N. Block (Ed.), Imagery, pp. 63-86. Cambridge, MA: MIT Press.

Hinton, G. (1979) Some demonstrations of the effects of structural descriptions in mental imagery. Cognitive Science, 3, 231-250.

Kolers, P. (1983) Perception and representation. Annual Review of Psychology, 34, 129-166.

Kolers, P. and Smythe, W. (1984) Symbol manipulation: Alternatives to the computational view of mind. Journal of Verbal Learning and Verbal Behavior, 23, 289-314.

Kosslyn, S. (1983) Ghosts in the mind's machine: Creating and using images in the brain. N.Y.: Norton.

Kosslyn, S. (1980) Image and mind. Cambridge, MA: Harvard University Press.

Monk, P. (1979) Orientation and symmetry. Perception and Psychophysics, 29, 230-234.

Reed, S. (1974) Structural descriptions and the limitations of visual images. Memory and Cognition, 2, 329-336.

Rock, I. and Leaman, R. (1963) An experimental analysis of visual symmetry. Acta Psychologica, 21, 171-183.

Slee, J. (1980) Individual differences in visual imagery ability and the retrieval of visual appearances. Journal of Mental Imagery, 4, 93-113.

METHODS FOR EVALUATING THE VALIDITY OF HYPOTHESIZED ANALOGIES*

John Clement

Department of Physics and Astronomy
University of Massachusetts/ Amherst

ABSTRACT

Evidence is presented indicating that spontaneously generated analogies can play a significant role in expert problem solving. Since not all analogies are valid, it is important for the subject to have a way to evaluate their validity. Three methods for evaluating analogical validity are identified using observations from thinking aloud problem solving protocols as well as examples from Newton and Galileo. In particular, this paper focuses on an evaluation strategy called bridging that has been observed in solutions to both science and mathematics problems. In constructing a bridge, the subject finds an intermediate case that is seen as "in between" the analogous case and the problem situation because it shares important features of both. Many of the bridges observed appeared to be novel inventions created by the subject. These empirical studies have led to the construction of a more detailed theory for how analogies can be used effectively in instruction. Some of the strategies observed in experts appear to have high potential for helping science students overcome persistent misconceptions in the classroom.

A number of authors have emphasized the important role of analogies in problem solving and learning, including Gentner and Gentner (1983), Rummelhart and Norman (1980), and Gick and Holyoak (1980). Expert subjects have been observed to generate and use analogies spontaneously during problem solving (Clement, 1981). However, since not all analogies are valid, it is important for the subject to have a way to evaluate their validity. This paper identifies methods used to evaluate and establish confidence in the validity of a hypothesized analogy. "Validity" is used here in a weak sense outside the context of deductive certainty. Since conclusions reached by analogy are viewed as always having a certainty level of less than 100%, establishing validity here means "raising confidence in the appropriateness of the analogy to a high level."

The focus in the study of experts is on qualitative observations from case studies and coarse grained modeling of underlying processes. The primary purpose is to identify important reasoning strategies that occurred across different subjects and problems and to propose an initial description of their form and function. This study contrasts with other studies where the subject is presented with all or part of an analogy and is given the opportunity to use it or complete it since the analogies studied here were not suggested by the experimenter, but were generated spontaneously by the subjects. The paper also attempts to form a direct link between three ordinarily distinct domains: expert problem solving, arguments in the history of science, and strategies

for instruction. Methods used to evaluate analogies in each of these domains will be discussed. The methods have important implications for education, because when students are presented with a "clarifying" analogous case, they may very well not understand why it is analogous to the target, unless they can convince themselves intuitively that the analogy is valid.

USE OF ANALOGIES IN EXPERT PROBLEM SOLVING

I first give two brief examples of the role of analogies in problem solving from thinking aloud studies involving experienced professionals in technical fields. The "Wheel Problem" illustrated in Fig. 1, is a question about whether one can exert a more effective uphill force parallel to the slope at the top of a wheel or at the level of the axle (as in pushing on the wheel of a covered wagon, for example). A spontaneous analogy occurs when the subject spontaneously shifts his attention to a different situation B that he believes may be structurally similar to the original problem situation A, with the intent of possibly applying findings from B to A. Subject S1 compared the wheel to the analogous case of pushing on a heavy lever hinged to the hill (Fig. 2b). He reasoned that pushing at the point higher up on the lever would require less force. He then made an inference by analogy that the wheel would be easier to push at the top (the correct answer). Apparently he used the lever to think about what was happening in the wheel.

WHEEL PUSH OR "SYSIPHUS" PROBLEM

YOU ARE GIVEN THE TASK OF ROLLING A HEAVY WHEEL UP A HILL. DOES IT TAKE MORE, LESS, OR THE SAME AMOUNT OF FORCE TO ROLL THE WHEEL WHEN YOU PUSH AT X, RATHER THAN AT Y?

ASSUME THAT YOU APPLY A FORCE PARALLEL TO THE SLOPE AT ONE OF THE TWO POINTS SHOWN, AND THAT THERE ARE NO PROBLEMS WITH POSITIONING OR GRIPPING THE WHEEL. ASSUME THAT THE WHEEL CAN BE ROLLED WITHOUT SLIPPING BY PUSHING IT AT EITHER POINT.

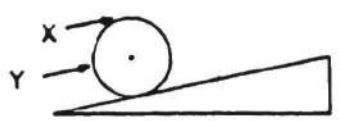


Figure 1

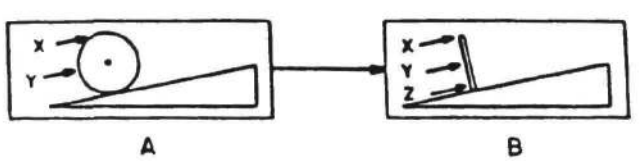


Figure 2

A second example concerns the "Spring Problem" shown in Fig. 3. Several subjects conjectured that this problem was analogous to the simpler case (Fig. 4b) of comparing long and short horizontal rods bent by equal weights hung at their ends. In most cases, a strong intuition that the longer rod bends more was used to predict the correct result that the wider spring stretches more. Thus the bending rod was used as an analogous case for thinking about the spring. That the wide spring will stretch farther seems to correspond to most people's initial intuition about the problem. However, carefully answering the question about <u>why</u> the wide spring stretches more (and explaining exactly where the restoring force of the spring comes from) is a much more difficult task.

Ten professors and advanced graduate students in technical fields were recorded while solving the spring problem in order to study the analogy generation process. They were told that the purpose of the interview was to study problem solving methods and were given instructions to solve the problem "in any way you can". After they reached an answer, subjects were asked to give an estimate of their confidence in their answer. They were then asked if there was any way they could increase their confidence, and this often led to further work on the problem. Probing by the interviewer was kept to a minimum, usually consisting of a reminder to keep talking. Occasionally the interviewer would ask for clarification of an ambiguous statement.

Some of the solutions were quite complex and took up to 90 minutes to complete. All subjects favored the (correct) answer that the wide spring would stretch farther. But the subjects varied considerably in the types of explanations they gave for their prediction. A considerable number of spontaneous analogies were observed, as described below.

<u>SPRING PROBLEM</u>

A WEIGHT IS HUNG ON A SPRING. THE ORIGINAL SPRING IS REPLACED WITH A SPRING

- MADE OF THE SAME KIND OF WIRE,
- WITH THE SAME NUMBER OF COILS,
- BUT WITH COILS THAT ARE TWICE AS WIDE IN DIAMETER.

WILL THE SPRING STRETCH FROM ITS NATURAL LENGTH, MORE, LESS, OR THE SAME AMOUNT UNDER THE SAME WEIGHT? (ASSUME THE MASS OF THE SPRING IS NEGLIGIBLE COMPARED TO THE MASS OF THE WEIGHT). WHY DO YOU THINK SO?

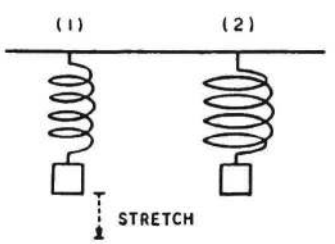


Figure 3

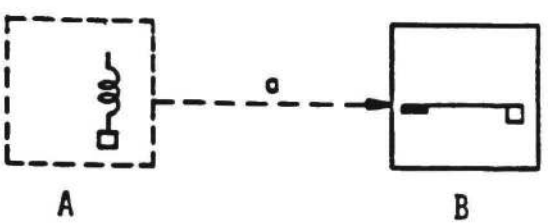


Figure 4

Number of Subjects	10
Total Number of Spontaneous Analogies Generated	38
Total Number of Significant Analogies Generated	31
Number of Subjects Generating at Least One Analogy	8
Number of Subjects Generating a Significant Analogy	7

An analogy was classified as significant if it appeared to be part of a attempt to generate or evaluate a solution, and as non-significant if it was simply mentioned as an aside or commentary. Thus there is evidence that scientifically trained individuals are capable of generating analogies during problem solving.

Subjects indicated varying degrees of certainty about the appropriateness of each proposed analogy. Sometimes they would decide that the new case was not analogous to the original problem in a useful way. In other instances further work would lead them to establish confidence in the validity of an analogy. These observations suggest that the following processes are involved in making a confident inference from a spontaneous analogy (Clement, 1981). (In this description we make a distinction between the analogous case, shown as "B" in Fig. 4, and the analogy relation, shown as the small letter "a".)

a) Generating the analogy. A representation of a situation B that is potentially analogous to A is accessed in memory or constructed. A tentative analogy relation "a" is set up between A and B (e.g. between the wheel and the lever).

b) Evaluating validity: Establishing confidence in the analogy relation. The validity of the analogy relation between A and B is examined critically and is established at a high level of confidence (confidence that the wheel works like the lever).

c) Understanding the analogous case. The subject examines and, if necessary, develops his or her understanding of the analogous case B, so that the behavior of the analogous case is well-understood, or at least predictable (the lever is well-understood).

d) Transferring findings. The subject transfers conclusions or methods from B back to A.

Steps b, c and d above were observed to occur in different orders in different solutions.

METHODS FOR EVALUATING THE ANALOGY RELATION

Methods for generating analogies (step (a) above), are discussed in Clement (1981). The remainder of this paper focuses on step (b) above. Three methods of evaluating validity will be discussed: matching key features; generating a bridging case; and using a conserving transformation. A fourth possible method, analyzing the two cases in terms of a higher level principle, is not treated here. This paper concentrates most on the second method:

generating a bridging case. Examples of evaluation methods will be drawn from a larger sample of 20 expert subjects who solved a variety of problems.

Evaluating the lever analogy for the wheel. A fairly obvious strategy for evaluating the validity of an analogy relation is to assess whether there is a structural match between cases A and B in terms of key features that are important for the behavior of the systems. This involves isolating the key features (especially higher order relationships as defined in Gentner and Gentner, 1983) in each of the cases A and B and comparing them explicitly. In the "Wheel Problem" discussed earlier subject S1 was confident that it would be easiest to move the heavy lever in Fig. 2b by pushing at point X, but he questioned whether there was a valid analogy relation between the wheel and the lever. Can one really view the wheel as a lever, given that the "fulcrum" at the bottom of the wheel is always moving and never fixed? In matching key features, he found a potential mismatch between the stationary fulcrum at the bottom of the lever and the moving fulcrum at the point of contact of the wheel that led him to doubt the analogy. Thus, matching key features is one method for evaluating the validity of analogies.

However, he also used a second, more creative method for evaluating validity. He considered a bridging case in the form of the spoked wheel without a rim shown in Fig. 5C. The spoked wheel allows one to view the original wheel as a collection of many levers. It is a bridge in the sense of being an intermediate case which shares features with both the wheel and the lever. The bridging case reduced the subject's concern about the moving fulcrum issue and raised the subject's confidence in the appropriateness of the lever model.

Presumably, this method works because it is easier to comprehend a "close" analogy than a "distant" one. The bridge divides the analogy into two small steps which are easier to comprehend than one large step. It is easier to see that the real wheel should behave like a rimless spoked wheel, and that the rimless spoked wheel should behave like a lever, than to make this inference in one step. The spoked wheel, then, is an example of a bridging case constructed by the subject in order to establish confidence in the

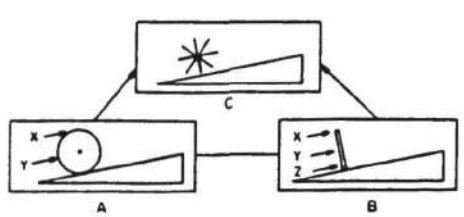


Figure 5

DOUGHNUT PROBLEM

COMPUTE THE VOLUME OF THE TORUS (DOUGHNUT) BELOW WITHOUT TAKING AN INTEGRAL. GIVE AN APPROXIMATE ANSWER IF YOU CANNOT DETERMINE AN EXACT ONE.

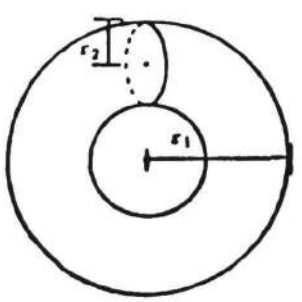


Figure 6

validity of the analogy relation between the lever and the wheel.

Bridging from doughnuts to cylinders. Another example of a bridge occurred in a solution to the mathematics problem (shown in Fig. 6) of finding the volume of a doughnut. Subject S3 first conjectured that the volume might be the same as the answer to the analogous problem of finding the volume of a cylinder (the "staightened out" doughnut). He thought the appropriate length for the cylinder would be equal to the central or "average" circumference of the torus ($2\pi(r_1-r_2)$) but was only "70% sure" of this. However, he then evaluated the plausibility of this choice by considering the bridging case of a square-shaped doughnut shown in Fig. 7. He then showed that the four sides of the square doughnut could be joined end to end to form a single long cylinder with slanted ends. He reasoned that the volume of this long cylinder would be exactly equal to its circular cross section times its central length and that therefore the appropriate length to use in the square doughnut was the average of its inner and outer perimeters. This raised his confidence in his original solution to "85%". He then reached the same conclusion for the case of a hexagonal doughnut, and this raised his confidence to "100%" for the problem. This is an example of a multiple bridge. Thus the bridging cases of a square and hexagonal doughnut helped the subject change his original conjecture about the cylinder into a firm conviction.

Bridging between rods and springs. In the spring problem subject S1 had generated the analogy of a horizontal bending rod. However, he was concerned about the apparent lack of a match between the increasing slope in a bending rod and the constant slope in a stretched spring (that a bug would experience in walking down the spring). This led him to question whether the analogy relation between the rod and the spring was valid.

An extremely successful attempt at a bridge between the case of a single coil of the spring and the bending rod analogy occurred when this subject generated the idea of a square-shaped coil in Figure 8. Visualizing the stretching of a square coil allowed him to recognize that some of the restoring forces in the spring come from twisting in the wire instead of bending -- corresponding to the way in which engineering specialists view springs. This discovery of a new causal variable in the system appears to be an example of a significant scientific insight. In this case the square spring not only helped him evaluate the bending rod model, it eventually

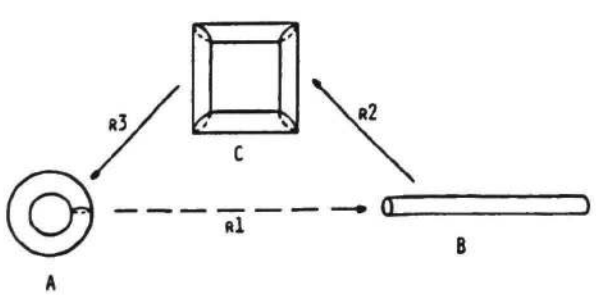


Figure 7

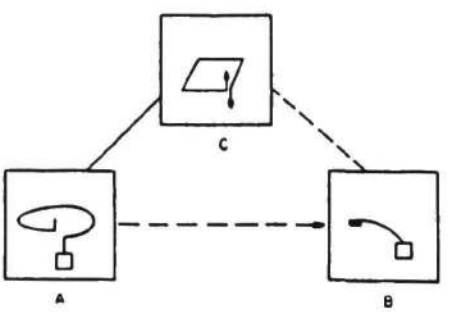


Figure 8

acquired the role of a preferred mental model which changed his conception of how springs work. This significantly increased the subject's confidence concerning the question of whether he had a good understanding of the spring.

Discussion of findings on bridging. Bridging is indicative of the recursive possibilities inherent in reasoning processes like the use of analogies. Since a bridging case is itself an analogous case, it can be described as an analogy used to evaluate a previous analogy (or, more precisely, as a second analogous case used to evaluate the analogy relation between a first analogous case and the original problem). We can summarize our view of bridging as one method for evaluating the analogy relation between a problem situation, A, and an analogous case, B, as follows:

1) The subject constructs a representation for an intermediate bridging situation C which shares important features with both A and B.

2) The subject asks whether the analogy relation between A and C is valid.

3) The subject also asks whether the analogy relation between C and B is valid with respect to the same relationships as in step 2.

4) If the subject can answer yes to both questions with high confidence, this constitutes evidence for the validity of the original analogy.[1]

Here A being analogous to C and C being analogous to B means that A is analogous to B. We can refer to this type of inference as analogical transitivity. However, it should be noted that analogical transitivity is considered a form of plausible reasoning which does not carry the force of a logical implication.

It is clear that many of the bridges discussed here are novel constructions in the sense that they are situations which the subject is unlikely to have studied or worked with before. This indicates that they are invented representations in the form of creative Gedanken experiments that have been constructed, not simply retrieved from memory. In theories of scientific discovery, hypothesis generation is ordinarily seen as a more creative process than hypothesis evaluation. However, in the case of bridging we are faced with a creative, non-empirical evaluation method which generates novel constructions. Thus we seem to have evidence for a type of "creative hypothesis evaluation" process.

ANALOGIES AND BRIDGES IN THE HISTORY OF SCIENCE

Legend has it that Galileo investigated the question of whether light bodies accelerate as rapidly as heavy bodies in an empirical manner by dropping objects from the tower of Pisa, but this legend has come under serious doubt. However it is known that Galileo and his predecessor, Benedetti, did use thought experiments like the following one to argue their side in this issue. Figure 9a shows two equal objects of one unit each being

dropped while Figure 9b shows a heavier object being dropped that is equal to the two smaller objects combined. According to Aristotle the one unit objects will fall much more slowly than the larger object. Galileo claimed that they will reach the ground at nearly the same time. In saying this he was effectively proposing an analogy between cases A and B in Figure 9 to the effect that each body falls according to the same rule irrespective of its weight.

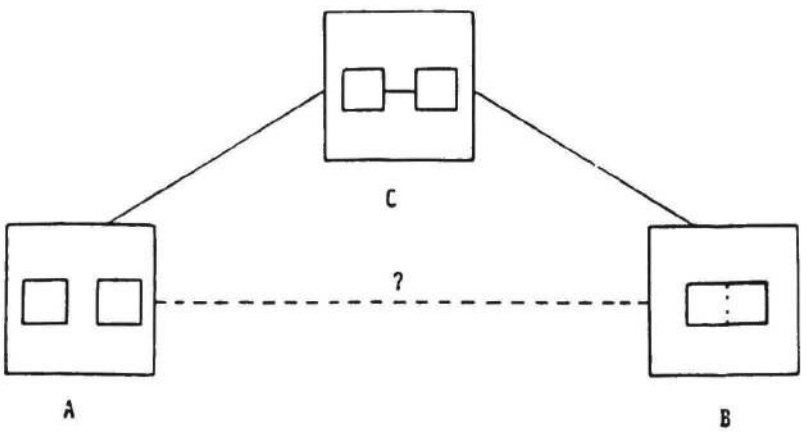


Figure 9

A marvelous bridging case used to support this analogy is the case shown in Figure 9c. The argument was first published by Benedetti (1969) and a similar argument was given by Galileo (1954). Imagine the two unit objects in A to be connected by a thin line or thread. Does the mere addition of this tiny thread, which makes the two objects become one, cause their rate of fall to increase by a large amount? Because this is implausible, the bridge argues that A and B are indeed equivalent with respect to rates of fall. In an insightful Gedanken experiment, the lightest thread can apparently make all the difference.

A bridge used by Newton: One of the most extraordinary scientific analogies of all time was propounded by Robert Hooke and Isaac Newton in the seventeenth century. They claimed that the moon falls toward the earth just as an everyday object (such as an apple) does. To a modern physicist, this may seem more like an obvious fact than a creative analogy, but to advocate such an idea in Newton's time was not an obvious step at all. One has only to imagine the consternation that would be produced by telling someone ignorant of science that the moon is falling.

The proposed analogy relation is represented by the dotted line in Figure 10. Essentially this conjecture says that the same causal mechanism is involved in making the moon revolve around the earth and making an apple fall. A multiple bridge used by Newton to support this analogy in his Principia is shown in Figure 10c. This is the idea of a cannonball fired faster and faster until it enters into orbit around the earth — a premonition of modern rocketry.

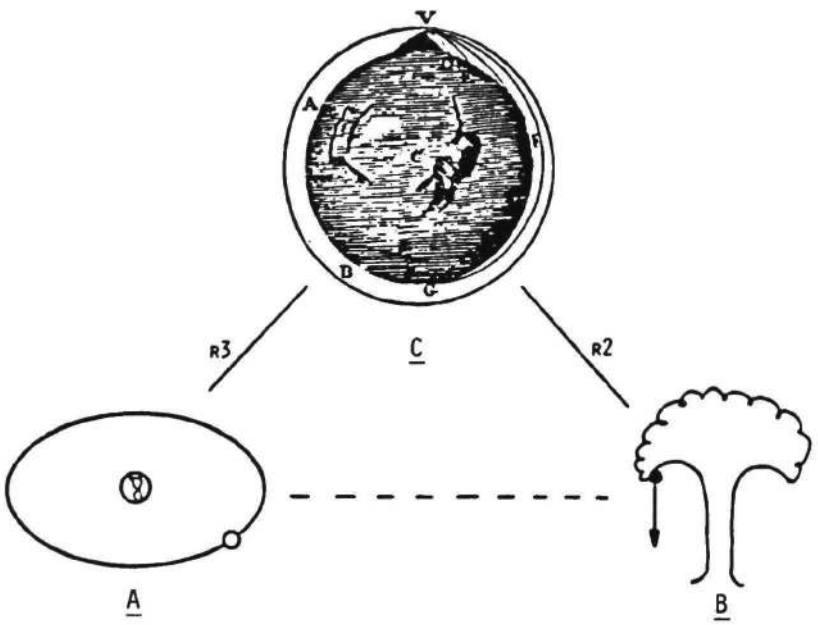


Figure 10

These bridging cases stand between the case of a cannonball dropped straight down and the case of the moon circulating in orbit. They help one see how the motion of a dropped object and the motion of the moon can have the same cause in the gravitational pull of the earth. Thus bridging cases are to be found in the history of science as well as in the protocols of expert problem solvers.

In fact the presence of a series of many bridging cases here suggests the possibility of smooth transition from the vertical drop to the orbiting object. In the case of multiple bridges we are approaching what can be considered a third method of analogy evaluation, namely, finding a conserving transformation. Such a transformation changes case A into case B while conserving important relationships that make A analogous to B. Thinking about continuously increasing the horizontal speed of launch in this case would constitute a conserving transformation since the major relationship of gravity causing the acceleration of the object remains unchanged.

APPLICATIONS TO SCIENCE TEACHING

Analogies and bridges may also have particular value in science teaching as an aid to helping students overcome persistent misconceptions. Recently, a fairly large literature has appeared on the problem of deep-seated preconceptions in students which can make the accepted scientific model seem counter-intuitive to them at a qualitative level (e.g., Helm and Novak, 1983; Clement, 1983). For example, in a survey of 112 high school chemistry

students who had not taken physics, we found that 75% of the students did not believe that a table pushes up on a book. The physicist believes that since all materials are deformable, the table will deform, acting like a spring, and provide an upward force just large enough to balance the weight of the books. The difficulty here is not just one of students lacking a particular fact. Pilot tutoring interviews and class discussions indicate that many students express disbelief in the physicist's view and have a deeply held belief that stationary objects are rigid barriers which cannot exert a force on their own. On the other hand, 96% of the students did believe that a spring pushes up when it is compressed with one's hand. The contrast between these results is interesting since the physicist views these two situations as essentially identical. diSessa (1983) refers to the concept of springiness as a "phenomenological primitive" and discusses the evolution of the individual's intuitions that is needed to become skilled in physics.

An interesting conjecture is that the right bridge may help a student believe in the validity of an analogy proposed in instruction. The following instructional strategy, suggested by our analysis of experts, attempts to build basic conceptual models that are grounded in intuitions the student already has.

Teaching Strategy:

(1) Draw out the conceptual difficulty in a concrete target situation A where the student makes a statement that is in conflict with accepted theory.
(2) Search for a simple analogous situation B where the student has a reliable intuition that is in agreement with physical theory and that is a relevant starting point for the area of difficulty.
(3) Stimulate discussion and encourage students to look for and match key features by asking them to describe how A and B are alike and different.
(4) Students often will still not believe that A is analogous to B. Finding an apt bridge that appeals to the students' intuitions is an important technique for remedying this.

The hand on the spring situation is a potential starting point for instruction since it draws out a correct intuition from students. For this reason we call it an "anchor". Figure 11 shows multiple bridging cases used to help convince students that the analogy between the "hand on the spring" anchor and the targeted "book on the table" case is valid. We have attempted to use this strategy in tutoring studies and class sessions with high school students and find that: 1) Students readily understand the anchoring case; 2) many students indeed do not initially believe that the anchor and the target cases are analogous; 3) the bridging cases sparked an unusual amount of argument and constructive thinking in class discussions; and 4) the bridging cases helped many students to believe in the analogy. We are in the process of collecting data to confirm these preliminary results. Minstrell (1982) has had some success with a related approach.

In this method then, we are trying to ground the student's understanding on a physical intuition about a familiar case. The strategy is to build on and extend the intuition by using analogical reasoning. The problem is that students will not be able to understand how other cases can possibly be

analogous to the familiar case. Presenting the right analogy is not enough -- the student must also come to believe in the validity of the analogy. The technique of bridging by using chains of analogies combined with discussion to encourage active thinking, appears to be helpful for this purpose.

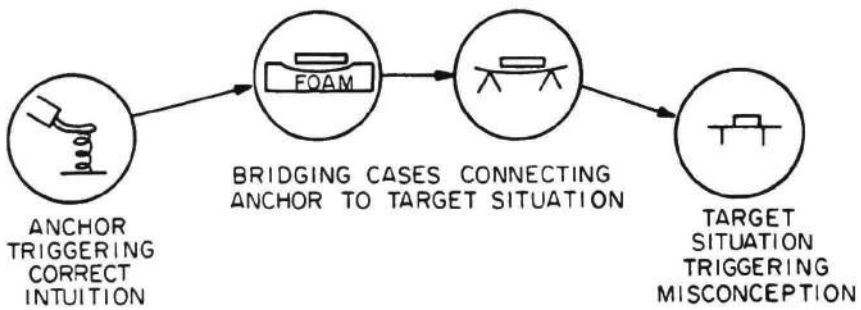


Figure 11

CONCLUSION

The ability to evaluate the validity of analogies appears to play as important a role in insightful problem solutions as the ability to generate analogies. Three methods for evaluating validity were discussed: matching key features, bridging, and finding a conserving transformation. These patterns of reasoning were observed to occur in different problem contexts in both science and mathematics. These empirical studies have allowed us to construct a more detailed theory of how analogies can be used effectively in instruction. A method for evaluating analogies used by expert scientists as well as by Newton and by Galileo and his predecessors appears to be a promising strategy for helping students overcome misconceptions in science.

Note 1. The subject may also use bridging recursively by bridging again between C and B or C and A as in the case of the square and hexagonal doughnuts.

*Preparation of this paper was supported by a grant from the National Science Foundation #MDR8470579.

References

Benedetti, G. (1969) Diversarum speculationum mathematicarum (1585), in S. Drake and I. Drabkin, Mechanics in sixteenth-century Italy (p. 206), Madison, Wisconsin: U. of Wisconsin Press.

Clement, J. (1981) "Analogy generation in scientific problem solving". Proceedings of the Third Annual Meeting of the Cognitive Science Society, Berkeley, California.

Clement, J. (1982). "Analogical reasoning patterns in expert problem solving". Proceedings of the Fourth Annual Meeting of the Cognitive Science Society (pp. 79-81). Ann Arbor, Michigan.

Clement, J. (1983). A conceptual model discussed by Galileo and used intuitively by physics students. In A. Stevens and D. Gentner (Eds.), Mental models (pp. 325-340), New Jersey: Lawrence Erlbaum Associates.

diSessa, A. (1983). Phenomenology and the evolution of intuition. In A. Stevens and D. Gentner (Eds.), Mental models (pp. 15-33), New Jersey: Lawrence Erlbaum Associates.

Galileo, G. (1954) Two New Sciences (1638). Translated by H. Crew and A. de Salvio (p. 62), New York: Dover.

Gentner, D., & Gentner, D. (1983). Flowing waters or teeming crowds: mental models of electricity. In A.L. Stevens and D. Gentner (Eds.), Mental models (pp. 99-101), New Jersey: Lawrence Erlbaum Associates.

Gick M. and Holyoak, K. (1980). Analogical problem solving. Cognitive Psychology, 12, 306 - 355.

Minstrell, J. (1982). Explaining the 'at rest' condition of an object. The Physics Teacher, 20, p. 10.

Rummelhart, D. and Norman, D. (1980). Analogical processes in learning. In J.R. Anderson (Ed.), Cognitive skills and their acquisition. Hillsdale, N.J.: Lawrence Erlbaum Associates.

USING FREQUENCY- AND ORIENTATION- TUNED CHANNELS TO DETERMINE SURFACE SLANT

PAUL KUBE

The Institute of Cognitive Studies
and
The Department of Electrical Engineering and Computer Sciences
University of California, Berkeley

INTRODUCTION

The thesis that spatial frequency analysis of the retinal image plays a fundamental role in the early stages of visual processing has received considerable evidential support recently, both from psychophysics (e.g., (Campbell & Robson 1968), (Sachs, Nachmias & Robson 1971), (DeValois & DeValois 1980), (Daugman 1984a)) and neurophysiology (e.g., (Campbell et al. 1969), (Maffei & Fiorentini 1973), (Schiller, Finlay & Volman 1976), (DeValois, Yund & Hepler 1982)). What role the spatial frequency information so extracted plays in later stages of processing remains an open question, though one that has increasingly received the attention of theoreticians and modellers (e.g., (Ginsburg 1978), (Watson 1983), (Janez 1983), (Daugman 1984b)). Here we present results which show how this information could be used in the determination of the orientation of environmental surfaces, and discuss the limits of such a model as an account of human perceptual capacities.

ANISOTROPY, SLANT, AND SPECTRUM

Suppose a planar environmental surface is oriented perpendicular to the line of sight and produces an irradiance pattern $I(x,y)$ on the retina, relative to some suitably chosen retinal coordinate frame. Rotating the environmental surface about a line parallel to the retinal y axis by an angle σ now foreshortens the retinal image: if the distance to the surface is large enough with respect to the size of the surface for orthographic projection to be a good approximation, the retinal irradiance pattern becomes $I'(x,y)=I(x,y/\cos\sigma)$.[1]

[1] Ignoring effects due to any change in angle between the surface normal and the illuminant vector.

If appropriate assumptions can be made about what the unforeshortened retinal image is like, the slant angle σ can be recovered from the foreshortened image I'. It has been suggested (e.g., by Witkin (Witkin 1981)) that if the retinal image is a foreshortened isotropic image, it should be so interpreted. That is, if $I'(x,y)$ is such that, for some σ, $I'(x,y\cos\sigma)$ is isotropic, then the imaged surface should be supposed to be slanted at angle σ.

Witkin counts an image isotropic if the distribution of tangent directions along contours in the image (or an edge-enhanced version of the image) are approximately uniform on $[0,\pi]$. Here, we suggest that an image be considered isotropic if its Fourier power spectrum is approximately circularly symmetric. This is an acceptable condition for isotropy for the following reason. Consider the definition of the two-dimensional Fourier transform of the retinal image $I(x,y)$:

$$f(\omega,\theta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-i2\pi(x\omega\cos\theta + y\omega\sin\theta)} I(x,y)\, dy\, dx$$

This expresses the spectrum f in polar coordinates of frequency ω and orientation θ. Fix $\theta=0$. Now

$$\begin{aligned} f(\omega,0) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-i2\pi x\omega}\, I(x,y)\, dy\, dx \\ &= \int_{-\infty}^{\infty} e^{-i2\pi x\omega} \int_{-\infty}^{\infty} I(x,y)\, dy \quad dx \end{aligned}$$

But this is just a *one*-dimensional Fourier transform of the function $I(x)$ such that

$$I(x) = \int_{-\infty}^{\infty} I(x,y)\, dy$$

$I(x)$ represents what $I(x,y)$ is like, 'on average', in the direction parallel to the x axis, and $f(\omega,0)$ just encodes this information as a function of frequency. Now since the orientation of the retinal coordinate frame was arbitrary, it's clear that a 'slice' through the two-dimensional spectrum $f(\omega,\theta)$ for any fixed θ represents what the image is like 'on average' in the direction making angle θ with the x axis; and thus if the spectrum is circularly symmetric, so will the image be. And a circularly symmetric image is intuitively isotropic.

But circular symmetry of images is too strong a necessary condition of isotropy. Requiring only that the Fourier *power* spectrum $F(\omega,\theta) = |f(\omega,\theta)|^2$ be circularly symmetric relaxes the condition in a satisfying way. Now isotropy is insensitive to translations of the retinal coordinate frame (since translation only affects the phase of the spectrum, not its magnitude) and to other perturbations of phase of components of the spectrum. Moreover, the power spectrum represents the foreshortening effects of slant in a systematic way. By the similarity theorem of Fourier analysis (cf. (Bracewell 1978) , p. 244), if $F(\omega,\theta)$ is the power spectrum of the image $I(x,y)$, then

$$F'(\omega,\theta) = \cos^2\sigma\, F\left((\omega^2\cos^2\theta+\omega^2\sin^2\theta\cos^2\sigma)^{1/2},\ \arctan\frac{\sin\theta\cos\sigma}{\cos\theta}\right) \qquad (1)$$

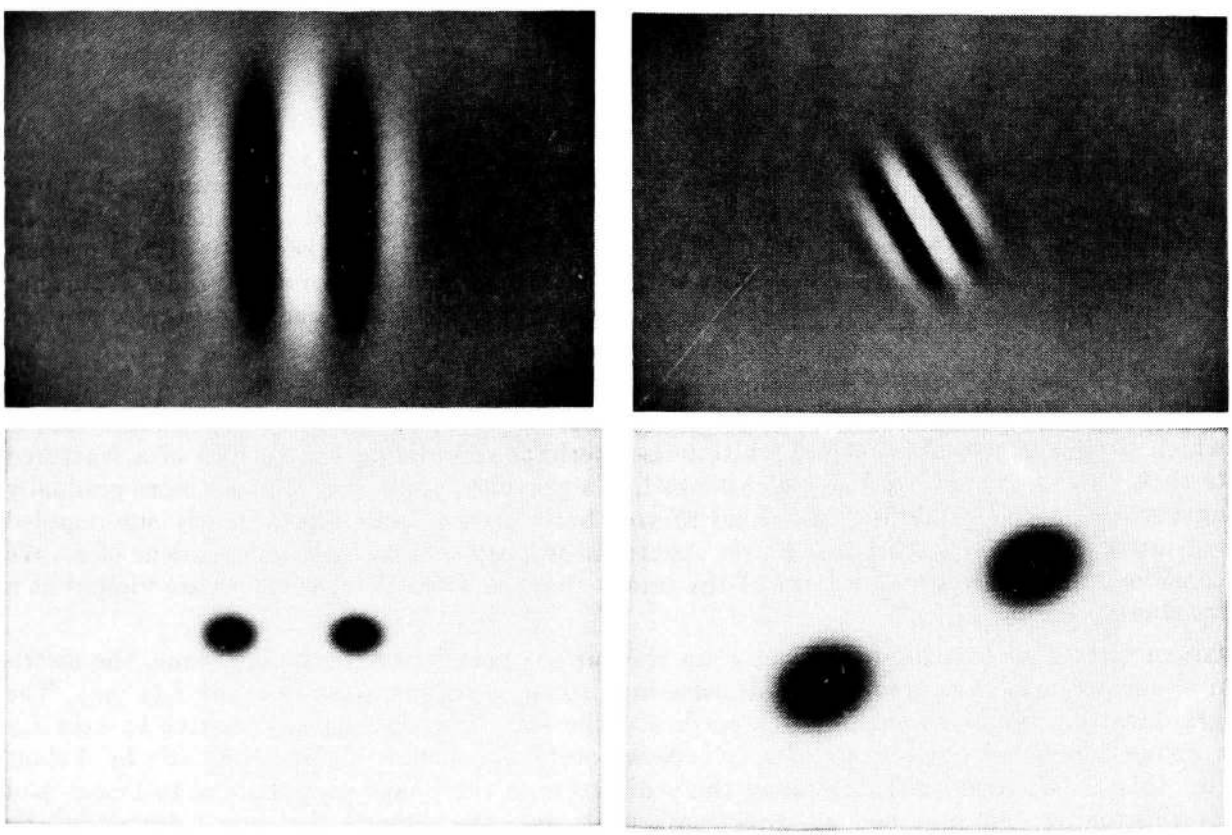
is the power spectrum of $I(x,y/\cos\sigma)$.

The power spectrum characterization of isotropy and Witkin's tangent distribution characterization agree for many images. We have adopted the spectrum characterization because it establishes a relationship between the presence or absence of isotropy in the retinal image and the activation of biologically plausible tuned channels in a way that permits the extraction of slant information from some kinds of images. This relationship will be developed in the remainder of the paper.

THE TUNED CHANNEL MODEL

To say that there are spatial frequency- and oriention-tuned channels in the visual system is just to say that there are neurons in visual cortex whose firing rates are elevated only when the power spectrum of the retinal image shows energy in certain regions of the Fourier plane. The power spectrum of the retinal image can be well represented by the activation level in many such channels.

The point spread functions ('receptive fields') and spectral response of two channels are shown in Figure 1. Channels vary in their nominal orientation and frequency but are assumed to be identical in orientation and relative frequency bandwidth at 30 degrees and one octave respectively. These bandwidth parameters are those used by Jañez (Janez 1983) and are close to the average observed in cortex (DeValois, Yund & Hepler 1982); Watson (Watson 1983) and Daugman (Daugman 1984a) propose slightly different parameters. In the present application, what's important is that the channels be fairly narrowly tuned in frequency and orientation, and that there be enough of them at enough different nominal frequencies and orientations to cover the Fourier plane over the frequency range of interest.



(a) (b)

Figure 1

Point spread (above) and frequency response (below) functions for two tuned channels: (a) frequency ω_0, orientation 0°; (b) frequency $2\omega_0$, orientation 30°. Each channel has half-amplitude bandwidth of one octave in frequency and 30 degrees in orientation.

The activation in a single tuned channel responding to a retinal image is calculated as follows. Let the power spectrum of the image I_i be described by the function of frequency and orientation $F_i(\omega,\theta)$, and that of the response of channel centered at frequency ω_0 and orientation θ_0 be $F_{\omega_0,\theta_0}(\omega,\theta)$; then we say that the activation produced by the image in the channel at f_0,θ_0 is

$$P_{\omega_0,\theta_0}(I_i) = \frac{\int_{\infty 0}^{\infty 2\pi} \int F_i(\omega,\theta)\, F_{\omega_0,\theta_0}(\omega,\theta)\, d\theta\, d\omega}{\int_{\infty 0}^{\infty 2\pi} \int F_{\omega_0,\theta_0}(\omega,\theta)\, d\theta\, d\omega} \qquad (2)$$

That is, channel power is the sum of image frequency power weighted by the response of the channel, and normalized by the total frequency-plane area of the channel.

DETERMINING THE SLANT OF SCALING NOISE SURFACES

The distribution of markings on surfaces in the natural environment is typically random in some sense; highly regular, structured patterns are rare. Here we consider the class of surfaces marked with *isotropic scaling noise*, a class of markings that seem to model some naturally occuring textures.[2] We define such surfaces as those which have spatial radiance functions whose power spectra are of the form

$$F_\beta(\omega,\theta) = k\, \omega^{-\beta}$$

with $\beta \geq 0$, and with random phase. Here β is a parameter that determines how rapidly changing the radiance across the surface is; as β increases, the power at high spatial frequencies on the surface decreases, and so the surface has a more slowly varying pattern marked on it (k is a proportionality constant that can be taken as unity if the markings are suitably normalized). Note that the spectrum is independent of orientation θ, so we should expect such a surface to look statistically the same in every direction (thus *isotropic* scaling noise).

Images of isotropic scaling noise surfaces with varying β are shown in Figure 2(a-d) viewed with zero slant, i.e., 'straight-on', with line of sight normal to the surface. An extreme case is $\beta=0$ which produces two-dimensional white noise, perhaps resembling the surface of a fractured granite rock. $\beta=1.0$ gives a surface which looks like a gravelled walk; $\beta=2.0$ has a more gradually varying texture, a fair rendering of a kind of tree bark; $\beta=3.0$ looks like a gently sun-dappled patch of lawn. As noted, each has a power spectrum proportional to $1/\omega^\beta$, independent of θ. We next consider how the power spectrum of the image changes when these surfaces are viewed at a *nonzero* slant.

Given Cartesian coordinate axes x',y' in the surface parallel to the image plane, the distribution of surface markings is some two-dimensional isotropic scaling noise function $I_\beta(x',y')$. The orthographically projected image of this surface is the same function $I_\beta(x,y)$ relative to axes x,y in the image which are projections of x',y'. Now rotate the surface about its x' axis by a slant angle σ: this causes foreshortening along the y direction in the image proportional to $1/\cos\sigma$, and no foreshortening along the x direction; that is, the image is now described by $I_{\beta,\sigma}(x,y) = I_\beta(x,y/\cos\sigma)$, which is no longer an *isotropic* scaling noise. (Examples of images of

[2] This class is a superset of the surfaces marked with isotropic fractional Brownian plane-to-intensity functions ((Mandelbrot 1983) , ch. 25-27), which restrict β to the range 2 to 4. The imaging of unmarked nonplanar surfaces whose shape is described by an isotropic fractional Brownian plane-to-*elevation* function is another matter, and it doesn't seem to have been yet adequately treated (Pentland's fractal Brownian surfaces (Pentland 1984) are not fractal in elevation).

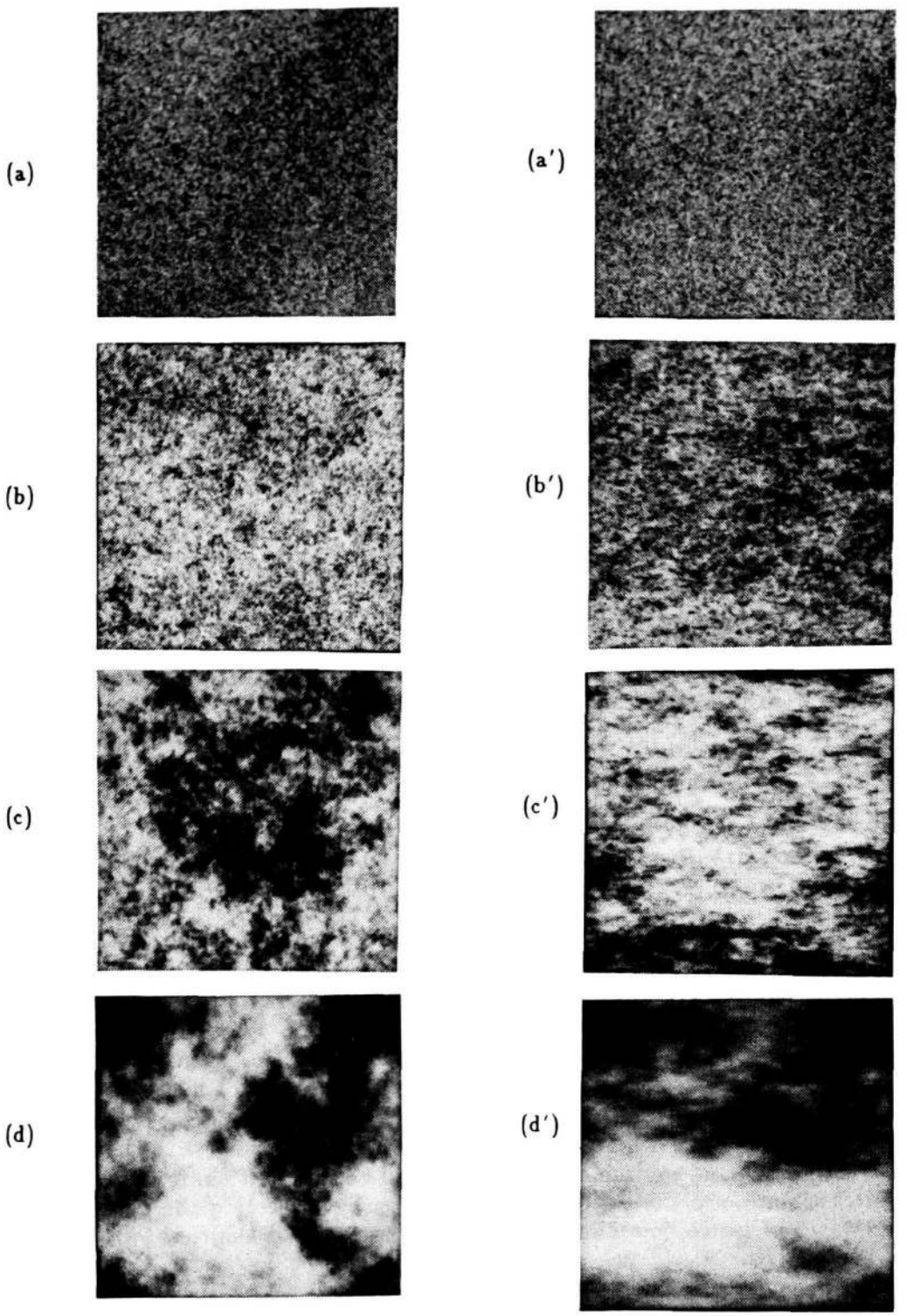


Figure 2

(a)-(d): Images of scaling noise surfaces, $\beta = 0, 1, 2$, and 3 respectively; slant $\sigma = 0$.
(a')-(d'): Images of scaling noise surfaces, $\beta = 0, 1, 2$, and 3 respectively; slant $\sigma = 70$ degrees.

scaling noise surfaces slanted at 70 degrees are shown in Figure 2(a'-d'); their anisotropy should be evident at least when β is large.) By the similarity theorem (cf. (1) above) the power spectrum of the image $I_{\beta,\sigma}$, which in the unslanted case $\sigma=0$ was

$$F_\beta(\omega,\theta) = k\ \omega^{-\beta},$$

becomes, for $\theta=\pi/2$ (i.e., parallel to the foreshortened y-axis):

$$F_{\beta,\sigma}(\omega,\pi/2) = \cos^2\sigma\ k\ (\omega\cos\sigma)^{-\beta},$$

and, for $\theta=0$ (i.e., normal to the foreshortened direction):

$$F_{\beta,\sigma}(\omega,0) = \cos^2\sigma\ k\ \omega^{-\beta}.$$

The ratio of the power in these two orientations, as a function of spatial frequency, is then[3]

$$R_{\beta,\sigma}(\omega) = \frac{F_{\beta,\sigma}(\omega,\pi/2)}{F_{\beta,\sigma}(\omega,0)} = (\cos\sigma)^{-\beta}, \tag{3}$$

and we can write

$$\sigma = \arccos\exp\left(\frac{\log R_{\beta,\sigma}(\omega)}{-\beta}\right). \tag{4}$$

Thus, we can determine the slant of the imaged surface if we can measure both β and $R_{\beta,\sigma}$. But this is easy to do, given the availability of frequency and orientation tuned channels of the sort introduced above. Say we have four channels with orientations at 0 and $\pi/2$ centered at each of two any distinct frequencies ω and $\alpha\omega$. Then, since α is known, a good estimate of β can be obtained from either the fact that

$$\frac{P_{\omega,0}(I_{\beta,\sigma})}{P_{\alpha\omega,0}(I_{\beta,\sigma})} \approx \frac{F_{\beta,\sigma}(\omega,0)}{F_{\beta,\sigma}(\alpha\omega,0)} = \frac{\cos^2\sigma\ k\ \omega^{-\beta}}{\cos^2\sigma\ k\ (\alpha\omega)^{-\beta}} = \alpha^\beta$$

or

$$\frac{P_{\omega,\pi/2}(I_{\beta,\sigma})}{P_{\alpha\omega,\pi/2}(I_{\beta,\sigma})} \approx \frac{F_{\beta,\sigma}(\omega,\pi/2)}{F_{\beta,\sigma}(\alpha\omega,\pi/2)} = \frac{\cos^2\sigma\ k\ (\cos^2\sigma\omega)^{-\beta}}{\cos^2\sigma\ k\ (\alpha\cos^2\sigma\omega)^{-\beta}} = \alpha^\beta\ ;$$

and then, with β known, and $R_{\beta,\sigma}$ estimated directly from

$$\frac{P_{\omega,\pi/2}(I_{\beta,\sigma})}{P_{\omega,0}(I_{\beta,\sigma})} \quad \text{or} \quad \frac{P_{\alpha\omega,\pi/2}(I_{\beta,\sigma})}{P_{\alpha\omega,0}(I_{\beta,\sigma})},$$

the slant σ can be determined by (4).[4]

Since the channels in the model have appreciable orientation bandwidth, they respond to image power in a range of orientations. This means that $P_{\omega,\pi/2}(I_{\beta,\sigma})$ will tend to be somewhat less than $F_{\beta,\sigma}(\omega,\pi/2)$, and $P_{\omega,0}(I_{\beta,\sigma})$ will tend to be somewhat larger than $F_{\beta,\sigma}(\omega,0)$, which in turn means that the estimate of $R_{\beta,\sigma}$ obtained in this way will be smaller than it should be, and so will lead to an underestimate of slant σ. But as can be seen by inspecting the results of simulating the model on images of isotropic scaling noise surfaces at various slants shown in Figure 3, the errors are not substantial.[5]

---

[3] The fact that $R_{\beta,\sigma}$ depends not on ω is the reason for calling these marking functions *scaling* noises.

[4] In practice, $P_{\omega,\theta}$ can be observed at a large assortment of frequencies, and the resulting estimates of β and σ can be subjected to averaging or a more sophisticated type of evidence composition.

[5] The case of the white noise surface is excluded; $\beta=0$ leaves (4) undefined. For such a surface, slanting does not break the isotropy of its image's power spectrum, and only changes its contrast: it is theoretically impossible to determine the slant of such a surface from a normalized image. Perhaps this is what Pentland ((Pentland 1984), p. 673) has in mind when he says "Characterization of an image in terms of radial slices of the Fourier domain ... constrains the shape of the 3-D surface hardly at all ... Illumination effects can account for most variation in

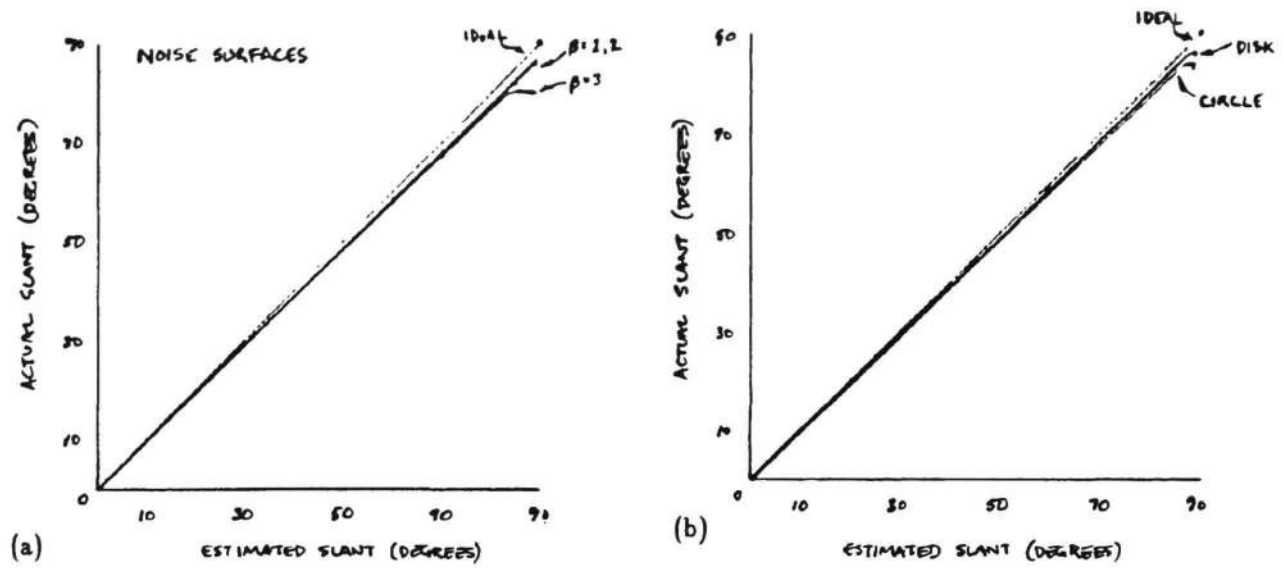


Figure 3

Estimates of slant generated by the 'adaptive β' model for (a) scaling noise surfaces, and (b) circles and disks.

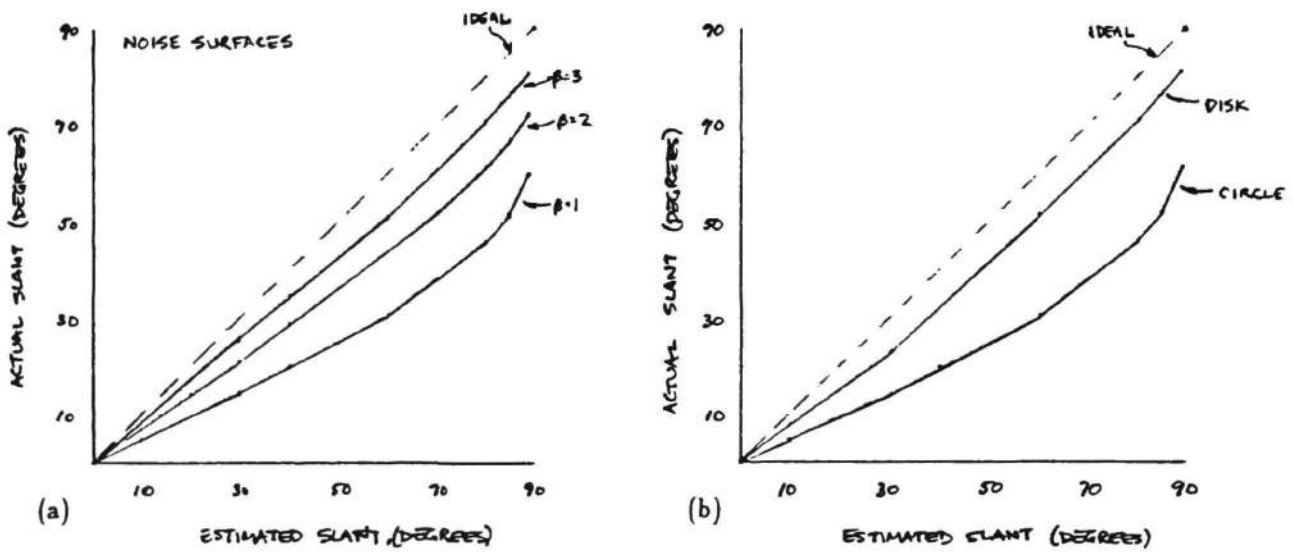


Figure 4

Estimates of slant generated by 'nonadaptive β' model based on Equation 5, $\beta_0 = 4$; for (a) scaling noise surfaces, and (b) circles and disks.

DETERMINING THE SLANT OF CIRCLES AND DISKS

The results of the previous section are not applicable only to stochastically textured surfaces; any surface marked with a function whose power spectrum is approximately proportional to $\omega^{-\beta}$ can have its slant estimated in the same way from a power spectrum of its image. In this section, we demonstrate how this works for isolated circles and disks.

Let a planar environmental surface be marked with a circle of radius r, ${}_C I_r(x',y') = \delta((x'^2+y'^2)^{\frac{1}{2}}-r)$, where δ is the unit impulse function. The orthographically projected image of this surface at zero slant is ${}_C I_r(x,y)$ which has power spectrum

$$F_r(\omega,\theta) = (2\pi r J_0(2\pi\omega r))^2 \quad .$$

Here J_0 is the Bessel function of the first kind of order zero, which can be closely approximated when $x > \pi$ by (cf. (Kreyszig 1972), p. 129)

$$J_0(x) \approx \frac{1}{\sqrt{\pi x}}\sqrt{2}\,\sin(x+\pi/4) \quad ;$$

so, if ω is not too small, we have

$$F_r(\omega,\theta) \approx \frac{4r}{\omega}\sin^2(2\pi\omega r+\frac{\pi}{4}).$$

Thus, an unslanted circle's power spectrum has an envelope like $k\omega^{-\beta}$ with $\beta=1$; and though the phase of the spectrum is not random as it is for scaling noise surfaces, the same tuned-channel techniques will work for slanted circles. The projection of a circle slanted at angle σ is the ellipse ${}_C I_r(x,y/\cos\sigma)$, which has power spectrum in the relevant directions $\theta=0,\pi/2$

$$F_{r,\sigma}(\omega,\pi/2) = \cos^2\sigma \quad \frac{4r}{\omega\cos\sigma} \quad \sin^2(2\pi\omega r\cos\sigma+\frac{\pi}{4})$$

$$F_{r,\sigma}(\omega,0) = \cos^2\sigma \quad \frac{4r}{\omega} \quad \sin^2(2\pi\omega r+\frac{\pi}{4}).$$

Now the ratio $F_{r,\sigma}(\omega,\pi/2)/F_{r,\sigma}(\omega,0)$ will be undefined for some ω since the denominator will vanish, but the averaging effect of the tuned channels (due to their having appreciable frequency bandwidth) will prevent the observation of zeros in $F_{r,\sigma}$ and will extract the envelope of the spectrum well enough to determine $\beta=1$ and then σ in the same way as for scaling noise surfaces.

A disk

$${}_D I_r(x,y) = \begin{cases} 1 & \text{when } x^2+y^2<r, \\ 0 & \text{otherwise} \end{cases}$$

has a $\omega^{-\beta}$ power spectrum envelope as does a circle of the same radius ${}_C I_r(x,y)$, but the scaling exponent β is, somewhat surprisingly, very different in the two cases. The disk has power spectrum

$$F_r(\omega,\theta) = (\frac{r}{\omega}J_1(2\pi\omega r))^2 \quad ,$$

where J_1 is the Bessel function of the first kind of order one, approximated when $x > \pi$ by

$$J_1(x) \approx \frac{1}{\sqrt{\pi x}}\sqrt{2}\,\sin(x-\pi/4) \quad .$$

Thus the power spectrum can be approximated by

$$F_r(\omega,\theta) \approx \frac{r}{\pi^2\omega^3}\sin^2(2\pi\omega r-\frac{\pi}{4})$$

---

such a description." But this is *not* true for values of β other than 0.

which has an envelope like $k\omega^{\beta}$ with $\beta=3$, compared to the circle's $\beta=1$. Again, the same tuned channel techniques permit an accurate estimation of the disk's slant. The results of a simulation of the model on these figures are displayed in Figure 3.

DISCUSSION OF THE MODEL

How suitable is this slant detection model as a model of human visual perception? One striking feature is its unrealistically accurate performance on images of scaling noise surfaces with small β: the model performs virtually flawlessly (cf. Figure 3) over nearly the entire range of slants when $\beta = 1$, but such images seem to produce very weak slant illusion for human observers and, even at a slant of 70 degrees do not have very salient anisotropies. As β increases, however, the salience of anisotropy seems to increase until, at $\beta=3$, Figure 2(d') looks fairly convincingly like a *slanted* patch of sun-dappled lawn.

This suggests that our model's 'adaptive' determination of the scaling factor β tailored for each image is unrealistic. Perhaps instead of (3), the visual system has

$$\frac{F_{\beta,\sigma}(\omega,\pi/2)}{F_{\beta,\sigma}(\omega,0)} = (\cos\sigma)^{\beta_0} \tag{5}$$

or

$$\frac{F_{\beta,\sigma}(\omega,\pi/2)}{F_{\beta,\sigma}(\omega,0)} = \frac{\beta_0}{\cos\sigma}$$

or

$$\frac{\pi}{2}\left(1-\frac{1}{1+\beta_0\,(F_{\beta,\sigma}(\omega,\pi/2)-F_{\beta,\sigma}(\omega,0))}\right) = \sigma$$

or some other monotonic mapping between a measure of anisotropy in the frequency plane and the possible range of slant angles with *fixed* parameter β_0. Given data from the psychophysical experiment, it would be possible to fit one or another such function; as an example, the results of a simulation of (5) with $\beta_0=4$ are shown in Figure 4, and they are closer than the adaptive model to what one might expect from human performance.

A problem with any such modification, however, is that to distinguish between the slant-detectability of scaling noise surfaces with $\beta=1$ and $\beta=3$ to fit human performance is *ipso facto* to distinguish to the same extent between the slant-detectability of circles and disks.[6] And while the experiments have not been done, it seems likely that circles and disks are very similar in respect of slant-detectability, and $\beta=1$ and $\beta=3$ noise surfaces are very different.

I suggest that the conclusion to draw is that the power in areas of an image's global frequency spectrum does not exhaust the information available to the perceptual system; spatial information also exists. In fact, as has been noted elsewhere,[7] if spatial frequency analysis is performed on the retinal image, it is performed locally, not globally; and the orientation- and

[6] This is true if the lowest spatial frequencies are ignored. At frequencies low enough to make the Bessel function approximations used here innacurate, or low enough to make the undulations in the circle and disk spectra observable by octave-bandwidth channels, a spatial frequency model could distinguish between noise on the one hand and circles and disks on the other. But low spatial frequencies are not the issue; low pass filtering the surfaces to remove these frequencies appears to change the slant-detectability of a large circle hardly at all, while practically eliminating the already scarcely detectable anisotropies in an image of a slanted $\beta=1$ noise surface. Anisotropies in images of a $\beta=3$ surface remain salient.

[7] For example, in (Daugman 1984a).

frequency-tuned channels that exist are almost optimally designed, from an information-theoretic point of view, to carry both spatial and frequency information. Now a scaling noise surface has a very uninteresting spatial organization; practically all its distinctiveness comes from the statistics of its radiance distribution, which is captured well in its power spectrum. Circles and disks and their projections, however, have a determinate shape, and it should play a role in the judging of slant. A parameterized tuned channel model of the sort sketched here may tell the story of slant detection for stochastic surfaces, but it will not be the whole story of slant detection.

Acknowledgements

The research reported here was supported by a grant from the Alfred P. Sloan Foundation to the Institute of Cognitive Studies at Berkeley. I want to thank Steve Palmer, Jitendra Malik, Gene Switkes and John Kruschke for helpful discussions.

REFERENCES

Bracewell, R. N. (1978). *The Fourier Transform and Its Applications*, McGraw Hill, 1978.

Campbell, F. W. & J. G. Robson (1968). Application of Fourier analysis to the visibility of gratings, *J. Physiol. 197* (1968), 551-566.

Campbell, F. W., G. F. Cooper, J. G. Robson & M. B. Sachs (1969). The spatial selectivity of visual cells of cat and the squirrel monkey, *J. Phsiol. 204* (1969), 120-121.

Daugman, J. G. (1984). Spatial Visual Channels in the Fourier Plane, *Vision Res. 24*, 9 (1984), 891-910.

Daugman, J. G. (1984). Representational Issues and Local Filter Models of Two-dimensional Spatial Visual Encoding, in *Models of the Visual Cortex*, D. Rose & V. G. Dobson (editor), Wiley, 1984.

DeValois, R. L. & K. K. DeValois (1980). Spatial Vision, *Ann. Rev. Psychol. 31* (1980), 309-341.

DeValois, R. L., E. W. Yund & N. Hepler (1982). The Orientation and Direction Selectivity of Cells in Macaque Visual Cortex, *Vision Res. 22* (1982), 531-544.

Ginsburg, A. P. (1978). *Visual information processing based on spatial filters constrained by biological data*, Cambridge University, 1978. Ph.D. dissertation.

Janez, L. (1983). Stimulus control of visual reference frame orientation: Quantitative theory, *Informes de Psicologia*, 1983, 133-147.

Kreyszig, E. (1972). *Advanced Engineering Mathematics*, John Wiley and Sons, 1972.

Maffei, L. & A. Fiorentini (1973). The visual cortex as a spatial frequency analyser., *Vision Res. 13* (1973), 1255-1267.

Mandelbrot, B. B. (1983). *The Fractal Geometry of Nature*, W. H. Freeman, New York, 1983.

Pentland, A. P. (1984). Fractal-based description of Natural Scenes, *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-6*, 6 (1984), 661-674.

Sachs, M. B., J. Nachmias & J. G. Robson (1971). Spatial frequency channels in human vision, *J. Opt. Soc. Am. 61* (1971), 1176-1186.

Schiller, P. H., B. L. Finlay & S. F. Volman (1976). Quantitative studies of single-cell properties of monhkey striate cortex , *J. Neurophysiol. 39* (1976), 1288-1351.

Watson, A. B. (1983). Detection and Recognition of Simple Spatial Forms, in *Physical and Biological Processing of Images*, O. J. Braddick & A. C. Slade (editor), Springer Verlag, Berlin, 1983.

Witkin, A. P. (1981). Recovering Surface Shape and Orientation from Texture, *Artificial Intelligence 17* (1981), 17-45.

ADAPTATION, BRIGHTNESS PERCEPTION AND THEIR CORRELATION WITH PHOTORECEPTOR RESPONSES

K.N. Leibovic: Department of Biophysics,

State University of New York at Buffalo 14214

ABSTRACT

In psychophysics, the effects of light backgrounds and photopigment bleaching can be equated in a variety of situations involving sensitivity, flicker, and the subjective perception of brightness. We have investigated such possible equivalences at the single unit level through intracellular recordings from vertebrate rods. At a given level of background or bleaching adaptation, vertebrate rods respond with characteristic waveforms to brief flashes of light of increasing intensity. The peak responses can be plotted vs. flash intensity to give a response curve in each state of adaptation. Two consequences of light adaptation are a shift to higher light intensities and a compression of the response curve. We have been able to establish equivalences in single units between bleaching and backgrounds in terms of threshold elevation and response compression. Moreover, we have found that such equivalent (background, bleaching) pairs have similar response curves. But these response curves are direct measures of stimulus flash intensity. Based on the close parallels between the psychophysical and the neurophysiological equivalences of bleaching and backgrounds, we therefore propose that brightness perception is mediated by the peak responses of photoreceptors. It is not often that behavioral observations can be traced back to the neuronal machinery and the latter suggest psychophysical tests which can further clarify brightness equivalence relations. This approach is discussed in the light of our results.

Introduction:

A major aspiration of neuroscience is to relate psychophysical and behavioral observations to the underlying neuronal machinery. Correlations with complex functions, such as memory, are difficult to come by. We are better off with adaptation and brightness perception which are considered in this paper.

Light adaptation is a fascinating phenomenon. We can see the light from a faint star in the night sky and we can see

objects some 10^{12} times as bright in full sunlight. When the ambient light level changes it takes time to adapt to it and the bigger the change, the longer it takes. At any given level of background light our visual system only operates over some three orders of magnitude of light intensity. This is the operating range of the photoreceptors, as well as of the neurons at the various stages of the visual pathway. The neuronal input-output ranges can be modulated somewhat through biochemical changes and through the interplay of excitation and inhibition. But they cannot be shifted over the enormous intensity range over which light adaptation operates. Moreover, as a general principle, there cannot be responses at any one stage of a serial information system to responses outside the range of a preceding stage. We must therefore conclude, as the signal is transmitted from photoreceptors to ganglion cells, LGN and cortex, that the capability for shifting the operating range over the adaptive range must reside in the photoreceptors (Leibovic, 1971, Dowling & Ripps, 1972).

Hecht (1937) had proposed a chemical basis for visual transduction and adaptation. In agreement with his ideas, the state of photopigment bleaching does affect adaptation, but not in the way he had envisaged. Nor is adaptation determined solely by the bleached pigment. It also depends on a biochemical cycle mediating transduction and affecting membrane conductance in a series of steps involving Ca^{++} and cGMP which have yet to be fully elucidated.

Crawford (1947) showed how background light and pigment bleaching both affect the state of adaptation, and he introduced

the notion of an equivalence between backgrounds and bleaching in terms of threshold elevation. This equivalence has been amply confirmed in a variety of different experimental situations (Rushton 1962). In particular, it has been extended to the perception of brightness: Barlow and Sparrock (1964) have shown that the apparent brightness of a bleached retinal area is the same as that of the background which elevates the visual threshold by the same amount.

If light adaptation is indeed mediated by photoreceptors, how do their responses correlate with the psychophysical findings and what measure of these responses can be assigned to brightness perception? We have addressed these questions by recording the intracellular responses of the rods in the isolated retina of Bufo marinus. In particular we have investigated the effects of backgrounds and bleaching on thresholds and response amplitude. The stimulus was a 100 msec flash. A background of a given intensity could be projected onto the retina, and bleaching of the photopigment could be accomplished by exposing the retina to a timed, intense background light. The materials and methods have been described fully elsewhere (Leibovic, Dowling, Kim: in preparation).

Clearly there are species differences and we know that, for example, the rods of rats adapt differently from those of frogs. Nevertheless, there are similarities among all vertebrate rods and, as we shall see, the results on Bufo bear quite a resemblance to psychophysical data.

Bufo Rod Responses:

A typical series of intracellularly recorded rod responses to flashes of increasing intensity is shown in Figure 1. At high

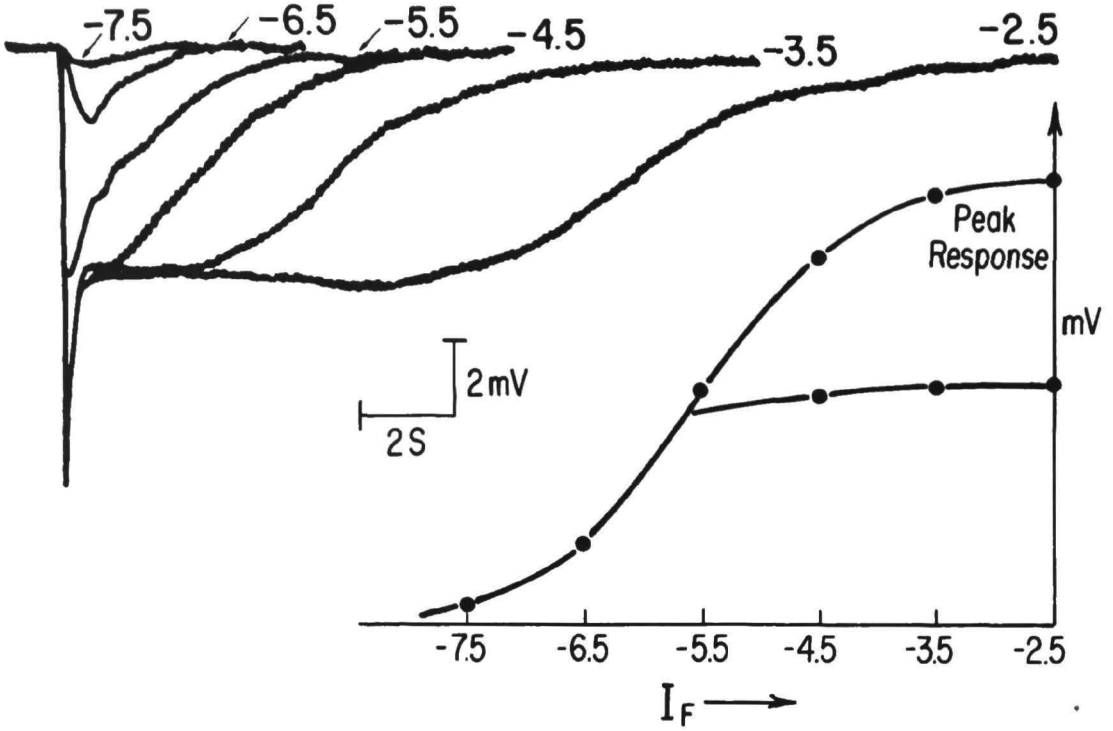


Figure 1: Responses of a dark adapted rod to 100msec flashes of increasing intensity.
Inset: The peak responses (mV) are plotted vs. relative flash intensity (I_F = neutral density of flash in log units). At high intensities the peak and plateau of the response give rise to two branches of the curve. This is the "operating curve".

intensities the response waveform develops a spike and a plateau before relaxing back to baseline. The response amplitude cannot increase indefinitely, but saturates at some level. The flash intensity producing the smallest detectable response (in practice, a criterion response of 1/2 mV or 1 mV) is the threshold intensity.

Both threshold and saturated response amplitude depend on the state of adaptation. Two consequences of either bleaching or background adaptation are a shift of the threshold to higher

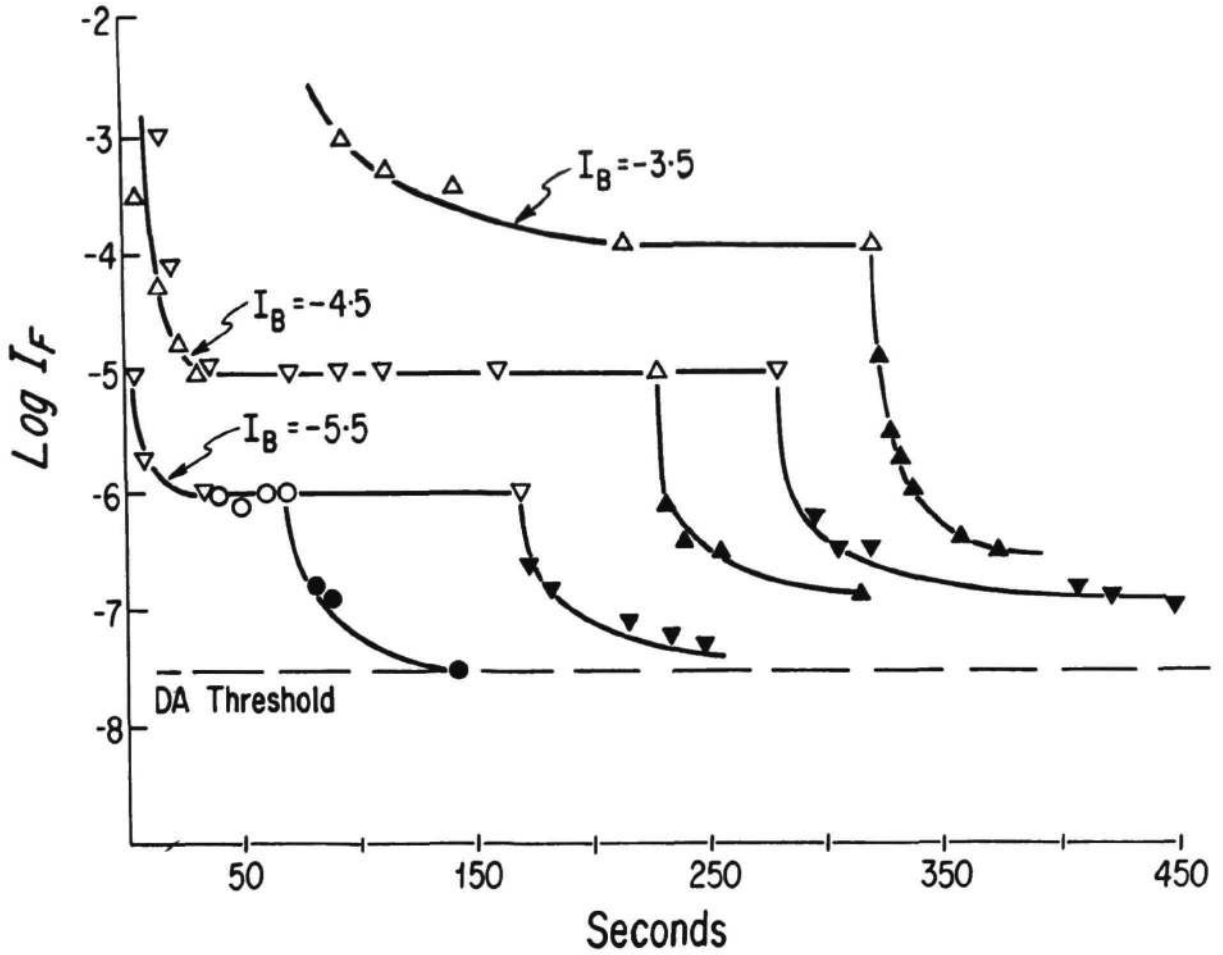


Figure 2: Threshold Recovery when a Background is turned ON and OFF:

A typical set of threshold recovery curves is shown, the lowest containing two experiments at a nominal background N.D. - 5.5, the middle ones two experiments at N.D. -4.5, and the upper curve one experiment at N.D. -3.5. The light was turned on at time 0 and, after a steady threshold was reached, the light was turned off. In the two experiments at N.D. -5.5 the light was turned off after different exposure times and similarly at N.D. - 4.5. There was no measurable threshold recovery at backgrounds of N.D. -2 or brighter. Log I_f on the ordinate is the relative intensity in N.D. units of a 100 mS flash which elicits a 1/2 mV response.

The original, dark adapted threshold of the cells was Log I_F = -7.5. The estimated fraction of pigment bleached at the termination of light exposure was less than 0.02 up to I_B = -4.5 and about .07 at I_B - -3.5, where I_B is the N.D. of the background.

NOTE: N.D. is short for the logorithm of neutral density attenuation.

light intensities and a reduction of the saturated amplitude. In fact, all response amplitudes are reduced. This can be displayed in the "operating curve", a graph of peak amplitude versus flash intensity for stimuli from threshold to saturation at each level of adaptation, illustrated in the inset to Figure 1.

Effects of Backgrounds on Thresholds:

When a background is projected onto the retina the intracellular threshold response falls rapidly from an initially high level to a lower value. When the background is turned off the threshold again falls rapidly to the dark adapted value, unless some measurable fraction of the pigment has been bleached, in which case the final threshold is elevated, depending on the amount bleached. This is quite analogous to the psychophysical observations. Our data are shown on Figure 2 for three different backgrounds. The complete range of threshold elevations is about five orders of magnitude (5 log units) of light intensity. This is comparable to the range of some 6 orders of magnitude for the scotopic range of human vision.

The steady state threshold elevation of Bufo rods in the presence of a background is shown in Figure 3. Again there is a close parallel with psychophysical results. At low background intensities the threshold is close to the dark adapted value. As the intensity rises the threshold becomes proportional to the square root of the background. This is precisely what one would expect for the detection of a signal in the presence of noise (see e.g. Leibovic 1972 p. 114). At still higher intensities, threshold is proportional to background (Weber-Fechner rule in psychophysics).

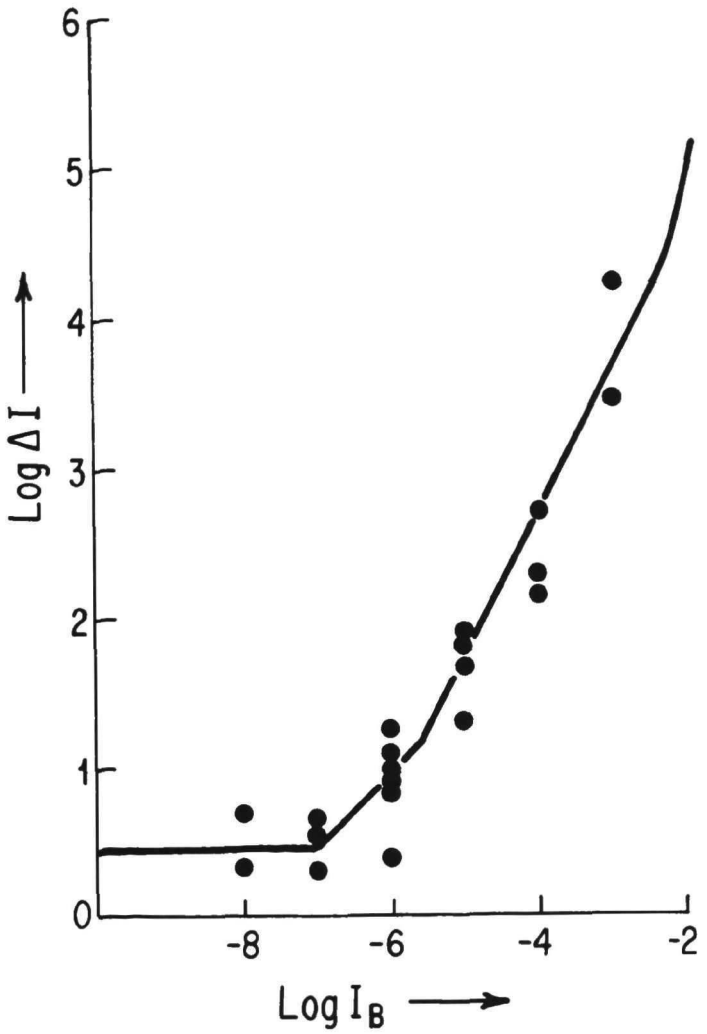


Figure 3: The ordinate is the threshold elevation above the dark adapted threshold in units of N.D. The abscissa is the relative background intensity in units of N.D.

We can see that there are close similarities between psychophysics and these single photoreceptor responses with regard to threshold elevation due to backgrounds.

Effects of Bleaching on Thresholds:

There is essentially no pigment regeneration in the isolated retina. Therefore it is an ideal preparation for studying bleaching. Our data for threshold elevation as a function of bleaching are shown in Figure 4. Up to almost 20% bleaching,

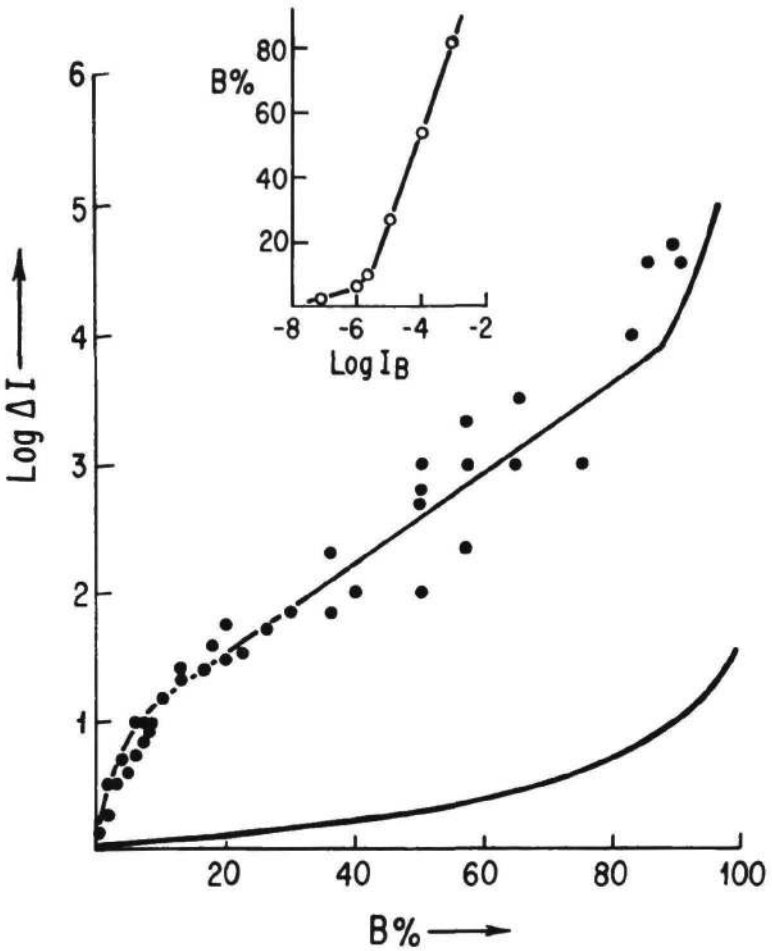


Figure 4: The ordinate is the same as in Figure 3. The abscissa is the bleached percentage. The lower curve is the threshold elevation due to the reduced photon absorptions in the presence of bleached pigment. In fact, the threshold elevation is much greater, showing that it is not simply due to the presence of bleached pigment.
Inset: Threshold equivalent I_B and B. For each point on the curve, I_B and B have the same threshold.

the threshold rises linearly and thereafter it rises exponentially. The linear region has not been reported previously, perhaps due to the difficulties of accurate measurement and the scatter of the data points. If there were fewer data points in that region on our graph, one might have been tempted to simply draw a straight line through the points, which would have implied an exponential relationship over the whole range. The value of the exponent depends on the range of thresholds covered. Thus, Rushton (1962) used an exponent of 20 (see his Figure 3) on a graph that has a maximum threshold elevation of log 20. But this is an extrapolation. The actual

thresholds he measured only went up to about log 7. Thus, the differences between our data and psychophysics may be more apparent than real. In any event, the qualitative similarities are close.

Effects of Background and Bleaching on Saturated Amplitude:

When we measure the saturated amplitudes, firstly in the dark adapted cell, secondly in the presence of a known background and thirdly after bleaching a given fraction of pigment, we can find the reduction of the saturated response in the last two conditions. Our data are shown in Figures 5 and 6. The points on these Figures are mean values obtained from the raw data. These Figures establish a relation between bleaching and backgrounds at the upper end of the operating curve, just as Figures 3 and 4 establish a relation at the lower end.

Is there any connection between these relations?

Equivalences between Backgrounds and Bleaching:

We can establish an equivalence between backgrounds and bleaching with respect to either thresholds or amplitues: Given a threshold elevation ΔI_T we can find a background I_B and a bleached fraction B which produce this threshold elevation. We write this equivalence as:

$$I_B \overset{\Delta I}{\sim} B \qquad (1)$$

We can similarly establish an equivalence

$$I_B \overset{A}{\sim} B \qquad (2)$$

where A is the ratio of the saturated amplitude in the presence

of the background I_B or due to the bleached fraction B to the saturated amplitude of the dark adapted cell.

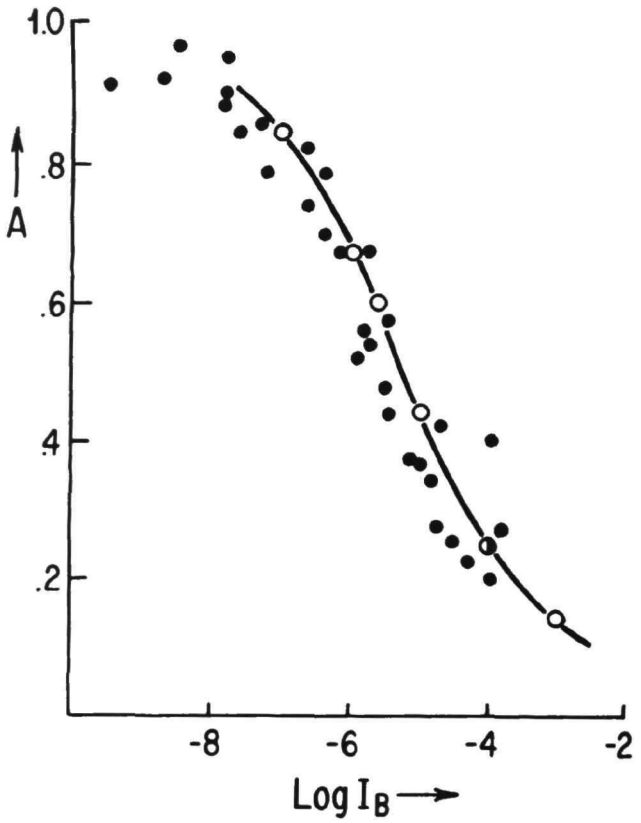


Figure 5: Amplitude reduction vs. background. See text for details.

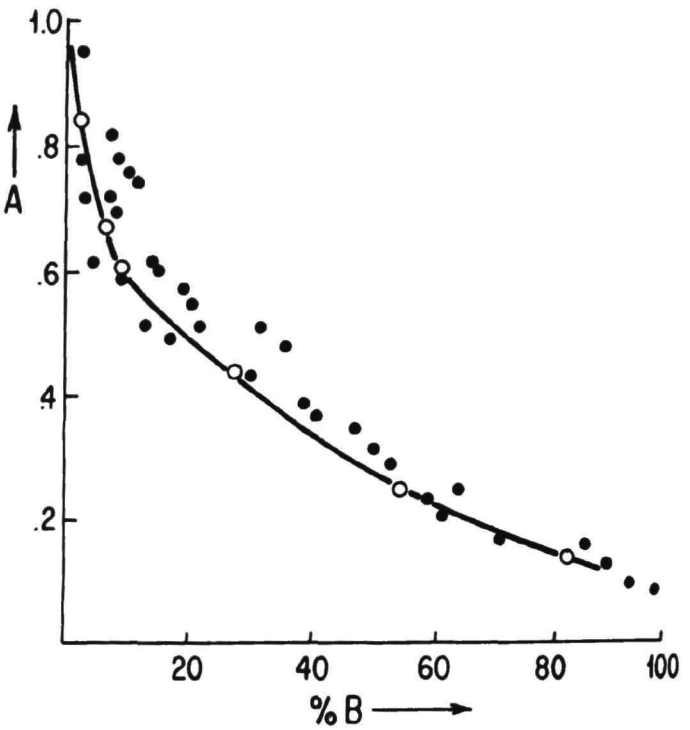


Figure 6: Amplitude reduction vs. bleaching. See text for details.

In the simplest case the two equivalences are the same. In other words, if I_{B1} and B_1 produce the same threshold elevation ΔI_1 then they also give rise to the same amplitude reduction A_1.

The inset on Figure 4 plots the (I_B,B) equivalence with respect to threshold elevation derived from the curves of Figures 2 and 3. These points are replotted as the open circles on Figures 5 and 6 as follows: From the (I_B,B) equivalence, log I_B= -4 corresponds to B = 54%. The open circles on Figures 3 and 4 with abscissae -4 and 54 respectively are then positioned as closely as possible to the experimental data to yield the same

value of A in the two Figures, which is 0.25 in this case. The other open circles are obtained similarly. The curves on Figures 4 and 5 are drawn by hand and it can be seen that they are close to the open circles as well as to the experimental points. Thus, the (I_B,B) pair with the same value of ΔI also has the same value of A. This implies a close agreement between the equivalences (1) and (2).

The agreement between the equivalence relations can also be demonstrated by a statistical procedure (Leibovic et al., in preparation). The result is the same. We therefore accept the proposition that the two equivalences coincide. But, if this is the case, then do the equivalences at the beginning and end of the operating curves extend also to the complete curve? More precisely, suppose the thresholds and saturated amplitudes coincide on the operating curves for a given background I_B and a given bleached percentage B. Do the curves then coincide everywhere?

Figure 7 shows this to be the case. In view of the statistical variability of the data evident in Figures 2 to 5, the (I_B,B) pair with coincident operating curves will not necessarily correspond to the exact pair on the inset of Figure 4. The significant point is that if a particular (I_B,B) pair have the same threshold, they have the same saturated amplitude and the same operating curves.

Some Implications for Psychophysics:

Equivalent effects of bleaching and backgrounds were first established in psychophysics. It is now evident that there are equivalences operating at the cellular level as well. We have

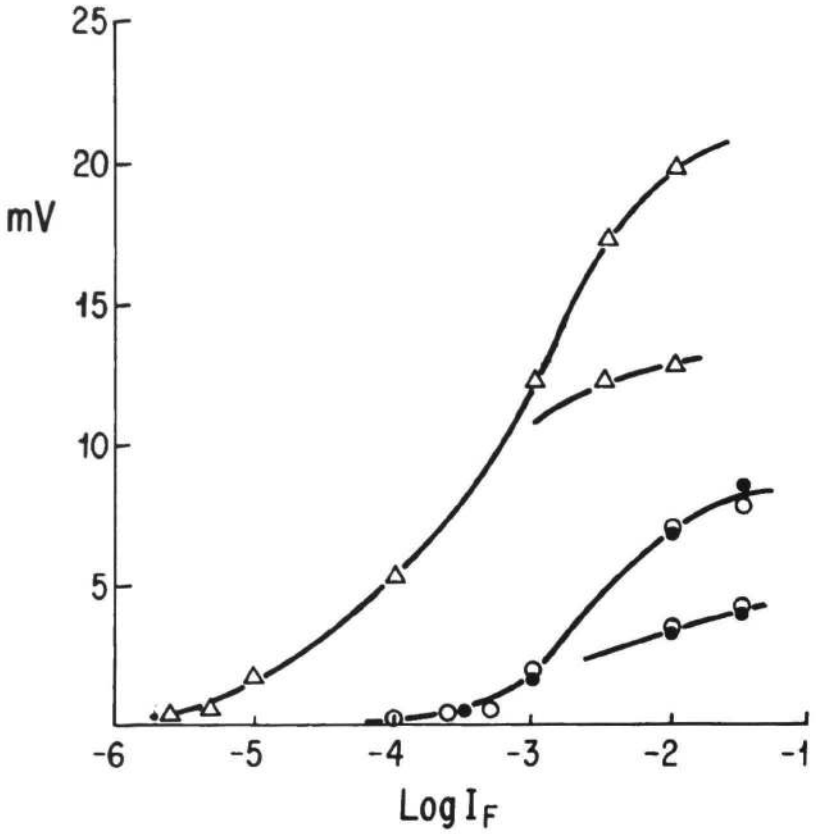


Figure 7: Operating curves for a cell in the dark adapted state (△), in the presence ofa back-ground of N.D. -5 (●) and after bleaching approximately 15% (o). Note the coincidence of the curves for N.D.-5 and 15% bleaching. Similar results were obtained for other (I_B,B) pairs up to background N.D. -3.5 and 70% bleaching.

demonstrated that the rod threshold in Bufo can be elevated some 5 or 6 log units before rising to saturation. This is the same as the psychophysically measured range in scotopic vision. At low background intensities the rod threshold obeys the Rose-de Vries rule and at higher intensities the Weber-Fechner rule, just like psychophysical threshold. In bleaching adaptation the rod threshold varies exponentially with the bleached fraction just as the psychophysical threshold, although the value of the exponent is different, in part at least due to the treatment of the data: in psychophysics too the value of the exponent differs in different experimental situations.

The compression of the rod responses due to adaptation is analogous to the fading of rod vision under photopic conditions.

Finally the equivalence between bleaching and backgrounds extends to the complete rod operating curve at any level of

adaptation. This result is especially interesting. For, the operating curve plots the peak responses to flashes of light. Each flash intensity is associated with a given peak response at any level of adaptation. There is therefore a 1-1 correspondence between light intensity and peak response. A given (I_B,B) pair which is equivalent with respect to threshold also results in equal peak responses to a flash of light. This is the cellular counterpart of the equal apparent brightness of a threshold equivalent (I_B,B) pair.

Can there be another feature of the response which could signal equivalent brightness? If so, it would not be the time integral of the response waveform or any other time dependent feature. A look at the different response waveforms due to bleaching and backgrounds suffices to confirm this: The response kinetics are significantly faster with a background than for the equivalent level of bleaching. But there could be another point on the waveforms in 1-1 correspondence with the peak response which would serve equally well. That is possible. However, from the point of view of signal detection, the peak response is the easiest to detect and transmit. We therefore propose that brightness perception is mediated by the peak photoreceptor response.

Our proposal would receive additional support if it could be shown psychophysically that not only is the subjective brightness the same for equivalent bleaching and background, but also that the perceived brightness of flashes of increasing intensity is the same, when superimposed on retinal areas of equivalent bleaching and background.

If the peak photoreceptor response conveys brightness information, we may ask what information is contained in the rest of the complex response waveform? It is a long lasting response which generates a visual persistence, a form of short term memory, which may be responsible for the perceived temporal continuity of the visual world.

A final point concerns the significance of electrophysiological signals: The neural response waveform reflects underlying biophysical and biochemical mechanisms. Thus the various phases of the action potential are shaped by sodium and potassium conductance activation and sodium inactivation. Similarly photoreceptor responses reflect changes in the electrical parameters caused by light absorption and the transduction biochemistry of phosphodiesterase, transducin, cyclic GMP, calcium and other products. As we have seen here, certain features of the response waveform can also be correlated with psychophysical phenomena. Electrophysiology, therefore, occupies the interface between sensory and molecular events.

Acknowledgements:

This work was supported by grants 5 R01EY0367203 from the National Eye Institute NIH; 2S07RR0540024 from the Division of Research Resources, and CAT857 from the NYS Science and Technology Foundation.

References:

Barlow, H.B., Sparrock, J.M.B. (1964): The role of afterimages in dark adaptation. Science 144:1309-1314.

Crawford, B.H. (1947): Visual adaptation in relation to brief conditioning stimuli. Proc. Roy. Soc. B. 134:283-302.

Dowling, J.E., Ripps, H. (1972): Adaptation in skate photoreceptors. J Gen Physiol. 60:698-719.

Hecht, S. (1937): Rods, cones and the chemicas basis of vision. Physiol. Rev. 20:831-850.

Leibovic, K.N. (1971): On the retinal basis of visual adaptation. Kybernetik 9(3):96-111.

Leibovic, K.N. (1972): Nervous system Theory. Academic Press.

Leibovic, K.N., Dowling, J.E., Kim, Y.Y. (in preparation): Background and bleaching equivalence in steady state adaptation of vertebrate rods.

Rushton, W.A.H. (1962): The Ferrier Lecture 1962: Visual adaptation. Proc. Roy. Soc. B 162:20-46, 1965.

MIDBRAIN MECHANISMS FOR ORIENTING VISUAL ATTENTION

ROBERT D. RAFAL
Division of Neurology
Brown University and
Roger Williams General Hospital
Providence, RI

ALBRECHT W. INHOFF
Department of Psychology
University of New Hampshire
Durham, NH

ABSTRACT

The role of midbrain visual centers for orienting attention was studied in chronometric experiments measuring the effect of pre-cues on simple reaction time to detect a peripheral luminance change. Two types of cues were tested: Exogenous cues--a peripheral luminance change which did not predict target location; and Endogenous cues--a central arrow which predicted the likely target location. Patients with peri-tectal midbrain degeneration from progressive supranuclear palsy showed deficits in orienting to both types of cues. In normal human subjects tested monocularly, we compared orienting into the temporal hemifield (which has more direct access to the midbrain superior colliculus) with orienting into the nasal hemifield. Exogenous cues produced equivalent speeding of detection at cued locations in both hemifields; but nasal cues produced more slowing of detection at uncued locations. Endogenous nasal cues produced earlier speeding of detection at cued locations than temporal cues; and at later intervals, they produced more slowing of detection at uncued locations. Both cortical and subcortical visual systems appear to be integrated in orienting to both exogenous and endogenous information. Whereas the subcortical pathway receives input mainly from the temporal hemifield, the cortical system is biased in orienting to the nasal hemifield; and its committment produces moreslowing of detection at unattended locations.

The selectivity of visual perception is guided by brain mechanisms which orient attention in the visual field. One manifestation of visual orienting is overt, consisting of movements of the head and/or eyes to achieve foveation of the attended spatial position. Covert orienting of attention can occur independent of motor activity, and serves to align internal attention systems with a spatial location. Neurophysiologic studies in behaving primates have indexed this covert orienting in terms of selective enhancement in neuron firing

rate in response to a visual cue which prepares the animal, while maintaining fixation, to detect a forthcoming target, or to prepare a saccade toward it (Wurtz, Goldberg & Robinson, 1980). Similarly, covert orienting can be studied in humans by measuring facilitations and inhibitions in simple reaction time (RT) performance resulting from preliminary visual cues which either correctly prepare the subject to detect the target at the cued location, or which summon attention elsewhere in the visual field (Posner,1980).

In one study using this approach, normal humans maintained fixation at the center of a video display, and made simple RT keypress responses on detecting a target which appeared, with equal probability, in either of two flanking boxes located equidistant to left or right. On each trial the taget was preceded by a precue which was the brief brightening of one of the flanking boxes. This cue predicted, with 80% probability, that the target would appear at the contralateral location. The results were quicker detection RTs at the location of the cue at early cue-target intervals; whereas for targets appearing more than 200 msec. following cue onset, RTs were quicker for targets contralateral to the cue, i.e. where the subjects had expected the target. This result suggests that attention may be summoned automatically by exogenous sensory signals, or it may be deployed endogenously based on a strategic set (Posner, Cohen & Rafal, 1982)

In the same communication we reported that covert orienting of visual attention was slowed in patients with progressive supranuclear palsy (PSP), and attributed this deficit to lesions of the phylogenetically older visual system of the midbrain superior colliculus and peri-tectal region. PSP, and the more frequently occurring Parkinson's disease, share the clinical and pathological features of degeneration of the substantia nigra and other basal ganglia structures. However, PSP is distinguished from Parkinson's disease pathologically by a conspicious degeneration of the superior colliculi and peri-tectal regions; and clinically by a distinctive impairment in making voluntary eye movements which affects vertical more than horizontal eye movements. Our experiment demonstrated that attention movements were also slower in the vertical plane. The cue used in that experiment, however, was the brightening of a peripheral box which predicted, with 80% probability, that the target would appear at the cued location. Since the exogenous and the endogenous contributions of the cue were redundant, we could not specify whether midbrain visual centers were involved in exogenous orienting, endogenous orienting, or both.

The main goal of the present experiments was to determine the role of midbrain visual centers in exogenous and endogenous orienting. The first experiment examined covert attention orienting in PSP patients and, as in the previous study of Posner Cohen & Rafal (1982), attention movements in the vertical plane were compared with those in the horizontal plane. However, two different types of precue conditions were also compared. In one condition, the exogenous cue condition, the precue consisted of the brightening of a peripheral box which did not predict (50% probability) the position of the target. Any effect of the cue on detection RT could, therefore, be attributed to automatic, stimulus driven orienting. In another condition, the endogenous cue conditon, a

central arrow was used which predicted, with 80% probability, the positon at which the target would appear. Since the arrow appeared at fixation and cued a peripheral locus only by its symbolic meaning, this condition was assumed to reflect endogenous control mechanisms.

Experiments 2 and 3 examined the effects of these two types of cues on covert orienting in normal human subjects tested under monocular conditions. These experiments took advantage of the different neuroanatomical connections of the subcortical retino-tectal pathway and the cortical geniculo-striate system. The geniculo-striate system is binoccular, and the temporal and nasal hemiretinas of both eyes are equally represented in the occipital lobe cortex. In contrast, the extrageniculate visual system is essentially monocular; this subcortical retino-tectal pathway is dominantly crossed, and each superior colliculus receives input mainly from the contralateral eye. Moreover, in each superior colliculus, the visual field of the contralateral eye is asymmetrically represented such that the receptive fields of the temporal hemifield dominate, whereas little input from the nasal hemifield is mapped on the colliculus. Therefore, under monoptic viewing conditions, a difference in covert orienting into the nasal and temporal hemifields should provide insight into the contributions of subcortical visual centers, which are relatively blind to signals occuring in the nasal hemifield.

EXPERIMENT 1

Subjects:

Six patients with progressive supranuclear palsy volunteered to participate in the study. Each patient had a moderate to severe impairment in the ability to move the eyes vertically, but little or no difficulty moving the eyes in the horizonal direction.

Apparatus and Procedure:

In a quiet, darkened room subjects faced a black and white TV monitor placed 37 cm in front of the patient at eye level. The index finger of the preferred hand rested on a single response key placed on a table between the subject and the TV monitor. The TV monitor and response board were interfaced with a microcomputer which controlled the stimulus display and the recording of RT.

Subjects fixated a central plus sign which was flanked by four unfilled, square boxes, one degree across, which were plotted 10 degrees of visual angle above, below, to the right and to the left of fixation. Subjects were instructed to maintain fixation at the central plus sign, and to respond with a quick manual keypress resonse whenever a target appeared in any of the boxes. The target was a large bright asterisk filling the box, and each of the four boxes was equally likley to contain the target on each trial. However, 50, 150, 350, or 550 msec prior to the onset of the target, a cue was presented to prepare the patient for target detection. In one precue condition, the

exogenous cue condition, the precue consisted of the brightening of one of the peripheral boxes for 300 msecs. This cue did not predict the location of the forthcoming target; the target was equally likely to occur at the location of the cue, or opposite to this position. Cues and targets were always in the same plane, i.e. horizontal or vertical. Trials in which the cue correctly indicated the target position are referred to as valid trials, and trials in which the cue incorrectly predicted the target position are referred to as invalid trials. Two blocks of 160 trials were run. In another precue condition, the endogenous cue condition, a central arrow, one degree of visual angle in length, was used to signal the position of target occurrence; e.g. a right directed arrow would indicate that the target would be likely to occur to the right of fixation. Eighty percent of trials were valid and 20% were invalid. Four blocks of 160 trials were run for each subject. The order of blocks with exogenous and endogenous cues was randomized within patients.

Results and Discussion:

Trials in which target detection responses were less than 150 msec or longer than 3500 msec. were excluded from analysis. Median RTs from the remaining trials for each subject in each condition were calculated and subjected to a repeated measures anlyaisis of variance (ANOVA). The within subjects' factors were type of precue (peripheral brightening vs. central arrow), validity (correct vs. incorrect signaling of the target position), direction (horizontal vs. vertical), and interval (50, 150, 350, and 550 msec).

Two of the main effects, validity and interval, were significant; $F(1,5)=14.21, p<.025$ and $F(3,15)=26.29, p<.001$; respectively. Detection RTs were shorter on valid trials than invalid trials, and decreased as the cue-target interval increased. Specifically, RTs were 952 msec, 910 msec, 882 msec, and 849 msec, for the 50 msec., 150 msec, 350 msec, and 350 msec, cue-target interval conditions, respectively. The main result was a significant interaction between validity and direction; $F(1,5)=6.71$, $p<.05$. Valid trials were responded to faster when cue and target appeared in the horizontal direction than when they appeared in the vertical direction (Table 1). There was, however, no difference between vertical and horizontal target detection RTs on invalid trials. This interaction was not qualified by the type of precue, and was present under exogenous and endogenous precue conditons.

These results reveal a direction specific impairment of covert attention orienting in our PSP patients. Covert orienting to the cued location occurred more effectively when the target occurred in the horizontal plane than when it occurred in the vertical plane, and this was true for both exogenous and endogenous cues These findings suggest that the subcortical visual system of the midbrain contributes to covert shifts of attention activated by either exogenous or endogenous information.

TABLE 1: ORIENTING IN PSP

		Cued	Uncued	Difference
Box Cue	Horizontal	793	904	111
	Vertical	868	898	30
	Difference			81
Arrow Cue	Horizontal	871	983	112
	Vertical	935	991	56
	Difference			56
Both	Horizontal	832	943	111
	Vertical	902	945	44
	Difference			67*

Mean RT in msec for PSP patients in Experiment 1 for horizontal and vertical targets appearing at cued and uncued locations. * $p<.05$

EXPERIMENT 2

In Experiment 1, we found that intact collicular function is necessary for normal orienting to both exogenous and endogenous cues. In a second experiment, endogenous orienting was examined under monocular conditions in normal subjects. A central arrow cue was used to indicate the peripheral position at which a target was likely to occur. Since this cue required endogenous processing, and since the central position of the cue provided information equally available to both superior colliculi, differences in orienting between temporal and nasal hemifields should reflect biases of cortical systems for orienting attention.

Subjects:

Fourteen neurologically unimpaired adults were payed to participate.

Apparatus and Procedure:

The apparatus and procedure of Experiment 2 were the same as the endogenous (arrow) cue of Experiment 1 except that only horizontal cues and targets were used. Cue duration was 200 msec, and cue-target stimulus onset asynchronies (SOA) were intervals of 50, 150, 300 and 500 msec. Subjects were tested monoptically by patching of one eye before each experimental block. The order of left versus right eye patching was counter-balanced across successively tested subjects who participated in four blocks of 160 trials each.

Results and Discussion:

After excluding RTs of less than 100 msec. and greater than 2500 msec, mean RTs for each subject in each condition were analyzed in an ANOVA. Within factors were: hemifield toward which the cue summoned attention (temporal versus nasal); cue (valid or invalid); and interval (50, 150, 300, 500 msec.). The mean RT for all fourteen subjects in each condition are shown in Fig. 1. As in Experiment 1, there were main effects of interval and of cue validity. RT decreased as the cue-target interval increased ($F(3,39)= 56.6$, $p< .001$.) RTs were faster at the validly cued locations at all intervals ($F(1,13)= 23.8$, $p< .001$.) The hemifield cued (nasal versus temporal) interacted with interval such that responses to targets which followed nasal cues were quicker at the two short intervals; whereas responses to targets which followed temporal cues were faster at the long two intervals ($F(3,39)= 3.3$, $p< .05$). The data from the two short intervals were cast into a separate ANOVA. This confirmed that RT to targets that followed nasal cues (whether the cues were valid or invalid) were significantly faster than those which followed temporal cues ($F(1,13)= 7.45$, $p< .05$). This advantage for targets which followed nasal cues was due to faster RTs on valid trials in which a cue summoning attention to the nasal hemifield was followed by a nasal target ($F(1,13)= 5.03$, $p< .05$). Reaction times to targets appearing in the temporal hemifield following nasal hemifield cues were also faster than those to uncued targets appearing in the nasal hemifield, but this difference was not significant ($F(1,13)= 1.29$, p= n.s.). The data from the two long intervals (300 msec and 500 msec) were also cast into a separate ANOVA which confirmed that, at these intervals, RTs following nasal cues were slower than those which followed temporal cues ($F(1,13)= 5.25$, $p< .05$). Fig. 3 shows that the faster RTs for targets following temporal cues were primarily a result of the invalid cue condition: nasal cues followed by targets in the temporal hemifield resulted in slower RTs than targets which appeared in the nasal hemifield following temporal cues.

The results of this experiment seem to indicate that endogenous control mechanisms for orienting attention are biased for orienting toward the nasal hemifield. This bias is reflected by the quicker responses to nasal targets following nasal cues at early cue-targets intervals; and by slower disengagement from the nasal hemifield to respond to temporal targets following invalid cues at the later intervals. Since the subcortical visual system of the midbrain can maintain surveilance of the visual periphery, but has little input from the nasal hemifields of either eye, such a bias by the cortical system toward the nasal hemifield seems to be to be a reasonable arrangement for efficient functioning of the visual system as a whole.

In Experiment 3, we examined orienting to exogenous signals in the temporal and nasal hemifields in normal subjects under monocular conditions. Since the nasal hemifield is relatively blind to the superior colliculus, we wanted to determine whether the cortical attention system was competent in responding to exogenous signals presented in the nasal hemifield.

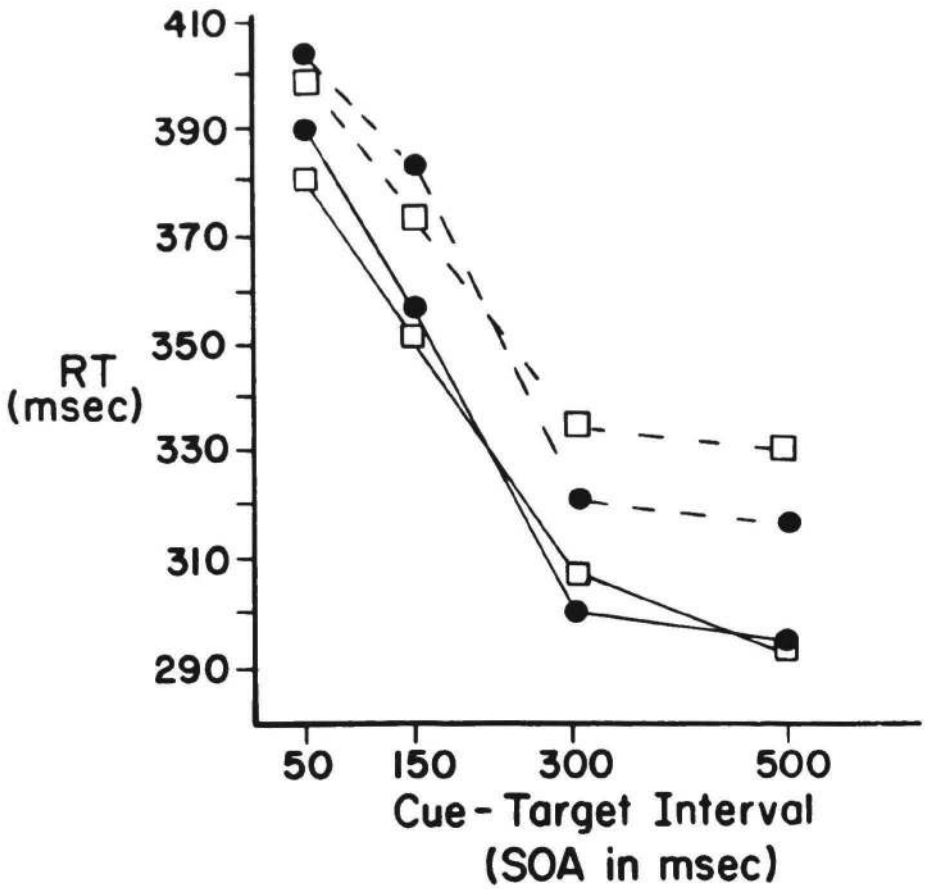


FIG. 1: ENDOGENOUS ORIENTING IN EXPERIMENT 2. Mean RT as a function of interval for target detection following arrow cues pointing toward temporal (solid circles) and nasal (open squares) hemifields. Solid lines indicate targets appearing at the cued location (valid trials); dashed lines indicate targets appearing contralateral to the cued location (invalid trials).

EXPERIMENT 3

Subjects:

Twenty-five neurologically unimpaired subjects were paid to participate, and were tested under monocular conditions as in Experiment 2.

Apparatus and Procedure:

The apparatus and procedure of Experiment 3 were the same as in Experiment 2, except that the cue was brightening of one peripheral box which did not predict (50% probability) target location.

Results and Discussion:

RTs shorter than 100 msec. and longer than 2500 msec were excluded from analysis. This led to an exclusion of approximately 2% of the data. The results from the remaining trials are shown in Fig. 2. Rsponses were faster on valid trials than on invalid trials ($F(1,24)= 6.05$, $p <.025$); and RTs decreased as the cue-target interval increased ($F(3,72)= 19.9$, $p <.001$). The two variables, validity and interval, interacted ($F(3,72)= 3.90$, $p< .025$), such that valid cues produced faster responses than invalid cues at the short intervals. This cue effect was reduced at the two longer intervals. Furthermore, validity interacted with hemifield ($F(1,24)= 4.37$, $p< .05$): Invalid trials resulted in slower responses when the cue appeared in the nasal hemifield and the target appeared in the temporal hemifield, than when the opposite sequence occurred. Inspection of Fig. 2 reveals that this interaction was present only at the two short cue-target intervals (50 msec and 150 msec). The interaction of validity and hemifield thus complements the finding of Experiment 2 which showed a deficit in the reorienting of attention in a temporal direction. In Experiment 2, this interaction was present at the longer cue-target intervals, presumably because the use of an endogenous cue required more time to fully commit attention, and encouraged subjects to maintain attention at the cued position. In Experiment 3, the interaction was present at the shorter cue-target intervals, presumably because the stimulus driven orienting of attention occurred relatively quickly and lasted briefly.

The major finding of this experiment is that exogenous signals which have no direct access to the superior colliculus do summon attention. This finding implies that cortical systems are involved in exogenous orienting. Their commitment, however, results in greater slowing to reorient to unattended locations.

GENERAL DISCUSSION

Attention may be oriented to facilitate the processing of information either internally in memory, or of events occurring at the sensory surface. When orienting to facilitate the processing of visual information, attention may be summoned either by an exogenous sensory signal, as when we turn toward a sudden movement seen out of the corner of the eye; or may be allocated endogenously, under internal control, in order to prepare to detect an expected stimulus, as when we decide to look both ways before crossing the street. In the studies reported here, we have attempted to relate both exogenous and endogenous mechanisms for orienting attention to neural systems; and specifically to determine what role the retino-tectal pathway of the phylogenetically older extrageniculate, midbrain visual system plays in visual orienting.

The results of the first experiment in patients with progressive supranuclear palsy were clear cut. Midbrain degeneration, including the superior colliculus and peri-tectal region, produced a deficit in orienting to

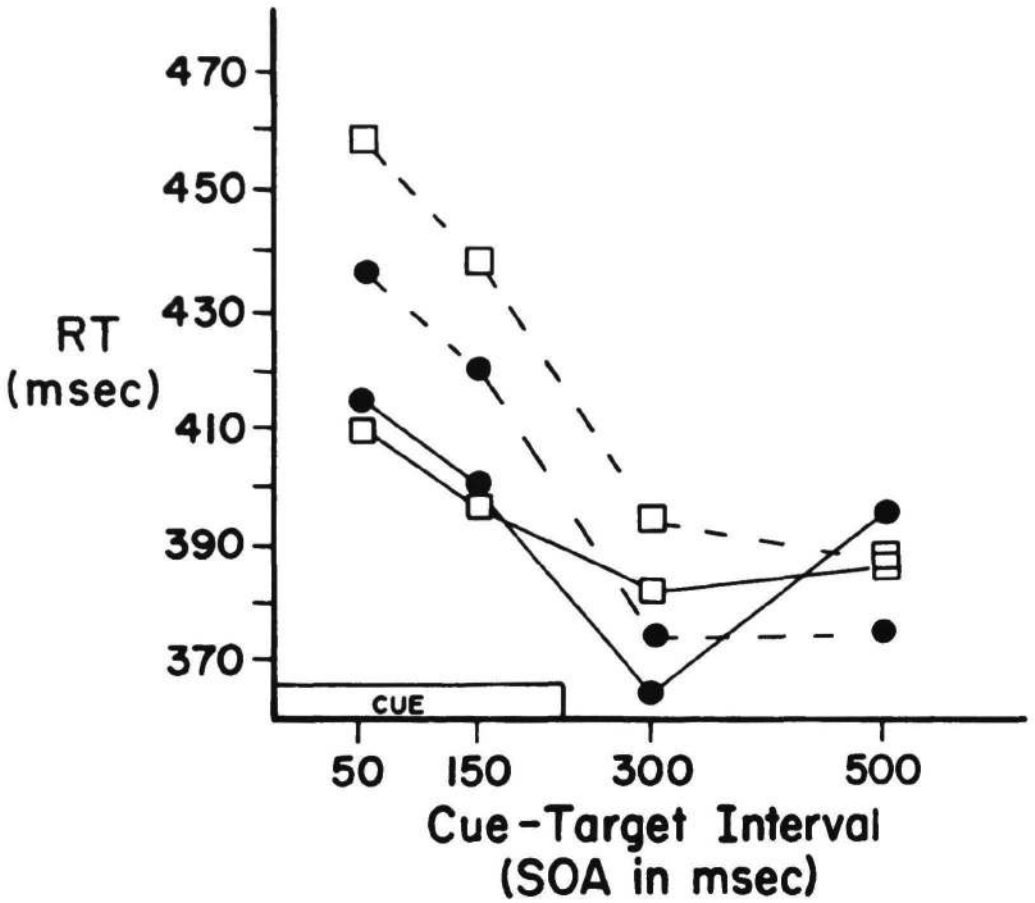


FIG. 2: EXOGENOUS ORIENTING IN EXPERIMENT 3. Mean RT as a function of interval for targets following a box brightening cue in the temporal (solid circles) and nasal (open squares) hemifields. Solid lines indicate targets appearing at the cued location (valid trials); dashed lines indicate targets appearing contralateral to the cue (invalid trials).

both exogenous and endogenous cues. Covert orienting of attention was found to be impaired in the direction in which eye movements were more limited in these patients. Moreover, the deficit in covert orienting was clearly related to the fact that cues summoning attention in the clinically affected direction produced less advantage for targets appearing at the cued locations; whereas there was no difference in reaction time to detect targets at the uncued locations. This result is consistent with the concept that the superior colliculus is involved in the operation of moving covert attention. It contrasts strikingly with observations in patients with cortical lesions of the parietal lobe in whom the opposite pattern was found: viz RTs to detect targets at cued locations were not greatly different in the visual fields ipsilateral and contralateral to the lesion; rather the differences found were at the uncued locations, implicating a disorder in disengaging attention to reorient toward the field contralateral to the lesion (Posner, et al, 1984).

In the PSP patients tested in Experiment 1, we found that collicular lesions affect both exogenous and endogenous orienting. In the hemifield experiments in normal subjects, we sought to determine how the retino-tectal system might contribute to exogenous and endogenous orienting, and what kinds of orienting behavior could occur independent of this system. We have recently found this hemifield method to be a useful approach which can provide converging evidence with patient studies. Posner and Cohen (1984) showed that an exogenous sensory signal which first summons attention, results in a subsequent inhibition of return which slows detection of targets at the recently cued location. We have shown that this inhibition of return is deficient in patients with progressive supranuclear palsy (Posner, et al , 1985). Recently we have also found that this inhibition of return effect is greater in the temporal than in the nasal hemifield of normal human subjects under monocular conditions (Rafal and Calabresi, unpublished). This same result is, in fact, reflected in our current results in Experiment 3, as may be seen from a close inspection of Fig. 2. At the 500 msec interval RTs at the validly cued location are slower than at the uncued location in the temporal hemifield, whereas this is not true for cues summoning attention into the nasal hemifield. We have reason to feel, therefore, that this hemifield method can be useful in examining the role of the retino-tectal pathway, and can converge with studies of patients with lesions of the superior colliculus.

We began these investigations with the hypothesis that exogenous orienting is specifically linked to the colliculus, whereas endogenous orienting is a cortical function. Our results cannot be interpreted to support such a simple division of labor. Collicular lesions affected both exogenous and endogenous orienting. In normals, hemifield differences were found for orienting to both exogenous and endogenous information. The pattern that emerges from these results is that cortical and subcortical mechanisms interact with one another in orienting in response to both exogenous and endogenous information.

Because of its neuroanotomic connections, the collicular system is competent to orient only to signals in the temporal hwmifield. In animals the nasal hemifield has little representation in the colliculus (Hubel et al, 1975; Pollack & Hickey, 1979) . Newborn human infants, in whom the retino- tectal pathway is well established, but who lack a fully developed geniculo-striate pathway, seemed to orient with saccadic eye movements only to stimuli in their temporal hemifield (Lewis et al, 1979). In considering a system in which cortical and subcortical centers interact, it would seem to be appropriately adaptive that cortical systems might compensate for the lack of nasal hemifield competence of the collicular system by being biased toward the nasal hemifield. The results of Experiment 2, using a central arrow (endogenous) cue, are consistent with such an arrangement. Orienting toward the nasal hemifield occured more quickly and required more committment of limited capacity attention resources.

In Experiment 3, we sought to determine whether the cortical system, operating on information from geniculo-striate input, played a role in

orienting to exogenous sensory signals which had no predictive value. Since nasal hemifield cues summoned attention (as indexed by speeding of RT to targets following valid cues) as efficiently as temporal hemifield cues, we concluded that cortical systems are involved in exogenous orienting. The committment of the cortical system to exogenous orienting was found, again, to require relatively greater limited capacity attention resources resulting in greater slowing to reorient to unattended locations.

In summary, the results of these experiments in neurological patients and normal human subjects support a model in which cortical and subcortical visual centers interact in orienting to both exogenous and endogenous information. Since subcortical centers have more direct access to input from the temporal hemifield, cortical centers are biased to orient toward the nasal hemifield. Furthermore, committment of the cortical system appears to require limited capacity attention resources to a greater degree than does the subcortical midbrain system, and its commitment results in slower reorienting to detect targets elsewhere.

This model must be considered to be tentative. In our hemifield experiments, no neutral cue was used, so that we did not measure costs and benefits directly. We did conduct a control experiment in six normal subjects tested monocularly in which no cue was given. No difference was found in RT to detect targets in the nasal and temporal hemifields. Nevertheless, in our cue experiments, we could not be sure whether the effects of cueing on detection in the two hemifields were due to the direction toward which attention was summoned by the cue, or were related to the hemifield toward which attention had to be reoriented after the appearance of the target. It will be necessary to measure the effects of nasal and temporal orienting on the reorienting to detect central targets to resolve this question. Finally, our use of the term "cortical" and "subcortical" are somewhat arbitrary. The colliculus and striate cortex are connected with one another both directly, and through relays through pulvinar nucleus of the thalamus. While our results suggest an interaction between cortical and subcortical centers, specification of the neural basis for this intereaction will require further investigation. The approach of seeking converging evidence from chronometric studies of neurological patients and normal subjects offers a promising avenue of investigation in attacking this kind of fundamental issue in cognitive neuroscience.

ACKNOWLEDGMENTS

We are grateful to Michael Posner, who provided guidance throughout the conduct of this study; and to James McIlwain and Catherine Downing for their helpful comments on the manuscript. The manuscript was typed by Corrine Hopp, and support for this work was provided by the Roger Williams General Hospital.

REFERENCES

Hubel DH, Le Vay S & Wiesel TW (1975). Mode of termination of retino-tectal fibers in macaque monkey: An autoradiographic study. Brain Res, 96:25-40.

Lewis TL, Maurer D & Milewski A (1979). The development of nasal detection in young infants. Invest Ophthal Visual Sci, Suppl. p. 271.

Pollack JG & Hickey TL (1979). The distribution of retino-collicular axon terminals in rhesus monkey. J Comp Neurol, 185:587-602.

Posner MI (1980). Orienting of attention. Qaurt J Exp Psychol, 32:3-25.

Posner MI & Cohen Y (1984). Components of visual orienting. In H. Bouma & D. Bouwhuis (Eds.) Attention and Performance X , pp. 531-556. London: Lawrence Erlbaum Assoc. Ltd.

Posner MI, Cohen Y & Rafal R (1982). Neural systems control of spatial orienting. Phil Trans Roy Soc B298:187-198.

Posner MI, Rafal R, Choate L & Vaughn J (1985). Inhibition of return: Neural basis and function. Cognitive Neuropsychol, 2:211-228.

Posner MI, Walker JA, Friedrich F & Rafal R (1984). Effects of parietal injury on covert orienting of attention. J Neurosci, 4:1863-1874.

Wurtz RH, Goldberg ME & Robinson DL (1980). Behavioral modulation of visual response in the monkey: Stimulus selection for attention and movement. Prog Psychobiol Physiol Psychol, 9:43-83.

REPETITION-SENSITIVE COMPONENTS OF NEURAL ACTIVATION: EVIDENCE FROM INTRACRANIAL RECORDINGS FROM THE HUMAN MEDIAL TEMPORAL LOBE

Michael E. Smith
Department of Psychology and Brain Research Institute,
University of California at Los Angeles

Eric Halgren
Department of Psychiatry and Biobehavioral Sciences,
and Brain Research Institute, UCLA; and,
V.A. Southwest Regional Epilepsy Center

Gary Heit
Program in Neuroscience, UCLA; and,
Stanford University School of Medicine

ABSTRACT

Neuropsychological studies indicate that hippocampal formation in the human medial temporal lobe (MTL) plays a crucial role in the formation and retrieval of memories for recent events. We have found that components of neural activity recorded from the MTL during recognition memory testing discriminate between repeated and nonrepeated words. Such recordings are made possible via intracranial electrodes implanted for the isolation of seizure foci in epileptic patients. Similar activity can be elicited during lexical decision tasks, where it is also sensitive to stimulus repetition. It has been suggested that repetition effects on lexical decision performance measures are a reflection of procedural learning, unlike the type of learning that underlies recognition memory performance, does not involve the MTL. However similar changes in the MTL response during both lexical decision and recognition memory suggests that a common mechanism contributes to repetition effects across tasks. These potentials provide a juncture for studies at both the psychological and synaptic levels of analysis.

INTRODUCTION

Many current biologically inspired models of memory and cognition represent mental states as patterns of activation occuring across a network of processing units (e.g. Hinton & Anderson, 1981). The units are more or less explicitly analogous to neural elements or integrated collections of neural elements. Many of these models share an assumption that the occurence of a pattern of activation can lead to an event-specific modification of the system. This modification is usually characterized as changes in the weighting of connections between units. Multiple sources of evidence indicate that in the human brain medial temporal lobe (MTL) structures-- in particular the hippocampus --mediate the initial enhancement of connection strength, at least during the acquisition of complex declarative knowledge. In order to explore this learning process we have made intracranial recordings from the hippocampus, parahippocampal gyrus, and emphasis on memory judgments for recognition of recent previous occurrence. In this paper some of our current findings from these studies will be

described.

NEUROBIOLOGICAL AND CLINICAL CONTEXT OF RECORDING STUDIES

The MTL is important for the formation and retrieval of memories for recent patterns of activation. Bilateral lesions to this area yield a dense and specific amnesia for recent events (Scoville & Milner, 1957). Unilateral lesions reflect hemispheric specialization for memory functions, thus, lesions to the MTL of the language-dominant hemisphere create a larger deficit on tests for memory for verbal materials than do lesions of the nondominant hemisphere. Stimulation studies indicate that some important MTL contribution to memory is made during the first 500 milliseconds following presentation of a stimulus (Halgren, Wilson, & Stapleton, 1985), both when that item is first presented for learning, and also when it later re-occurs as a recognition probe. Anatomically the MTL is characterized by converging excitatory inputs from both modality specific and transmodal association cortices (coding the results of perceptual and cognitive analyses), and diverging outputs to these same areas, creating a circuit for the cycling of activation between limbic and cortical areas (Van Hoesen, 1982). It is thus well situated to affect changes in the neural activation that presumably underlies experience. Hippocampal synapses are characterized by a plasticity potentially critical for trace formation. That is, activation of hippocampal neurons can result in morphological changes in their connectivity (Lynch & Baudry, 1984) and physiological changes in their conductivity that are specific to the pattern of activated synaptic inputs (McNaughton, 1983; Teyler & Piscenna, 1984).

Formal models of the neural basis of human cognitive memory (e.g. Gardner-Medwin, 1976; Halgren, 1984; Marr, 1971) suggest that the MTL helps retain information about novel juxtapositions of past inputs, functioning in concert with the association cortices and other areas to modify the connectivity of neural elements such that subsequent partial inputs will be sufficient to progessively reconstruct a pattern of activation similar to that of the original experience. In effect, the hippocampus might be viewed as a source of weak randomly wired connections between disparate cortical units. Selective enhancement of these connections might yield dynamic links to support memory until more permanent structural changes in synaptic connectivity are formed (cf. Feldman, 1982). Grossberg (1976) has emphasized the need for a mechanism to stabilize new patterns of activation if subsequent similar input patterns are to be coded into the same recognition category. Halgren (1984) has proposed that new ensembles of cortical elements are stabilized as a result of the emergence of excitatory association cortex-MTL-association cortex feedback loops from these random connections, leading to potentiation of the hippocampal synapses in this circuit.

Unilateral surgical excision of the MTL is a technique often employed in an effort to control medically intractable complex partial epilepsy. In many cases the surgical prognosis can be greatly improved

if the epileptic focus has been previously unambiguously localized to one MTL by monitoring the activity of the MTLs during the onset and spread of seizures (Delgado-Escueta & Walsh, 1983). Such monitoring is made possible by recording from chronically implanted intracranial electrodes. While awaiting the spontaneous occurence of seizures, the opportunity sometimes exists to record the electrical activity of the MTL while the patient is engaged in controlled cognitive tasks. Such recordings provide a unique window for monitoring neural activity possibly central to memory processes.

A common technique in human electrophysiological studies is to repeatedly sample activity elicited in conjunction with some sharp onset stimulus event, such as the presentation of words in relatively simple judgment tasks. Stimulus-locked recordings can be subjected to signal averaging techniques to produce a representation of the prototypical sequence of neural activity specific to the processing of stimuli of that class. Such representations are referred to as event-related potentials (ERPs). ERP components are typically identified as characteristic task-related peaks on the averaged waveform and referred to by their typical polarity and latency to peak amplitude. Isomorphisms are often made between such components of ERPs, and inferred concommitant cognitive processes. In contrast to recordings of the action-potentials from individual neurons, these field potentials reflect the summation of relatively co-temporal synaptic potentials across a population of neurons. They must then more closely represent gross envelopes of activation that parse or modulate mental events rather than closely reflecting specific information transactions. However even this gross measure is highly sensitive to item specific features such as the intrinsic meaning of a stimulus, its relationship to local context, and whether it has also occurred in the recent past.

INTRACRANIAL ERPs DURING RECOGNITION MEMORY TESTING

Scalp-recorded ERPs, inasmuch as they are one of the few measures by which neural activity can be noninvasively monitored in normal humans, have been intensively studied. Among many other things, such recordings have indicated that the waveforms elicited in normal subjects during testing for recognition of previous occurrence discriminate between repeated words and foils (foils in this task are words that are well known in the lexicon, but have not occurred before in the test context, i.e. 'nonrepeated' words). This discrimination is eliminated in subjects who have received surgical lesions that include medial structures in the temporal lobe of the language dominant hemisphere (Smith & Halgren, in preparation). Such an effect underscores a global contribution of the MTL to patterns of neural activation. However scalp-recorded ERPs represent the smearing of activity from multiple intracranial current sources, and thus are seldom informative with respect to the specific activity of any particular brain system. Recording directly from the MTL allows the monitoring of the activity of local neuronal populations. The hippocampus is a laminated structure, with the apical dendrites of its

pyramidal cells arranged in a highly regular parallel fashion. This architecture is appropriate for the summation of synchronous synaptic potentials, and thus potentially for the generation of ERPs. MTL-generated ERPs have been known for some time to be elicited in simple perceptual discrimination tasks (e.g. Halgren, Squires, Wilson, Rohrbaugh, Babb, & Crandall, 1980).

In recent years we have been recording MTL activity during more demanding cognitive tasks, such as the recognition memory (RM) task mentioned above. In this task subjects are presented with a series of 180 common words organized into nine randomly ordered blocks. Ten words are common to all blocks and serve as the target set. The other 90 are presented only once. Following the first block, subjects indicate by a keypress whether each word has also occurred earlier in the test. Words are exposed for 300 msec every 2500 msec, and a feedback tone occurs 1200 msec after word onset. Although this test is very easy for normal subjects, who consistently perform it with a hit rate of well over 90% across blocks and very few false recognitions, it is highly sensitive to MTL dysfunction. Subjects who have received unilateral left-sided MTL lesions score substantially below the normal range in RM, and hit rate approached the chance level in one amnesic patient who was tested.

Intracranial recordings are typically made simultaneously from multiple electrodes and recording contacts, sampling activity from both MTLs across multiple anatomical structures. Such recordings during RM are characterized by long-latency (300-800 msec to peak amplitude) and large amplitude (often 50 to 250 microvolts) ERP components, that often discriminate between repeated and nonrepeated words. The voltage and polarity of these ERP components can vary greatly both within and across MTL structures, with large differences often seen between even closely spaced electrode contacts (for a more detailed explanation of recording method and results, see Smith, Stapleton, & Halgren, 1986). Potentials recorded from the left amygdala of seven patients during RM are illustrated in figure 1. As can be seen in the figure a robust negative component was observed between 300-500 msec after stimulus onset (N460) in response to words on their first presentation. Although the synaptic basis of this component is speculative, it possibly represents the summation of synchronized synaptic activation of hippocampal pyramidal cells in response to diffuse cortical input. The N460 is much attenuated when it is elicited in response to repeated words, suggesting that the MTL (or its input) was modified in some way by the recent experience with the initial patterns of activation elicited by the set of repeated words. This difference is not simply due to a state change or nonspecific habituation, in that repeated and nonrepeated words are intermixed in a random sequence. Following the N460 to correctly recognized repeated words, a broad, typically positive potential (P620) was often observed at many MTL sites. A component with a similar MTL voltage distribution is also elicited by attended infrequent stimuli in simple perceptual discrimination tasks (e.g.

Halgren et al, 1980). Current evidence suggests that it is associated with a decrease in local neuronal firing (Altafullah, Halgren, Stapleton, & Crandall, 1986), possibly as a result of recurrent inhibition of hippocampal pyramidal cells (cf. Andersen, Eccles, & Loyning, 1964).

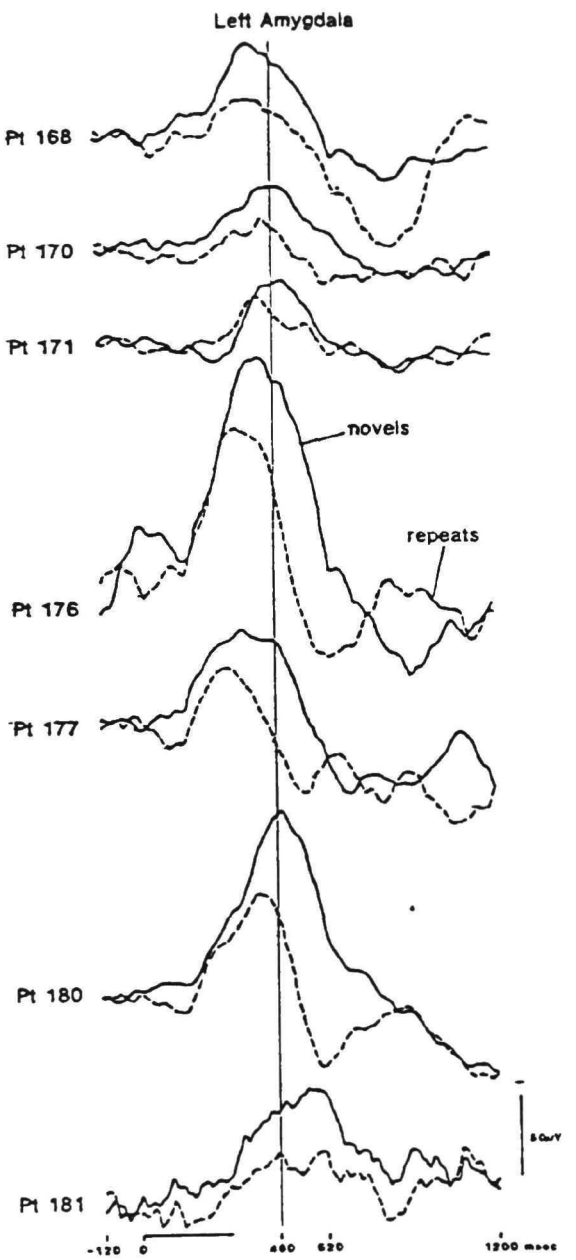


Figure 1. Average waveforms recorded from the left amygdala of 7 patients during recognition memory testing. Negativity is up in this and subsequent traces (from Smith, Stapleton, & Halgren, 1986).

Large amplitude, steep voltage gradients, and polarity reversals over short distances, strongly suggest that the N460 and P620 are generated within the MTL. Further evidence for local generation is provided by recordings from multicontact electrodes that allow neural activity to be sampled at regular spatial intervals from the medial temporal lobe sites out to the ipsilateral temporal cortex. Such recordings indicate the these potentials are consistently of largest amplitude at MTL contacts. In addition, in patients with severe MTL pathology, these components are often absent or extremely corrupted, a result that would not be expected if they originated from extra-MTL sources.

Thus, these potentials appear to be generated in a brain structure critical for cognitive memory, at a time when that structure might reasonably be assumed to be making some important contribution to performance. That, and their consistent sensitivity to recent occurences of similar past inputs, provide strong circumstantial evidence that these potentials directly reflect the neural activity inherent in the early stages of memory formation/retrieval. The N460 might well represent a pattern of activation that results in the enhancement of weak latent connections between cortical elements that is probably central to memory for the novel aspects of recent experience. The P620 might represent a complementary process of dampening that helps prevent the recoding of recently repeated items and promotes context reconstruction. Gardner-Medwin (1976) found that optimal reconstruction of a pattern from memory was obtained when the threshold for neural firing is initially low and then gradually increases over successive iterations in a progressive recall. With a low threshold, a cue such as a recognition probe is able to activate more of its original contextual associates. As the emerging pattern stabilizes, an increase in threshold inhibits the activation of tangential elements. The temporally progressive enhancement of the P620 for repeated words might thus reflect this type of threshold incrementation in the recognition process.

INTRACRANIAL ERPs RECORDED DURING LEXICAL DECISION

We have attempted to extend our findings by contrasting the components elicited during RM with those obtained from recordings in comparable tasks. Intracranial recordings from the MTL yield waveform records with a distinct voltage distribution across recording sites for each component in each subject. Thus, within individual subject's records it is possible to make systematic comparisons of the depth voltage topography of a component in order to ascertain its identity across different eliciting tasks. By utilizing this strategy we have found that N460 component is also elicited during lexical decision (LD) tasks, where subjects are required to discriminate words from nonsense letter strings (Smith et al, 1986).

Reaction time studies have indicated that repetition of items facilitates performance speed and accuracy in LD. In this task no overt judgment of recent previous experience is required, and learning

of the stimulus set is incidental rather then intentional. To test whether the repetition related decrease in MTL N460 amplitude could be generalized to these conditions, intracranial recordings were obtained during both RM and LD in the same subjects. Individual trial timing and visual stimulus quality was the same in both tasks. In this version of LD stimuli were structured in an initial block of 32 and three subsequent blocks of 64 trials. Order was randomized within blocks. Half of the trials in each block were letter strings that did not form words. The 32 items in the first block were subsequently

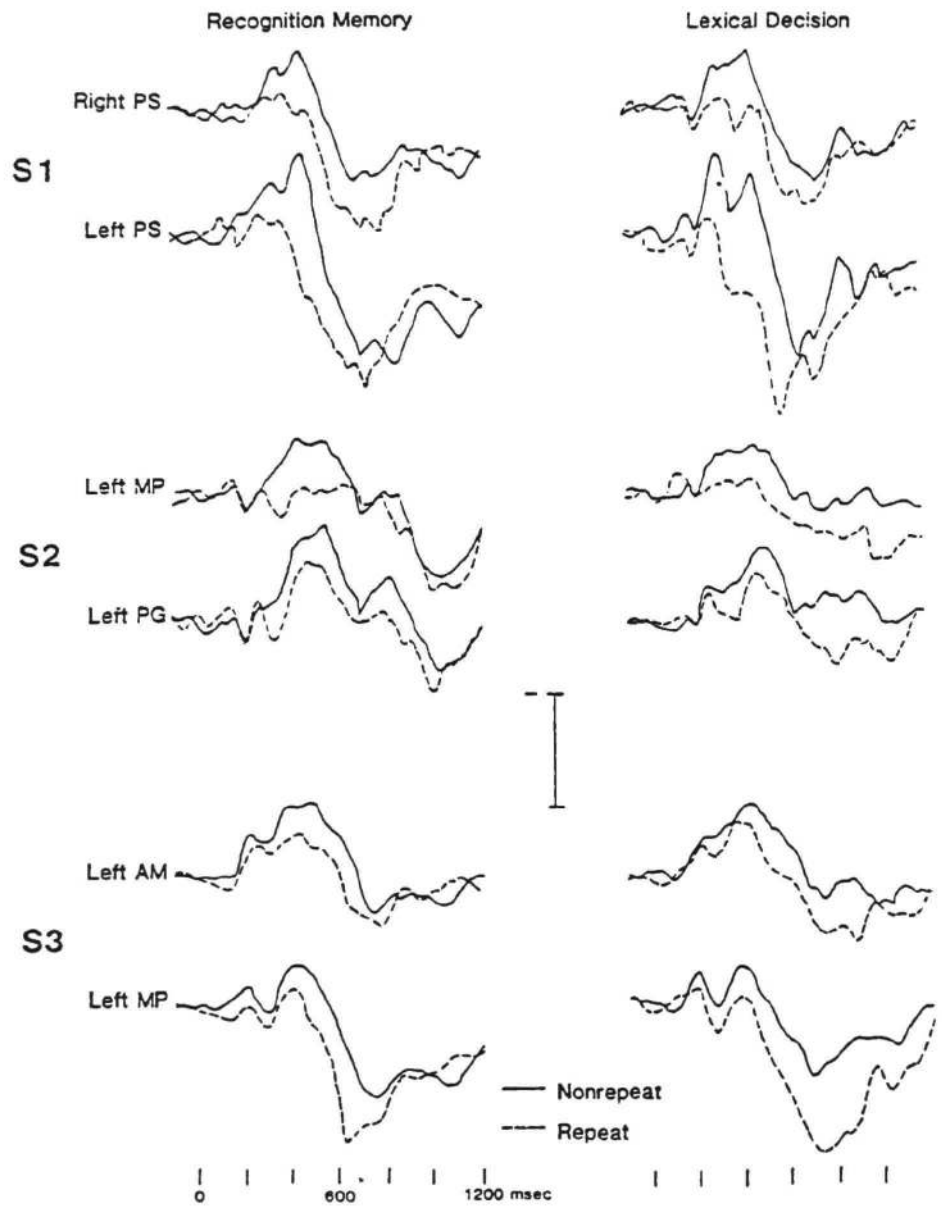


Figure 2. Comparison of repetition related changes in MTL recorded ERPs between recognition memory and lexical decision tasks. Scale is 50 microvolts. Negativity is up. PS=presubiculum, MP= middle pes hippocampus, PG= posterior parahippocampal gyrus, AM= amygdala.

repeated in each of the other three blocks. All other items were presented only once.

We have found that in normal subjects scalp-recorded ERPs elicited during LD discriminate between repeated items and those presented a first time. These differences are very similar to those observed between nonrepeated words and targets in RM (Smith & Halgren, 1986). Intracranial ERPs recorded during RM and LD in three patients are compared in Figure 2. In these three subjects seizure onset was localized to the nondominant MTL. They performed in the normal range on RM and LD, and displayed a substantial enhancement of response speed and accuracy following word repetition in LD. As can be seen in the figure a negativity extending from about 300-500 msec was elicited by words presented for a first time in RM. This potential was followed by a typically positive component that extending between 500-700 msec post-stimulus onset. The first component is analogous to the N460 potential observed in the earlier series of patients, again being smaller in amplitude when elicited by repeated words. A negativity similar in amplitude, MTL voltage distribution, and latency, can also be seen in response to items presented for a first time for lexical judgment. An effect of repetition of similar magnitude to that observed in RM was also observed in the late negative component recorded in LD. Similarly, the second component is analogous to the P620 potential described above, and can also be seen in LD. In both tasks, it is relatively larger in response to repeated items. It is of interest to note that this effect was obtained even though the set of repeated items in LD was much larger than in RM (32 vs 10), they were repeated fewer times (4 vs 9), repetitions were separated by a much longer average lag (over 2 min vs 50 sec, with many more intervening items), subjects were not instructed to remember the items, and no overt judgment of recent occurence was required.

Parallel voltage distributions and task correlates strongly support the idea that identical populations of neurons are being activated in a similar manner in both tasks. Comparable effects of repetition on ERP components in both RM and LD indicate that these effects are robust and that the MTL makes a general contribution to cognitive processing in both tasks. Moscovitch (1982) found that a group of subjects with memory disorders, including institutionalized elderly subjects and patients in the early stages of Alzheimer's disease, still show repetition related facilitation in LD. He thus proposed that facilitation in LD reflects changes in the 'procedural' system, and were independent of intact MTL function (cf. Cohen, 1984). However the extent of MTL pathology in those subjects was unknown. In contrast, the famous amnesic patient H.M., who suffers from well documented bilateral surgical removal of the MTL, has been reported to show no facilitation in LD when judging repeated words (Gabrielli, Cohen, Huff, Hodgeson, & Corkin, 1984). Further, recent psychological studies have found that 'episodic' sources of priming can have effects on LD performance paralleling priming from other sources (McKoon & Ratcliff, 1986). Thus, in addition to extending the conditions under which repetition effects on the MTL N460 might be obtained, these

results are consistent with the idea that repetition priming effects in lexical decision do not solely reflect procedural learning, but instead share task related components of neural activation with overtly MTL-dependent tasks such as recognition memory (for a similar conclusion derived from an insightful series of behavioral studies, see Ratcliff, Hockley, & McKoon, 1985).

SUMMARY

We have recorded field potentials from the human hippocampus and associated structures in order to better understand how these brain areas contribute to memory. These recordings have revealed repetition-sensitive components of MTL activation. One component is largest when a stimulus occurs for a first time within a context, and is smaller on subsequent presentations. This component might represent a pattern of activation that results in the enhancement of weak latent connections between cortical processing elements. A second component is instead enhanced to repeated items, possibly reflecting a dampening process that decreases the noise level in the network, or alternatively, that provides a mechanism to reset the system after it reaches a stable point (cf. Anderson & Silverstein, 1978; Grossberg, 1978). These components can be identified in both recognition memory and lexical decision, suggesting that the medial temporal lobe makes a wide-ranging contribution to repetition effects. Analyses of these components at both the synaptic and the psychological levels might provide a juncture for linking these domains of memory.

Supported by USPHS grant number NS18741 to E.H., and the Veterans' Aministration.

REFERENCES

Anderson, J.A. & Silverstein, J.W. (1978). Reply to Grossberg. Psychological Review 85:597-603.

Altafullah, I., Halgren, E., Stapleton, J.M., & Crandall, P.H. (1986). Interictal spike-wave complexes in the human medial temporal lobe: Typical topography and relation to cognitive potentials. Electroencephalography and Clinical Neurophysiology, in press.

Andersen, P., Eccles, J.C., & Loyning, Y. (1964). Location of postsynaptic inhibitory synapses on hippocampal pyramids. Journal of Neurophysiology 27:592-607.

Cohen, N.J. (1984). Preserved learning capacity in amnesia: Evidence for multiple memory systems. In L.R. Squire and N. Butters (eds.) The Neuropsychology of Memory. New York, Guilford Press, pp 83-103.

Delgado-Escueta, A.V., & Walsh, G.O. (1983). The selection process for surgery of complex partial seizures: Surface EEG and depth electroencephalography. Research Publications of the Association for Research on Nervous Mental Disorders 61: 295-326.

Feldman, J.A. (1982). Dynamic connections in neural networks. Biological Cybernetics 46:27-39.

Gabrieli, J.D.E., Cohen, N.J., Huff, F.J., Hodgeson, J., & Corkin, S. (1984) Consequences of recent experience with forgotten words in amnesia. Society for Neuroscience Abstracts 10:383.

Gardner-Medwin, A.R. (1976). The recall of events through the learning of associations between their parts. Proceedings of the Royal Society of London. B. 194:375-402.

Grossberg, S. (1978). Do all neural models really look alike? A comment on Anderson, Silverstein, Ritz, and Jones. Psychological Review 85:592-596.

Grossberg, S. (1976). Adaptive pattern classification and universal recoding, II: Feedback, expectation, olfaction, and illusions. Biological Cybernetics 23: 187-202.

Halgren, E. (1984). Human hippocampal and amygdala recording and stimulation: Evidence for a neural model of recent memory. In L. Squire & N. Butters (Editors) The Neuropsychology of Memory. New York, Guilford, pp. 165-181.

Halgren, E., Squires, N.K., Wilson, C.L., Rohrbaugh, J.W., Babb, T.L. & Crandall, P.H. (1980). Endogenous potentials generated in the human hippocampal formation by infrequent events. Science, 210: 803-805.

Halgren, E., Wilson, C.L., & Stapleton, J.M. (1985). Human medial temporal lobe stimulation disrupts both the formation and retrieval of recent memories. Brain & Cognition 4: 287-295.

Hinton, G.E. & Anderson, J.A. (1981). Parallel Models of Associative Memory. Hillsdale, NJ, Erlbaum.

Lynch, G., & Baudry, M. (1984). The biochemistry of memory: A new and specific hypothesis. Science 224: 1057-1063.

Marr, D. (1971). Simple memory: A theory for archicortex. Philisophical Transactions of the Royal Society of London B. 262, 24-81.

McKoon, G. & Ratcliff, R. (1986). Automatic activation of episodic information in a semantic memory task. Journal of Experimental Psychology:Learning, Memory, & Cognition 12:108-115.

McNaughton, B.L. (1983). Activity dependent modulation of hippocampal synaptic efficacy: Some implications for memory processes. In W. Seifert (ed.) Neurobiology of the Hippocampus. New York, Academic Press, pp. 233-249.

Moscovitch, M. (1982). A neuropsychological approach to perception and memory in normal and pathological aging. In F.I.M. Craik & S. Trehub (eds.) Memory and Cognitive Processes in Aging. New York, Plenum.

Ratcliff, R., Hockley, W., & McKoon, G. (1985). Components of activation: Repetition and priming effects in lexical decision and recognition. Journal of Experimental Psychology: General 114: 435-450.

Scoville, W.B. & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. Journal of Neurology, Neursurgery, and Psychiatry 20: 11-21.

Smith, M.E., & Halgren, E. (in preparation). Scalp-recorded field potentials during memory after anterior temporal lobectomy.

Smith, M.E., & Halgren, E. (1986). ERPs during lexical decision: interaction of repetition with concreteness, frequency, and pronounceability. Proceedings of the Eighth International Conference on Event-Related Potentials of the Brain, Palo Alto, CA, June 22-28, 1986.

Smith, M.E., Stapleton, J.M. & Halgren, E. (1986). Human medial temporal lobe potentials evoked in memory and language tasks. Electroencephalography and Clinical Neurophysiology 63: 145-159.

Teyler, T.J. & Piscenna, P. (1984). Long-term potentiation as a candidate mnemonic device. Brain Research Reviews 319:15-28.

Van Hoesen, G.W. (1982). The parahippocampal gyrus: New observations regarding its cortical connections in the monkey. Trends in Neuroscience 5:345-350.

Default Defeaters in Explanation-Based Reasoning

Gilbert Harman, Department of Philosophy, Princeton University
Richard Cullingford, Information and Computer Science, Georgia Institute of Technology
Marie Bienkowski, Bell Communications Research, Morristown, New Jersey
Ken Salem, Department of Computer Science, Princeton University
Ian Pratt, Department of Philosophy, Princeton University

The purpose of this paper is to illustrate an approach to the theory of reasoning that takes all reasoning to be "explanation-based". In particular, we consider how to treat "default reasoning" as a special case of explanation-based reasoning and we indicate what implications this treatment of default reasoning has for handling cases where the legitimacy of default reasoning is defeated by special considerations.

We are particularly interested in the following question about default reasoning. Given a default principle of the form, "Normally A's are B's," one can normally infer that a given A is a B. But sometimes further information about an A can block this inference. The question is: How should the rules of inference accommodate these exceptional cases?

One method that is used in certain production systems is to have several rules, one for the default rule and one for each of the exceptional cases: "From x is an A, infer x is a B." "From x is an A and x is a C, infer x is not a B." Etc. Furthermore, a restriction is placed on the rules of inference saying that, if the left hand side of a rule R is satisfied, one can use R only if there is no satisfied rule whose left hand side includes all the conditions of R's left hand side plus some further conditions. Given A only, one can then use the first rule to infer B. But given A and C one cannot use the first rule, since the second rule's left hand side is now satisfied.

This method supposes that one has already discovered whether the case is exceptional before deciding whether to infer from "x is an A" to "x is a B" using the default rule. This does not account for the case in which the current evidence would allow the inference that x is an exception but this has not yet been inferred. McDermott and Doyle (1980) handle this case by using rules of the following form: "Given that x is an A and that x cannot be inferred to be a C, infer that it is a B." But even this approach does not handle a case in which the evidence indicates that there is a significant chance that x is a C, without being so strong as to allow the inference that x is a C.

The research reported here was supported in part by a research grant from the James S. McDonnell Foundation, by a research grant (487906) from IBM, by the Defense Advanced Research Projects Agency of the Department of Defense and by the Office of Naval Research under Contracts Nos. N00014-85-C-0456 and N00014-85-K-0465, and by the National Science Foundation under Cooperative Agreement No. DCR-8420948 and under NSF grant number IST8503968. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the McDonnell Foundation, IBM, the Defense Advanced Research Projects Agency or the U.S. Government.

In some cases of this sort, the possibility of x's being a C is relevant because the reason why A's are normally B's is that A's are normally C's, and C's are always B's. We suggest that in order to infer from x's being an A to x's being a B one must be able simultaneously to infer that x is a C. If the evidence indicates that there is a significant chance that x is not C, then it will not be possible to conclude that x is C, and that will prevent the inference that x is B. As we will now indicate, this way of handling certain default defeaters fits in with a general framework of explanation-based reasoning.

REASONING

A preliminary account of explanation-based reasoning occurs in Harman (1986). Reasoning is identified with a nonmonotonic process of "change in view". Such a change occurs only in the presence of an interest or goal of the agent, for example, an interest in the answer to a particular question. The process of reasoning tries to respond to this interest or goal by making a minimal change in the agent's beliefs that improves the explanatory coherence of the whole set of beliefs by addition to and subtraction from that set. The process is subject to a number of constraints discussed in Harman (1986) that will not be discussed here.

We envision a computer program, AR (for Artificial Reasoner), that modifies representations of beliefs in accord with the principles of explanation-based reasoning (Cullingford, et al., 1985).. When AR draws a new conclusion, this conclusion will normally take the form of a complex explanatory structure, in which beliefs are linked together by relations of intelligibility. Sometimes, explanation-based reasoning will involve new beliefs that are inferred as the best explanation of the truth of certain old beliefs. For example, when doctor AR infers that a patient has a particular disease, AR's conclusion is that the patient's having this disease explains why the patient has such and such symptoms. Sometimes explanation-based reasoning will introduce new beliefs whose truth is inferred to be explained by certain old beliefs. For example, when predictor AR predicts that an agent will do a particular action, AR's conclusion is that such and such motives will lead the agent to do that action and so will explain the agent's action. Other more complex cases are discussed in Harman (1986).

The unit of inference is an explanatory structure. When AR is considering whether or not to accept a given conclusion C, AR must consider whether there is some explanatory structure AR can accept of which C is a part. So, AR will have some principles for determining what possibly acceptable explanatory structures there are. AR will also have principles for deciding among competing explanatory structures (e.g. for choosing the best explanation). (These principles need not be wholly separate, because the possibly acceptable explanatory structures might be produced in an order that indicates how good they are. For example, there might be a preference for explanatory structures that involve fewer rather than more new beliefs. There might also be a preference for explanatory structures that account for more rather than less.)

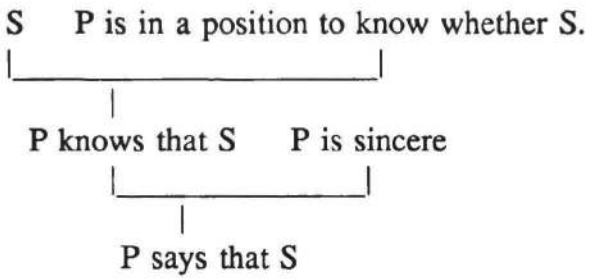


FIGURE 1: A (SLIGHTLY) COMPLEX EXPLANATION

STRUCTURES

In this view, beliefs are organized into complex explanatory structures. The basic links in such structures are immediate explanations. Immediately intelligible links represent connections that AR grasps without having to note intermediate links. An immediate explanation or *e-node* has two components, an *explanans* or list of (pointers to) immediately explaining beliefs and an *explanandum* or (pointer to) something immediately explained. Every belief is associated with *explainer* links to e-nodes of which the belief is the explanandum and *explained* links to e-nodes of which the belief is one of the explanans.

A complex explanation is a structure of immediate explanations, perhaps a tree, with the *ultimate explanandum* (thing explained) at the root, where it and other propositions in the tree have as immediate descendents e-nodes of which they are the explananda, where these e-nodes have as their descendents the propositions that are the explanans of the e-nodes. (Figure 1)

Some e-structures are more complex than this, since a new hypothesis might allow the explanation of more than one thing. For example, doctor AR should prefer a diagnosis that accounts for several of a patient's symptoms over a diagnosis that accounts for only one symptom. That involves an explanatory structure with more than one root. So we have to allow for cases in which there is more than one e-node below a proposition. It is not clear what to call this structure, but it consists basically in links among propositions and e-nodes.

So, we assume that AR has procedures that produce possible explanatory structures of this sort containing the proposition it is "considering" and minimizing the number of new beliefs it adds and old beliefs it gets rid of.

INFERRING STRUCTURES

Suppose AR knows that A is F and AR also knows that, normally, x is F only if x is G. But AR does not know why this is so. In particular AR does not know whether something's being F is responsible for its being G, whether something's being G is responsible for its being F, or whether some other thing is responsible for this correlation. Still, AR can infer that A is G. But how is that to be represented as an explanatory inference?

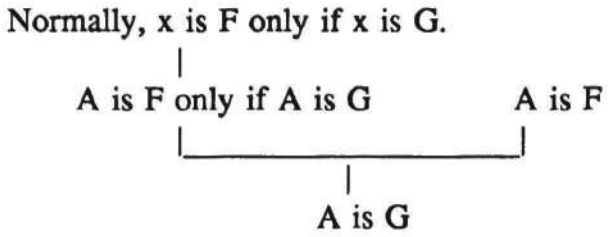


FIGURE 2: EXPLANATION USING A DEFAULT PRINCIPLE

We suggest the following answer: AR infers that the existence of the general correlation between something's being F and something's being G will account for the correlation in this particular case. That is, AR will add an e-node whose single explainer is (a link to) "Normally, x is F only if x is G" and whose explanandum is (a link to) "A is F only if A is G". AR will also add an e-node whose explainers are links to "A is F only if A is G" and "A is F" and whose explanandum is "A is G." These two e-nodes and the propositions they are linked to make up the explanatory structure that AR infers on this occasion. (Figure 2) (This is not to say that AR has to retain this structure in memory as time goes on. Rather: this is what AR accepts for the moment in coming to accept "A is G".)

How does AR infer from "S says that P" to "P"? Perhaps via the generalization, "Usually, x says that m only if m." Then this is an instance of the sort of inference just discussed.

However, AR might also have a view as to why the generalization holds. AR might believe that the generalization holds because, usually, when x says that m, that is because x believes that m and x wants to say whether m, and, furthermore, x believes that m because m and x is in a position to know whether m. It is only because AR accepts such an explanation that AR can avoid inferring "P" from "S says that P" on those occasions on which AR believes S does not want to say whether P or on those occasions on which AR thinks S is not in a position to know whether P. (Eventually, we consider how AR's acceptance of this sort of explanation allows AR to avoid these bad inferences.)

To represent this we need to allow for explanatory structures that link propositions with variables in them. The relevant structure here contains an e-node whose explanandum is (a link to) "x says that m" and whose explainers are (links to) "x believes that m" and "x wants to say whether m". It also contains an e-node whose explanandum is (a link to) "x believes that m" and whose explainers are (links to) "m" and "x is in a position to know whether m." (Figure 3)

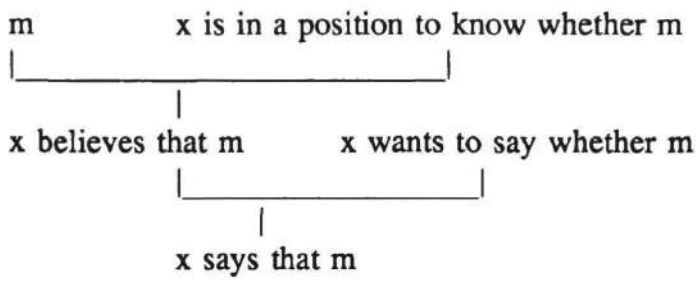


FIGURE 3: EXPLANATORY STRUCTURE WITH VARIABLES

Call this structure ST, then the relevant belief is that, usually, when x says that m, then ST. Notice that this involves quantifiers whose scope is the whole explanatory structure. (Also the phrase, "usually when x says that m," is a kind of quantifier here. Its scope is also the whole structure.)

LIMITED REASONING

The amount of processing required to tell whether a conclusion is currently inferable in N steps is an exponential function of N. Furthermore, anyone who has taught a logic course knows that students often have trouble with more than one or two steps of inference at a time. Proofs have to be broken down into manageable stages, each of which must be absorbed before going on to the next. In general, people are capable of only a few steps of inference at any given time. So we wanted AR to be subject to the same limitations.

Our first implementation of this idea was to limit AR to N steps of immediate implication or immediate inductive projection, where N is 1 or, anyway, small. But the reasoning involved in many elementary activities, e.g. story understanding, often involves rather complex chains of inference.

Reflection on various examples seemed to us to indicate that how many steps of immediate inference are possible depends on how familiar the area is. So, our second implementation of AR replaced the limitation to N steps of inference with a set of rules specifically spelling out what deductions or projections AR was capable of at any given time, where these rules might allow (in principle) for unlimited chains of implication and projection, if the chains were of a "familiar" sort. (What chains of reasoning were to count as familiar dictated the choice of rules. The notion of familiarity did not play an explicit role in the rules.)

Now, in order to make further progress here, we saw we had clearly to distinguish between a step of *inference* (i.e. a step of *belief revision)* and a step of *implication* or *inductive projection.* This distinction is easy to appreciate for the case in which belief revision includes the elimination of some prior belief, because eliminating a belief is clearly different from inferring an implication or projection from prior beliefs. But the distinction is also important for reasoning that does not eliminate any old beliefs, in other words, for reasoning that adds new beliefs that are implications

and/or projections of old beliefs. What gets added in such a case is (we claim) an implicational-explanatory structure. There will often be a complex chain of implication and projection in the structure. In a sense this represents a complex chain of reasoning. But in another sense (we want to say) this might be only one step of reasoning.

For example, a person may accept as a background "belief" a complex general explanatory structure. A single step of inference might involve the acceptance of a particular instance of that general structure, accepted as an instance of that structure. Then the conclusion accepted can involve a number of steps of implication and projection even though there is only one step of inference between the general explanatory belief and that conclusion.

So, we can combine the original idea that there is a limit to the number of steps of reasoning a person can do at any one time with the observation that some reasoning involves a complex chain of considerations, once we distinguish between a step of reasoning and a link in an implicational-explanatory chain.

DEFAULT REASONING

Suppose AR believes -

Normally, if x is an F, x is a G.
Normally, if x is an F and x is an H, x is a Q.
G and Q are contraries.
a1 is an F.

Then AR can accept the conclusion "a1 is a G." More specifically AR can accept the explanatory structure in Figure 4. This involves adding one new belief, namely "a1 is a G", which is linked to beliefs previously accepted.

The competing explanatory structure involves adding two new beliefs (Figure 5). The two new beliefs added here are "a1 is an H" and "a1 is a Q". The explanatory power of this structure is the same as that of the previous structure, so the previous structure is preferred to this one.

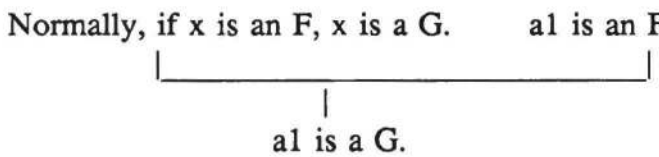


FIGURE 4: DEFAULT STRUCTURE

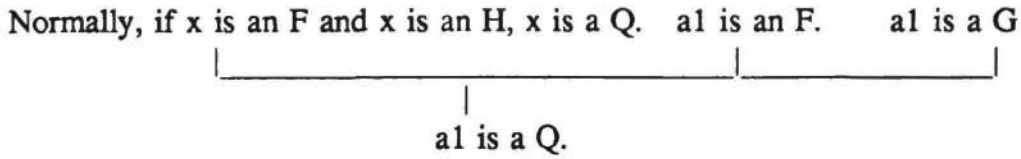


FIGURE 5: COMPETING STRUCTURE

If AR's evidence included both "a1 is an F" and "a1 is an H", then the second structure would be preferred because it links "a1 is an H" to the newly inferred belief, whereas the first structure would only link "a1 is F" to the newly inferred belief.

McDermott and Doyle (1980) would allow the first inference only given the further premises, "I have no reason to infer 'a1 is H'". But this premise is not needed on the explanation-based reasoning approach.

DEFEATERS

Suppose AR believes

(1) Normally, if x is F, then x is G.
(2) a is F.

Then AR ought to be able to infer a is G. But not if AR also believes:

(3) (1) holds because, normally, if x is F, then x is H, and any H is G.

together with other things that prevent AR from using (3) to infer that a is H, either because AR believes things that imply a is not H, or because AR believes things that imply that there is a significant chance that a is not H despite (2) and (3).

Notice that we cannot capture this within a "nonmonotonic logic" by asserting that, when something like (3) is believed, (1) must be replaced with

(4) Normally, if x is F and *it is not inferable that* x is not H, then x is G.

This is not enough of a modification in (1), because it would not prevent the inference in certain cases in which the inference should be prevented, namely, those in which it is not inferable that x is not H but it is also not inferable from (3) that x is H (e.g. because we have other evidence that indicates a significant likelihood that x is not H (for example the evidence might indicate that there is a 50-50 probability that x is not H).

ALGORITHM

Here is a quasi-algorithm that AR uses, given an interest in answering a question. It first forms a list of possible e-structures it might infer that contain an answer. It orders this list in terms of how good these explanations are as measured by the extent

of change required (the more change the worse the explanation) and effect on explanatory connections (the more the better). This is discussed in Harman (1986) without settling on a precise measure. We do not have space here to discuss possible measures. In any event, call the ordered list L. AR takes the first item from L. Call this item I. AR forms a list of competing e-structures that might be inferred. It considers whether I is better than any e-structure in this last list. If so, it infers I. If not, it considers the next item in L and goes back three steps. If L is empty, no answer to the question can be inferred.

INTERSECTION EXAMPLE

We conclude with a different sort of example which we have been examining. Suppose AR starts with an interest in answering the question, "Do the two streets Harrison and Aiken intersect?" It collects a list of e-structures it might infer that contain an answer, using backward chaining. In this case the possible answers are yes and no, i.e. "Harrison and Aiken do intersect" and "Harrison and Aiken do not intersect".

AR discovers the following possible e-structure that it might infer: Harrison and Aiken do intersect, because Harrison and Aiken are near each other and perpendicular to each other, and, normally, when roads are near each other and perpendicular, they intersect. This e-structure is inferable only because AR already believes (1) that Harrison and Aiken are near each other and perpendicular to each other and (2) normally, when roads are near each other and perpendicular, they intersect.

AR discovers no other inferable e-structure.

AR next considers whether there are competing e-structures that might be inferred and discovers none. So AR infers that Harrison and Aiken intersect.

The rules of backwards chaining applicable in a case like this are quite similar to ordinary backward chaining inference rules EXCEPT that they lead to an e-structure containing all the "premises".

It might be good to say a bit more about "near" and "perpendicular". In saying that Harrison and Aiken are near and perpendicular what is meant is that there is a point X on Harrison and a point Y on Aiken such that X is near Y and the orientation of Harrison at X is perpendicular to the orientation of Aiken at Y. AR might know about such points, e.g. the intersection of Aiken and Princeton and the intersection of Harrison and Nassau.

Now suppose that AR starts by believing this:

(*) Normally, if
X is a point on S1,
Y is a point on S2,
X and Y are near each other, and
the orientation of S1 at X is perpendicular to the orientation of S2 at Y,
then: S1 intersects S2.

Then AR comes to believe that (*) holds because

(a) the lines going through X and Y with the orientations of S1 at X and S2 at Y intersect in a point Z that is near or at both X and Y and
(b) normally, given a point P on a road R, the road continues at least a short distance along in the same direction from that point, so that, if P' is near or at P and P' is on the line through P that has the same orientation that R has at P, then R continues from P to R -- so
(c) S1 continues to Z and S2 continues to Z, so
(d) S1 and S2 are both at Z, so
(e) S1 and S2 intersect at Z.

Now, if AR is considering whether Harrison and Aiken intersect, backchaining leads to a much more complicated e-structure, which might be expressed in words as follows: Harrison and Aiken do intersect because there is a point X on Harrison and a point Y on Aiken such that X is near Y and the orientation of Harrison at X is perpendicular to the orientation of Aiken at Y and that means there is a point Z that is near X and near Y and Harrison continues beyond X to Z and Aiken continues beyond Y to Z, where Z is the intersection of the lines going through X and Y that are oriented as Harrison and Aiken are at X and Y, and where all this is so because of the generalizations alluded to above.

This modification of backchaining requires that some change be made to the original generalization linking it to its explanation. The backchaining rule has to be sensitive to this link in such a way that it yields the e-structure just given.

This means that AR will be unable to infer that Harrison and Aiken intersect if it cannot infer the more complex e-structure, perhaps because AR believes that Aiken does not continue on in the indicated way but instead dead ends into a park.

BIBLIOGRAPHY

Richard Cullingford, Gilbert Harman, Marie Bienkowski, & Ken Salem (1985) "Without Logic or Justification: Realistic Belief Revision," *Proceedings of the National Academy of Sciences Workshop on Artificial Intelligence and Distributed Problem Solving*. Washington, D. C. National Academy of Sciences.

Gilbert Harman (1986) *Change In View: Principles of Reasoning*. Cambridge, Massachusetts. M.I.T. Press.

Drew McDermott & Jon Doyle (1980) "Non-monotonic logic, I," *Artificial Intelligence* 13: 41-72.

SPHEXISHNESS, EPISTEMIC BOUNDS, AND A PRIORI PSYCHOLOGY

M. P. Smith
Department of Computer Science
Duke University

We are all too aware of limitations on human cognitive capacities. All of us routinely run up against the boundaries of our abilities to calculate swiftly, to take several factors into account at once, to recall information. Many of the experimental techniques of cognitive psychology involve the measurement of these limits. Lately it has been suggested that we may expect the human mind to have limitations of a different order, qualitative rather than quantitative in character. These are not merely upper bounds on the size, speed, reliability, etc. of calculation, but limitations on the scope of human understanding. I have in mind here Fodor's claim that our minds are likely epistemically bounded and the Dennett-Hofstadter thesis that they are sphexish.

These claims are not without precedent. Fodor cites Aquinas, who borrows from Aristotle, who attributes to Anaxagoras the argument that, since the intellect can think all things, it must be "unmixed with body" (De Anima 111.4).[1] Dennett cites Descartes (Dennett 1978, p. 245). Traditionally the claim that materialism implies limitations on the understanding has been used to argue for immaterialism. This immaterialist modus tollens has become the materialist's modus ponens. In this paper I consider the justification for the implication itself, however it may be used, and also try to clarify the content of the claim being made.

I

Sphexishness has as its paradigm case the behavior of the sphex wasp as described in an oft-quoted passage from Wooldridge (Wooldridge 1963, p. 82). The wasp's behavior of leaving her prey at the entrance of the burrow, entering unencumbered at first, and then retrieving her burden, smacks of prudent foresight. One can't be too careful! But if, while she is checking things out inside, a meddlesome finger nudges her prey a few inches from the doorway, she will repeat the procedure, even though presumably she can see that no enemy has snuck into her burrow in the meantime. Further, she will repeat this routine endlessly, or at least to the end of an experimenter's patience. So one <u>can</u> be too careful. After a few dozen repetitions, however, it becomes obvious that the wasp is not being careful. At first she seems prudent, then stupidly over-cautious. Eventually, though, we abandon the whole framework of rational appraisal and realize that the poor wasp is just stuck like a broken record (Hofstadter 1985, p. 530). Sphexishness is apparently rational behavior revealed to be really merely reflexive through its rigid persistence where it is no longer appropriate. Hofstadter calls the sphex's performance "a rather shocking revelation of the mechanical underpinning ... of what looks like quite reflective behavior" (Hofstadter, p. 529).

Note that sphexish behavior is not merely sub-optimal or even irrational, but non-rational. Admittedly the line between behavior which is irrational and that which falls outside the pale of rational criticism is not sharp. To claim that some human behavior is sphexish, however, is to say that it is only apparently done for reasons, not that it is done for poor reasons.

Wooldridge's original analysis is that the rigidity of the sphex's apparently thoughtful behavior reveals it to be pre-packaged and triggered,

rather than deliberately chosen. He compares it to an interrupt-driven subroutine (Wooldridge, pp. 83 - 85). Sphexishness would arise in such a system when undiscriminating interrupts are in charge of extended behavioral routines.

At the end of Part V of the Discourse on the Method Descartes proposes two features which separate people from both animals and machines. The first and more famous is the ability to use language. The second has not been well understood, but I believe that it amounts to a claim that people are antisphexish.

About the abilities of animals Descartes says:

> ...although many animals show more skill than we do in some of their actions, yet the same animals show none at all in many others; so what they do better does not prove that they have any intelligence, for if it did then they would have more intelligence than any of us and would excel us in everything. It proves rather that they have no intelligence at all, and that it is nature which acts in them according to the disposition of their organs. (Cottingham et al. 1985, p. 141)

Clearly Descartes must be referring to apparently rational abilities on the part of animals, and not just, say, the cheetah's speed or the monkey's agility, for his denial that these skills prove intelligence to be apposite. Descartes is claiming that animal cunning is in fact just sphexishness.

Why Descartes thinks that the singularity of the spider's skill at construction or the bird's navigational prowess shows these to be merely mechanical rather than truly intelligent behaviors comes across more clearly when Descartes distinguishes between the abilities of people and machines.

> ...even though such machines might do some things as well as we do them, or perhaps even better, they would inevitably fail in others, which would reveal that they were acting not through understanding but only from the disposition of their organs. For whereas reason is a

> universal instrument which can be used in all kinds of situations, these organs need some particular disposition for each particular action; hence it is for all practical purposes impossible for a machine to have enough different organs to make it act in all the contingencies of life in the way in which our reason makes us act. (Cottingham et al., p. 140)

Before the term 'sphexishness' was coined by Hofstadter, Daniel Dennett, citing Descartes, claimed that the lesson of the sphex was that "any behavior controlled by a finite mechanism must be tropistic" (Dennett 1978, p. 245). Apparently Dennett meant by a "finite mechanism" an elaboration of the sphexian architecture as imagined by Wooldridge: a system of hard-wired patterns of response set off by pre-arranged signals. The Dennett-Descartes claim then would be that no finite system of routines and interrupts could act with the resourcefulness of a being endowed with universal reason.

This thesis is indefensible, however, simply because a finite system can still be very large. The number of situations a human being can expect to meet, with however small a delta factor of probability one chooses, is still finite. (I say this while cheerfully admitting I have no idea how to count possible situations, much less estimate their probability. But the product of the number of nanoseconds in the maximum human life span times the number of cubic angstroms within, say, thirty parsecs of earth, times the number of distinct physical states each of these space-time parcels can be in, is still comfortably finite.) Finitude is no constraint at all when there are only finitely many opportunities. Dennett should have followed Descartes' lead and talked about the limitations of biologically or physically feasible mechanisms, rather than merely finite ones.

A system with a finite number of primitive responses is still capable of an infinite number of responses anyway, provided that we allow combinations of

responses of any length. The advantage of universal reason over finite mechanism is not to be determined by the cardinality of behavioral repertoires. If the alleged human sphexishness amounts to more than the admission of upper bounds on the size, speed, and accuracy of the information processing that goes on within them, then it must be based on more than human finitude.

II

It is tempting to read Descartes' argument for the sphexishness of animals as anticipating Fodor on the consequences of modularity. If intelligence consists in computationally discrete, informationally encapsulated, special-purpose organs, "then it is surely in the cards that there should be some problems whose structure the mind has no computational resources for coping with" (Fodor 1983, p. 120). In particular, Fodor proposes that a likely consequence of the mind's modularity is that it is epistemically bounded: "our cognitive organization imposes epistemically significant constraints on the beliefs that we can entertain" (ibid.). If there are beliefs that we cannot entertain, then there are reasons for acting that we cannot have. In situations where those reasons would be significant we should act sphexishly. Anticipating objections on this point, however, Fodor goes on to argue that epistemic boundedness is not unique to modularism, but in fact should be a consequence of any psychological theory. If Fodor is right, we can know a priori that we or any cognitive beings would be sphexish.

Fodor offers two sorts of reasons apart from modularity to expect the mind to be epistemically bounded. The first sort involves bounds on executional parameters: certain beliefs might be so complex that their mental tokens are unparseable, too large for short-term memory, or the like (p. 121). The more

interesting possibility is that the mind's basic conceptual repertoire might be limited. Empiricism -- nothing in the mind unless first in the senses -- poses such a restriction on the concepts available to the mind, and Fodor points out that empiricists like Hume are saved from epistemic boundedness only by verificationism, "a semantic thesis that has the convenient property of entailing that psychologically inaccessible beliefs are ipso facto nontruth-valuable [sic]" (p. 124).

Without some form of verificationist semantics or other retreats from realism, Fodor points out, significant conceptual constraints will imply epistemic boundedness. "All cognitive psychologies thus far proposed," Fodor declares, impose such constraints (p. 125). Nor is this likely to change. "The point," according to Fodor,

> is that any psychology must attribute some endogenous structure to the mind (really unstructured objects -- bricks, say -- don't have beliefs and desires and they don't learn things). And it's hard to see how, in the course of making such attributions of endogenous structure, the theory could fail to imply some constraints on the class of beliefs that the mind can entertain. (ibid.)

Part of the implication seems plausible enough. Fodor's claim a little later on that "as long as the class of accessible concepts is endogenously constrained, there will be thoughts that we are unequipped to think," is surely defensible, at any rate. (ibid.) But why must the attribution of an endogenous structure to the mind constrain the concepts available to it? Aside from a footnote to Aquinas' version of the Aristotelian argument (as paraphrased by Geach), Fodor gives us no further explanation. Perhaps it would help to look at the original.

III

The Aristotelian account seems unpromising at first. It depends on the communication of forms theory of sensory perception. On this theory, perception involves the transmission of a sensible form from the object to the sense organ (usually) via a medium. When George sees the red ball, both the ball and George's eye share a form; both the ball and the eye are literally, occurrently red. On a Lockean account, by contrast, George's sensation is the end of a causal chain beginning with the ball, and it is the only component of the chain that is occurrently red.

It is crucial for the Aristotelian account that the sense organ be potentially whatever quality it is to be sensitive to (De Anima 418a3). This requires that, prior to sensing, the organ actually possess none of those sensible qualities it is supposed to respond to. The eye must be in itself uncolored, the tongue neither moist nor dry, the flesh[2] neither hot nor cold, and so on.

Understanding is like perceiving for Aristotle, in that the communication of forms explains it as well. So, he infers, "as that which is capable of perceiving is to the objects of perception, so must be the intellect similarly to its objects" (429a13 - a17). Here Anaxagoras' point applies: the intellect, which can think all things, must be potentially all things (cf. 431b21), and so actually none of them. The intellect must potentially have all forms, and actually have none. The intellect, Aristotle proclaims, "must have no other nature than this, that it is potential" (429a20).[3]

Aristotle concludes that the intellect cannot be a material thing, a composite of matter and form, not because it cannot be material per se, but

because it can have no form. Thus the argument would tell equally against a dualism which differed from the view that the intellect is a material substance only in having it made of different stuff. Aristotle turns next to other related arguments that the intellect cannot be material, and there is a tendency in the medieval Aristotelians to focus on this aspect of Anaxagoras' insistence that the mind be "unmixed"; but it is clear that if Aristotle's argument shows anything, it shows that the intellect is no sort of substance, or at least no sort of intelligible substance, either material or immaterial.

So far, Fodor's argument is simply Aristotle's turned on its head. Fodor's claim that no psychology can escape epistemic boundedness, since any psychology must impose an endogenous structure on the mind, is not limited to materialist psychologies or structures, either. Also the possibility that the mind may possess an unintelligible form recalls Fodor's claim that, to the extent that cognition is not modular, it is probably inexplicable (Fodor, pp. 126 -129). There seems no great leap in equating 'forms' and 'structures', so that both Aristotle and Fodor wind up accepting the claim that if the intellect has an endogenous structure, then it is incapable of certain thoughts. They differ only in that Aristotle's <u>modus tollens</u> has become Fodor's <u>modus ponens</u>.

The difficulty in attributing Aristotle's argument for the Anaxagorean point to Fodor, of course, is that it is based on an archaic theory of perception, involving as it does the communication of forms, which neither Fodor nor any psychologist today would accept. Or would they?

On the standard reading, communication of forms theories of perception were displaced by representative theories sometime in the seventeenth century. The crucial difference was that representative theories do not require, as do communication theories, that the objects of sensation share the intrinsic

qualities of their sensations. Instead, the objects in themselves are supposed to be bereft of sensible forms such as colors, sounds, odors, tastes, and feels. Our perceptions represent their objects by modeling them through their internal relations, not by qualitiatively resembling the objects in their sensory contents. Perceptions capture the structural relations of their objects, not their intrinsic character.

Another way of putting all this, however, would be to say that it is the form and not the content that is communicated in perception. This would be misleadingly anachronistic, admittedly, since sensible forms just were the forms of sensory contents such as red. But it would be neither misleading nor anachronistic, I would argue, to view modern representative theories as differing from their predecessors in their reduction of qualitative to quantitative forms, and not in a rejection of the communication of forms. However this works out as history, there seems nothing in the claim that perception or understanding involves a transmission of forms, where 'form' is used in its current sense as a synonym for 'structure', to which Fodor or most psychologists could object.

Unfortunately, this version of the communication theory won't work in Aristotle's argument for the formlessness of the intellect. If the intellect's taking on the form of a tiger means only that it represents one, and not that it becomes one (albeit intentionally), then the formlessness of the understanding demanded by the communication theory means only that the intellect in itself represents nothing, and not that in itself it is nothing, i.e., that it is not a substance. It might even be a material substance, provided that a material substance can be a universal representer. Digital computers, as instantiations of Turing machines, seem to be such substances.

The point can be made by reference to Aristotle's famous comparison of the potential intellect to a blank tablet which potentially contains all inscriptions (429b29). Aristotle understood the analogy quite differently from Locke. Aristotle introduces the analogy only to explain what he means by the potential of the potential intellect. For Aristotle, the activation of this potential involves more than the inscription of the name of something on the blackboard of the intellect: the intellect must become another instance of the species understood, in the sense that it actually takes on its form. The possession of any form at all by the intellect would mean that it does not have the potential to take on that form, or forms incompatible with it, and so would limit the scope of understanding.

But on the modern view the potential intellect is just like the blank slate, a representational medium. The communication of forms now means the modification of the structure of the intellect to match that of the object of cognition. On Fodor's view this would involve tokenings in the language of thought. My point is that the ascription of an endogenous structure to a representational medium need not limit what it can represent. There of course will be what Fodor calls "boring parametric considerations" (p. 121): the size of the slate places an upper bound on the length of possible inscriptions, for example. But there is no reason to expect these parametric limitations to translate into categorical limitations in the content of what can be represented. For instance, the materiality of the mind need pose no barrier to its taking on either material or immaterial forms, if this means only that it adopts a certain representational structure, rather than that it actually becomes an immaterial thing.

The upshot then is that the Aristotelian argument for Fodor's claim that the ascription of any endogenous structure to the intellect would necessarily circumscribe the scope of its understanding does not work under a representative theory of mind. There may be other a priori arguments against the possibility of a universal representative medium, perhaps involving Godelian self-reference.[4] We might also inquire whether specific theses on the structure of minds -- materialism, computationalism, and the like -- entail conceptual constraints.

IV

In Elbow Room, Dennett proposes an argument for human sphexishness based on a computational view of mind that proceeds through something like epistemic boundedness. It goes like this (Dennett 1984, pp. 28 - 29):

1. Avoiding sphexishness requires responding to semantic features.
2. Physical systems can respond only to syntactical features.
3. Semantics is not determined by syntax.
4. Human minds are physical systems.
5. Therefore, human minds cannot avoid sphexishness.

None of the premises is uncontroversial, naturally, but I doubt whether there is a single understanding of 'syntax' and 'semantics' throughout the first three premises.

In the first premise, 'semantic' is used in the same broad sense as when Dennett says that "brains are meaning manipulators, information processors, or, as I shall say, semantic engines" (Dennett 1984, p. 28). Here it has no intrinsic connection to language, but could perhaps be replaced by 'evidentiary' or 'inferential'. The second premise is stated by Dennett thus: "physical mechanisms [are] syntactic engines, responding only to structural or formal

properties" (ibid.). Clearly again the sense of 'syntactic' is not connected to language, but how it is to be understood is unclear to me. The third premise, however, although a truism when both 'syntax' and 'semantics' are understood in the usual way as aspects of language, is not obvious when the terms are used in the extended senses of the first two premises One reason that syntax doesn't determine semantics, for example, is that context is important. But this distinction uses 'syntax' in a narrower sense than would be plausible in the second premise. If 'syntactic' means just physical, and 'semantic' evidentiary, on the other hand, it is not obvious that syntax doesn't determine semantics. Isn't the fact that 'smoke' means fire determined by physical features?

One reason a syntactic emulator would lag behind a semantic engine is quantitative. A finitely axiomatizable grammar will have only countably many expressions. But there are uncountably many truths, not to say meanings. (Consider real number theory.) So if my mind is a syntactic engine in this sense -- a finitely axiomatizable automatic formal system[5] -- then there will be uncountably many meanings it cannot capture, meanings which presumably would be directly apprehended by a semantic engine. But this is just epistemic boundedness, quod erat demonstrandum.

Not quite. A condition of epistemic boundedness was that the unthinkable thoughts be "epistemically significant" (Fodor, p. 120). Only if the inaccessible meanings were significant would their inaccessibility imply sphexishness, too. So both the epistemic boundedness and the sphexishness claims await some reason to think that the inaccessible meanings are not so due to their outrageous complexity or other "boring parametric considerations." So this is just another cardinality argument, which assumes that perfect

intelligence must be infinite, when surely finite intelligence is all finite creatures require.

* * * * *

There are undoubtedly an infinity of truths which lack expressions in Mentalese because they are too long or complex. Epistemic boundedness and cognitive sphexishness seem to be claiming more than this, however. It may well be that the endogenous structure of our minds makes us all Mentalese aphasics as dramatic, say, as the man who mistook his wife for his hat, and that only the universality of the phenomenon and the tendency of the mind to paper over its blind spots keeps the dread secret hidden. There seems little reason to think so a priori, however. At least it is hard for me to see how it is hard for Fodor to see how it could be otherwise.[6]

NOTES:

---

1. Fodor doesn't realize how close he is to Aquinas, because he reverses the direction of the Angelic Doctor's reasoning: he takes him to be arguing from the immateriality of the intellect to its universality, instead of the converse.

2. Or whatever is the proper organ of touch: see 422b34ff.

3. This actually describes only what came to be known as the potential intellect; in the following chapter, Aristotle notoriously proposes as well an intellect whose essence is pure activity.

4. Cf. Hofstadter pp. 534 - 36 on the relevance of the halting problem to the sphexishness issue.

5. Cf. Haugeland 1986 Chapter 2.

6. Read to the North Carolina Philosophical Society and to the Faculty Research Group at Davidson College. Thanks to Christopher Bell, Irwin Goldstein, Richard E. Grandy, Michael Resnik, and Richard Smyth for helpful comments.

BIBLIOGRAPHY

Cottingham, John et al. The Philosophical Writings of Descartes Vol I. Cambridge University Press, 1985.

Dennett, Daniel. Brainstorms, MIT Press, 1978.

Dennett, Daniel. Elbow Room. MIT Press, 1984.

Fodor, Jerry. The Modularity of Mind. MIT Press, 1983.

Hamlyn, D.W. Aristotle's De Anima, Books II, III. Oxford University Press, 1968.

Haugeland, John. Artificial Intelligence: The Very Idea. MIT Press, 1986.

Hofstadter, Douglas. Metamagical Themas. Basic Books, 1985.

Wooldridge, Dean. The Machinery of the Brain. McGraw Hill, 1963.

COUNTERFACTUAL CONDITIONALS AND THE CONJUNCTION FALLACY

John M. Miyamoto
Department of Psychology
University of Washington

Emily Dibble
Department of Psychology
University of Washington

INTRODUCTION

A counterfactual conditional is a statement of the form:

(1) If X were the case, then Y would be the case.

where X is a proposition that is not true in the current situation. The following are examples of counterfactual conditionals:

(2) If John F. Kennedy had not been assassinated, he would have run for the presidency in 1964.

(3) If John F. Kennedy had not been assassinated, U.S. involvement in Vietnam would have ended by 1968.

For brevity, we will refer to such statements as counterfactuals.

We are concerned here with the mental processes underlying judgments of whether a counterfactual is true or false. We will argue that intuitive verification of counterfactuals is based on judgments of representativeness like those described by Kahneman and Tversky (1972, 1973). In ways to be made precise, we propose that a counterfactual is judged to be true to the extent that the hypothetical situation described in the counterfactual is thought to be similar to the actual world as we know it. The idea that the truth of a counterfactual depends on the similarity of alternative, hypothetical situations to the actual world is not new. This notion is captured in a possible worlds semantic analysis of counterfactuals (Lewis, 1973; Stalnaker, 1968).

What is new, however, is that we have found systematic fallacies in the intuitive verification of counterfactuals that are analogous to representativeness errors in probabilistic reasoning. Experimental evidence will be presented, showing that intuitive judgments of the truth of counterfactuals systematically violate implications of the possible worlds semantics for counterfactuals. To explain our results, we first discuss possible worlds semantics for counterfactuals. Next we describe conjunction fallacies in probabilistic reasoning discovered by Tversky and Kahneman (1983), and suggest how analogous fallacies might be found in counterfactual reasoning. We then present experimental evidence demonstrating conjunction fallacies in counterfactual reasoning. Our results also establish the existence of disjunction fallacies in counterfactual reasoning.

STALNAKER'S POSSIBLE WORLDS SEMANTICS FOR COUNTERFACTUALS

Philosophers have long recognized that counterfactual inference is fundamental to scientific epistemology (Chisholm, 1946; Goodman, 1955). Here, we cannot discuss the general role of counterfactual inference in epistemology, but will restrict our discussion to a semantic analysis of counterfactuals developed by Stalnaker and Thomason (Stalnaker, 1968; Stalnaker & Thomason, 1970; Thomason, 1970). Our purpose is to derive certain semantic relations among counterfactuals that are subject to experimental test. These relations can also be derived in Lewis's (1973) semantics for counterfactuals, but his analysis contains complexities whose discussion must be omitted for the sake of brevity.

Stalnaker (1968) proposed a logical analysis of counterfactuals using possible worlds semantics as developed by Kripke (1963). In Kripke's semantics, propositions are not simply true or false; rather they are true or false relative to a possible world. A proposition might be true in some worlds, and false in others. Stalnaker's analysis assumes that possible worlds are related by degrees of similarity. Furthermore, he makes the strong assumption that for any world i and proposition X, if there exists any world in which X is true, then there exists a unique world that is most similar to i in which X is true (X is false in every world if it is self-contradictory).

Let X ==> Y symbolize a counterfactual of the form (1). Stalnaker proposed that X ==> Y is true at the world i if and only if either (a) there is no world at which X is true, or (b) Y is true in the world k, where k is the unique world such that X is true in k and k is most similar to i. Clause (a) of this truth condition merely excludes trivial cases, e.g., "If 2 + 2 = 3, then the national debt would be eliminated" is true because there are no worlds in which 2 + 2 = 3. Clause (b) of the truth condition is the heart of Stalnaker's analysis. To evaluate the truth of X ==> Y at the world i, we find the most similar world k where X is true. If Y is true in k, then X ==> Y is true; if Y is false in k, then X ==> Y is false. For example, to decide whether statement (2) is true relative to the actual world, we must check whether John F. Kennedy ran for president in 1964 in the most similar world where he was not assassinated. If he ran, then (2) is true; otherwise (2) is false.

Let Y AND Z and Y OR Z denote, respectively, the truth functional conjunction and disjunction of Y and Z. In Stalnaker's analysis, truth functional implications are logically valid. Thus if Y AND Z is true in world i, then Y must be true in i and Z must be true in i. Similarly, if Y OR Z is true in i, then Y must be true in i, or Z must be true in i, or both. This leads to two critical implications of Stalnaker's theory. First, if X ==> Y AND Z is true at world i, then X ==> Y and X ==> Z must also be true at world i. Second, if X ==> Y OR Z is true at world i, then either X ==> Y

or X ==> Z must be true at world i. For example, consider the following statements:

(4) If John F. Kennedy had not been assassinated, he would have run for the presidency in 1964, and U.S. involvement in Vietnam would have ended by 1968.

(5) If John F. Kennedy had not been assassinated, he would have run for the presidency in 1964, or U.S. involvement in Vietnam would have ended by 1968.

Statements (2)-(5) have the respective forms, X ==> Y, X ==> Z, X ==> Y AND Z, and X ==> Y OR Z. Let k denote the most similar world where John F. Kennedy was not assassinated. Stalnaker's analysis claims that (4) is true if and only if in the world k, John F. Kennedy ran for the presidency in 1964 and U.S. involvement in Vietnam ended by 1968. Thus, if (4) is true, (2) is true because John F. Kennedy ran for the presidency in k, and (3) is true because U.S. involvement in Vietnam ended by 1968 in k. Similarly, if (5) is true, then in the world k, either Kennedy ran for president in 1964, or U.S. involvement in Vietnam ended by 1968, or both. Therefore, if (5) is true, either (2) is true or (3) is true, or both. These examples suggest the following:

(6) If a counterfactual with a conjunctive consequent is true, then the counterfactuals with either clause of the conjunction as consequent are also true.

(7) If a counterfactual with a disjunctive consequent is true, then the counterfactuals with either clause of the disjunction as consequents cannot both be false.

It is easy to show that (6) and (7) are implied by Stalnaker's theory, and also by Lewis's (1973) generalization of Stalnaker's theory. Our experimental results suggest that (6) and (7) are both violated in intuitive reasoning.

CONJUNCTION FALLACIES IN PROBABILISTIC REASONING

Tversky and Kahneman (1983) have shown that intuitive judgments of probability systematically violate the principle that the probability of a conjunction of events can never exceed the probabilities of its constituent events. Let P(Y AND Z), P(Y) and P(Z) denote the probabilities of Y AND Z, Y, and Z, respectively. When P(Y AND Z) is judged to be greater than either P(Y) or P(Z), the judgment is said to constitute a conjunction fallacy. Tversky and Kahneman found that if Y is a highly representative outcome, and Z is mildly unrepresentative, then P(Y AND Z) will often be judged greater than P(Z).

For example, suppose that Linda is described to be 31 years old, single, and deeply concerned about social issues. A representative outcome would be the outcome Y that Linda is active in the feminist movement, and an unrepresentative outcome would be the outcome Z that Linda is a bank teller. By the

laws of probability, it is more likely that Linda is a bank teller than that she is both a bank teller and active in the feminist movement. Tversky and Kahneman (1983) found that 85% of 142 subjects judged Y AND Z to be more probable than Z alone. Similar examples of conjunction fallacies have been elicited in many contexts, using a variety of stimulus materials and modes of response (Tversky & Kahneman, 1983; Leddo et al., 1984).

Why do conjunction fallacies occur? Kahneman and Tversky have argued that the judged probability of events is often determined by the representativeness of the events (Kahneman & Tversky, 1972, 1973; Tversky & Kahneman, 1971, 1982). Although the concept of representativeness itself has several aspects to it (see Tversky & Kahneman, 1982), for present purposes representativeness can be construed as the degree of similarity of an instance to typical members of a class or category. For example, the description of Linda is similar to that of a prototypical feminist, hence being active in the feminist movement is representative of Linda. Linda's description is not particularly similar to that of a bank teller, thus working as a bank teller is mildly unrepresentative of Linda. An independent sample of subjects judged Linda to be most similar to an active feminist, least similar to a bank teller, and of intermediate similarity to a bank teller who is active in the feminist movement (Tversky & Kahneman, 1983). Note that the ranking of outcomes by judged similarity coincides with the rankings of the outcomes by judged probability, as predicted by the hypothesis that probability judgments are based on the representativeness of outcomes.

As demonstrated by the Linda example, an instance (Linda) can be more similar to a conjunction of categories (bank teller and active feminist), than to one of the categories alone. Tversky (1977) has proposed a theory of similarity according to which the similarity of objects is an increasing function of the features shared by the objects, and a decreasing function of the features that distinguish the objects. If the category Y is representative of an instance and Z is not, the conjunction Y AND Z will share more features with the instance than Z alone. Thus, an instance will be more similar to a conjunction of representative and unrepresentative categories, than to the unrepresentative category alone. Tversky and Kahneman explain the occurrence of conjunction fallacies by combining the representativeness hypothesis with Tversky's theory of similarity. Events will be regarded as more probable if they are more representative. A conjunction of events will appear more representative of an instance than one of the events in the conjunction if the conjunction shares more features with the instance than does the single event. When this occurs, the conjunction of events will appear more probable than the single event.

CONJUNCTION FALLACIES IN COUNTERFACTUAL REASONING

We propose that representativeness also plays a fundamental role in counterfactual reasoning. In judging whether X ==> Y is true, we postulate that the individual evaluates whether X AND Y or X AND NOT Y is more representative of the actual situation. If X AND Y is much more representative, X ==> Y will be judged to be true. If X AND NOT Y is much more representative, X ==> NOT Y will be judged to be true. Finally, if neither X AND Y nor X AND NOT Y is clearly more representative, then neither counterfactual will be judged to be true, but the counterpossible ("If X were the case, then Y might be the case") will be judged to be true.

If intuitive verification of counterfactuals is based on representativeness, conjunction fallacies in counterfactual reasoning should be easy to construct. Consider statements (2)-(5) above. If John F. Kennedy had not been assassinated (X), then his running for president in 1964 (Y) is a representative outcome, but the ending of U.S. involvement in Vietnam by 1968 (Z) is rather unrepresentative. We conjecture that the combination X AND Y is more representative than X AND Y AND Z, which is more representative than X AND Z. Hence the counterfactual (2) should be more plausible than (4), which should be more plausible than (3). Symbolically, X ==> Y should be more plausible than X ==> Y AND Z, which should be more plausible than X ==> Z.

The Stalnaker/Lewis semantics implies that in no instance can X ==> Y AND Z be true if X ==> Z is false. Here we will present evidence that violations of this principle can be found in the judged degree of truth of counterfactuals. In our experiment, subjects were presented with a story which they were told to regard as reliable information, and then were asked to rate the degree of truth of counterfactuals pertaining to the story. The counterfactuals were chosen as matched sets of the form:

X ==> Y	(Representative statement)
X ==> Z	(Neutral statement)
X ==> Y AND Z	(Conjunction statement)
X ==> Y OR Z	(Disjunction statement)

We chose propositions X, Y and Z such that in the context of the story, Y would have been a representative outcome, and Z would have been neither clearly representative nor unrepresentative if X had been the case. For any matched set, Y will be called the representative statement and Z the neutral statement of that set. Our hypothesis is that a subject will rate X ==> Y AND Z as more true than X ==> Z. If this pattern of ratings is observed, we will say that the subject has committed a conjunction fallacy.

METHOD

Subjects. Subjects were 280 undergraduates at the University of Washington who participated in the experiment for credit in a psychology course. Subjects were run in small groups in sessions lasting about 15 minutes. Eight subjects were dropped from the experiment because they failed to follow instructions.

Materials. Two stories, "Donald" and "Harper City," were written for the experiment. For each story, two matched sets of four related counterfactuals were prepared, along with two anchoring statements. One anchoring statement was obviously true and the other was obviously false. When subjects were asked to rate the truth of counterfactuals, the anchoring statements were presented first in order to elicit extreme judgments that would mitigate end effects in the rating response (cf. Anderson, 1982). An abbreviated version of one story and a matched set of related counterfactuals is presented below.

> Harper City is a port town on the Gulf Coast that has enjoyed a booming economy because of productive oil wells and prosperous fishing industry. Local tax revenues created a large surplus in the city treasury that the citizens targeted for city improvement. Because interest in sports was intense and widespread, a coalition of citizens and business interests lobbied for the construction of a football stadium, with the intention of attracting a professional football franchise to the city. There was also some support for replacing the public library with a new facility. Unfortunately, Harper City was struck by a hurricane last year. The damage was severe. The cost of cleaning up the city and replacing buildings that had been destroyed, including the library, was so great that the people of Harper City were forced to drop all plans for civic improvement.

R1. If the hurricane had not struck, the people of Harper City would have decided to build a professional football stadium. (Representative)

N1. If the hurricane had not struck, the people of Harper City would have decided to build a new public library. (Neutral)

C1. If the hurricane had not struck, the people of Harper City would have decided to build a professional football stadium and a new public library. (Conjunction)

D1. If the hurricane had not struck, the people of Harper City would have decided to build a professional football stadium or a new public library. (Disjunction)

Each subject received one story and rated the truth of the statements associated with the story. All subjects rated the anchoring statements, two representative statements, two neutral statements, and two conjunction statements. A subset of subjects also rated the two disjunction statements. The order of the statements was varied twelve ways for each story, with anchoring

statements preceding all other statements. Each statement was rated by marking a slash through a horizontal response line, labeled "virtually impossible," "somewhat unlikely," "somewhat likely," and "virtually certain," from left to right. Responses were coded by dividing the response line into 20 equal intervals, with 1 denoting the greatest certainty that the statement was false, and 20 the greatest certainty that the statement was true.

We should mention a typographical error in one disjunction statement of the "Donald" story. Donald is a high school football player who broke his leg before the season and thus could not play. The counterfactual antecedent X asks what would have happened if he had been able to play football. The representative outcome Y is that he would have been captain of the football team. The neutral outcome Z is that his grade point average would have been 3.4. The representative, neutral and conjunction statements have the respective forms X ==> Y, X ==> Z, and X ==> Y AND Z. Unfortunately, the disjunction statement had the form X ==> Y OR Q, where Q asserts that Donald's grade point average would have been 3.5. This typographical error does not affect the comparisons of conjunction statement to neutral statement, or of the disjunction statement to representative statement. The Stalnaker/Lewis semantics predicts that X ==> Y AND Z should receive a lower rating than X ==> Z, and that X ==> Y OR Q should receive a higher rating than X ==> Y. We predict that the opposite will occur, based on considerations of representativeness.

Design and Procedure. Subjects were informed that the study concerned reasoning about hypothetical events. Subjects read a story which they were instructed to regard as reliable information from a news magazine, and were asked to rate the truth of the counterfactuals associated with the story. The use of the response line was explained, the anchoring statements were identified, and subjects were instructed to give the true anchor the highest rating, and the false anchor the lowest rating among all statements. A total of 133 subjects rated representative, neutral, and conjunction statements based on "Donald", with 57 of these subjects also rating disjunction statements. Representative, neutral, and conjunction statements based on "Harper City" were rated by 139 subjects, with 56 subjects also rating disjunction statements. All dependent variables were within subject variables.

RESULTS AND DISCUSSION

In the following analysis, the term conjunction fallacy will be reserved for cases where the rating of a conjunction statement is greater than the rating of a corresponding neutral statement. This operational definition of a conjunction fallacy is conservative, for it underestimates the true prevalence of conjunction fallacies (it omits cases where the conjunction statement is rated higher than the representative statement, but not the neutral state-

TABLE 1
Median Ratings of Statements: All Subjects

	Harper City N = 139		Donald N = 133	
	Set 1	Set 2	Set 3	Set 4
Representative	15.4	15.9	16.1	14.1
Neutral	10.6	10.3	9.4	7.0
Conjunction	12.5	13.2	11.4	7.9

TABLE 2
Conjunction Statements versus Neutral Statements: All Subjects

	N	Ties	% Conjunction Fallacies	Signed Ranks Test z score	p value
Set 1	139	16	67%	-4.148	.000
Set 2	139	20	71%	-6.168	.000
Set 3	133	20	56%	-1.145	.252
Set 4	133	21	64%	-2.236	.025

ment). The adoption of this conservative policy is suggested by statistical issues which cannot be discussed here*.

Table 1 presents median ratings of conjunction, representative and neutral statements. Note that in every case, the conjunction statement receives a higher median rating than the corresponding neutral statement. Three of the four median ratings were significantly higher for the conjunction statement ($p < .05$, two-tailed Wilcoxon signed ranks test). Table 2 lists z scores for the signed ranks tests, and the percentage of subjects who rated conjunction statements over corresponding neutral statements. In every case, the majority of the subjects produced conjunction fallacies. The three largest percentages are significantly greater than .5 ($p < .01$, two-tailed sign test).

These findings conclusively demonstrate the existence of conjunction fallacies, for even if some conjunction fallacies were due to random variation in ratings, one would not expect the median rating of conjunctions to exceed the median rating of neutral statements, nor would one expect the probability of conjunction fallacies to exceed .5. It might be argued that subjects have misinterpreted "and" as "or," i.e., interpreted the conjunction statements as

* We wish to thank Bob Frick for useful criticism of our statistical analysis, and Elizabeth Moore for advice on computation.

TABLE 3
Median Ratings of Statements:
Subjects Who Rated Disjunction Statements

	Harper City N = 56		Donald N = 57	
	Set 1	Set 2	Set 3	Set 4
Representative	16.2	15.5	16.9	16.0
Neutral	10.0	8.5	8.3	6.7
Conjunction	13.5	13.5	12.8	8.2
Disjunction	15.5	15.5	14.6	8.0

TABLE 4
Conjunction Statements versus Disjunction Statements[a]:
Subjects Who Rated Disjunction Statements

	N	Ties	Signed Ranks Test z score	p value
Set 1	56	14	-2.882	.004
Set 2	56	7	-3.293	.001
Set 3	57	5	-4.107	.000

[a] The conjunction and disjunction statements of set 4 are not directly comparable due to a typographical error. (See methods section.)

disjunction statements. Tables 3 and 4 show that three of four disjunction statements received significantly higher median ratings than corresponding conjunction statements ($p < .005$, signed ranks test).

The Stalnaker/Lewis semantics also predicts that a disjunction statement will receive an equal or higher rating than the corresponding representative statement. Contrary to this prediction, three of four disjunction statements received lower median ratings than the corresponding representative statements (Tables 3 and 5). Two of the four differences in median rating were highly significant ($p < .005$, two-tailed signed ranks test). It should be noted that the statistical test is conservative in that it tests the null hypothesis of no difference in median ratings. This hypothesis essentially claims that subjects do not distinguish disjunction statements from representative statements. Under the more plausible assumption that subjects do distinguish these statements, the Stalnaker/Lewis semantics would predict that disjunction statements would receive higher median ratings than representative statements. Thus, even in the two cases where the difference in median ratings were not significantly different, the percentage of disjunction fallacies was close to

TABLE 5
Disjunction Statements versus Representative Statements:
Subjects Who Rated Disjunction Statements

	N	Ties	% Disjunction Fallacies	Signed Ranks Test z score	p value
Set 1	56	10	52%	-1.903	.275
Set 2	56	9	49%	-0.825	.409
Set 3	57	8	69%	-2.860	.004
Set 4	57	2	98%	-6.422	.000

50%, a much higher percentage than one would expect if the fallacies were due only to random variation in rating responses.

To develop a representativeness analysis of disjunction fallacies, one needs an account of the representativeness of disjunctions. Tversky (1977) proposed that that the similarity of objects A and B increases with the number of features common to A and B, and decreases with the number of features that distinguish A from B, and B from A. Although the situation described by a disjunction statement would seem to have more features in common with the actual world, it also has more distinguishing features. Hence, whether disjunction fallacies occur may depend on the balance of the common and distinguishing features. Although disjunction fallacies have not been demonstrated in probabilistic reasoning, our results for counterfactuals suggest that similar results can also be obtained for probability judgment. If this is so, a representativeness analysis of disjunctive events will be required whether or not one accepts the extension of representativeness to the analysis of counterfactuals.

CONCLUSION

The Stalnaker/Lewis semantics fails to describe the intuitive verification of counterfactuals because truth functional consequences are valid in any possible world, and conjunction and disjunction fallacies violate truth functional implications. We propose that intuitive verification of counterfactuals is based on a judgment of relative representativeness. The individual compares the representativeness of a situation or scenario where the antecedent and consequent of the counterfactual are true, to the representativeness of a situation where the antecedent is true and the consequent is false. Conjunction fallacies occur when the antecedent and conjunctive consequent are more representative than the antecedent and one clause of the conjunction. Disjunction fallacies occur when the antecedent and disjunctive consequent are less representative than the antecedent and one clause of the disjunction.

One way to contrast the two analyses is to consider what entities are claimed to enter into similarity relations. The Stalnaker/Lewis semantics proposes that degree of similarity applies to possible worlds, i.e., entities that completely determine the truth or falsity of every proposition. Our analysis proposes that degree of similarity applies to sets of features that characterize situations or scenarios. Tversky's (1977) similarity model suggests how a situation where a conjunction is true might be more similar to the actual world than a situation where a single clause of the conjunction is true. This is paradoxical from the possible worlds standpoint, for the most similar world where a conjunction is true could never be more similar than every world where one clause of the conjunction is true. Our results thus suggest that the mental representation of similarity, and the entities to which similarity applies are rather different from structures assumed in the possible worlds semantics.

A common objection to studies attempting to demonstrate logical errors in human reasoning is the claim that the subjects may have failed to interpret the questions in the manner intended by the experimenter (Henle, 1962). The study reported above contains no control for the possibility that X ==> Z was interpreted as X ==> Z AND NOT Y, a circumstance which could yield apparent conjunction fallacies without logical error. Tversky and Kahneman (1983) and Pennington (1984) have conducted studies of probabilistic reasoning which attempt to exclude this interpretation. In one study, ratings of conjunctions were elicited from one group, and ratings of neutral statements from another group. The between subjects design also yields conjunction fallacies. Another study explicitly asked subjects to evaluate the probability of Z whether or not Y was true. Again, conjunction fallacies were prevalent.

Counterfactual conjunction and disjunction fallacies are symptomatic of the difference between truth in a possible world as the logician understands it, and realization in the mental construction of a situation or scenario. Mental constructions are affected by representational and processing constraints that are usually ignored in model theoretic semantics. When we better understand the factors influencing judgments of the truth of counterfactuals, it may be possible to use the cognitive theory of counterfactual judgment to elucidate causal and dispositional conceptions that appear to depend on counterfactual inference.

REFERENCES

Anderson, N. H. (1982). Methods of information integration theory. New York: Academic Press.

Chisholm, R. M. (1947). The contrary-to-fact conditional. Mind, 56, 289-307.

Goodman, N. (1955). The problem of counterfactual conditionals. In N. Goodman, Fact fiction and forecast. New York: Bobbs-Merrill.

Henle, M. (1962). On the relation between logic and thinking. Psychological Review, 69, 366-378.

Kahneman, D. & Tversky, A. (1972). Subjective probability: A judgment of representativeness. Cognitive Psychology, 3, 430-454.

Kahneman, D. & Tversky, A. (1973). On the psychology of prediction. Psychological Review, 80, 237-251.

Kripke, S. A. (1963). Semantical analysis of modal logic 1. Normal modal propositional calculi. Zeitschrift fur Mathematische Logik und Grundlagen der Mathemathik, 9, 67-96.

Leddo, J., Abelson, R. P., & Gross, P. (1984). Conjunctive explanations: When two reasons are better than one. Journal of Personality and Social Psychology, 47, 933-943.

Lewis, D. (1973). Counterfactuals. Oxford, UK: Blackwell.

Pennington, N. (1984). Technical note on conjunctive explanations. Center for Decision Research, Graduate School of Business, The University of Chicago.

Stalnaker, R. C. (1968). A theory of conditionals. In N. Rescher (Ed.), Studies in logical theory. Oxford, UK: Blackwell.

Stalnaker, R. C., & Thomason, R. H. (1970). A semantic analysis of conditional logic. Theoria, 36, 23-42.

Thomason, R. H. (1970). A Fitch-style formulation of conditional logic. Logique et Analyse, (New Series Vol. 13), 397-412.

Tversky, A. (1977). Features of similarity. Psychological Review, 84, 327-352.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. Psychological Bulletin, 2, 105-110.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185, 1124-1131.

Tversky, A., & Kahneman, D. (1982). Judgments of and by representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), Judgment under uncertainty: Heuristics and biases. Cambridge, UK: Cambridge University Press.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. Psychological Review, 90, 293-315.

Should we use Probability in Uncertain Inference Systems?

Max Henrion

Department of Engineering and Public Policy,
and Department of Social and Decision Science,
Carnegie Mellon University.

Abstract

Criticisms of probability as being epistemologically inadequate as a basis for reasoning under uncertainty in AI and rule-based expert systems are largely misplaced. Probabilistic schemes appear to be the best way to deal with dependent evidence, and to properly combine diagnostic and predictive inference. Suggestions that expert systems should duplicate human inference strategies, with their documented biases, seem ill-advised. There is evidence that popular schemes perform quite poorly under some circumstances and there is an urgent need for careful study of when they can be relied upon. Some promising probabilistic alternatives are available, but they need to be demonstrated in realistic applications.

Introduction

Historically, probability has been by far the most widely used formalism for quantifying uncertainty and making inferences about it. However, for various conceptual and pragmatic reasons the majority of AI researchers have not, hitherto, found standard probabilistic techniques very appealing for use in rule-based, expert systems. Among the many alternatives they have used are the Certainty Factors used in Mycin (Shortliffe & Buchanan, 1975) and its descendants, Fuzzy Set Theory (Zadeh, 1984), the quasi-probabilistic scheme of Prospector (Duda et al, 1976), the Belief functions of Dempster-Shafer theory (Shafer, 1976), Paul Cohen's theory of endorsements (Cohen, 1985), Doyle's theory of reasoned assumptions (Doyle, 1983), and non-numerical, linguistic representations of uncertainty (Fox, 1986). We shall refer to both probabilistic and alternative methods, generically, as *uncertain inference schemes*, or UISs.

Each of these techniques has its partisans and its detractors, and discussion about their various merits and flaws seems to be heating up of late (Kanal & Lemmer, 1986, Gale, 1986). Much of the discussion hitherto has focussed on the theoretical issues of *epistemological adequacy* -- how well can each UIS represent the different aspects of reasoning with uncertainty? System developers have understandably been more concerned with the pragmatic issues of *heuristic adequacy* -- how easy is it to use and what are its computational demands?

The purpose of this paper is to give a personal view of some of the basic issues in evaluating and comparing UISs. Following a very brief account of the subjective or personalist view of probability and its applications in AI, I shall summarize the most common objections to probabilistic schemes, and attempts to rebut them that have appeared in the literature. The main focus will be on four issues that seem to have attracted less attention. The first concerns the treatment of correlated sources of evidence, and assumptions about dependence. The second is the issue of combining diagnostic and predictive reasoning, and the separation of inference rules from domain knowlege. The third is the vexed question of whether or not UISs for expert systems should try to to approximate human reasoning. The last is the question of whether it matters much which approach you use, and I shall argue the importance of more systematic comparison of alternative UISs to find out. The

issue is not simply what can each UIS do or not do, but how much practical difference to the conclusions can it make which you use?

There has been considerable controversy on several of these topics. Researchers are operating under different paradigms with different programmatic goals, so I cannot expect agreement with all my arguments. What I hope is that it will be a contribution towards creating a more focussed debate, as a prerequisite for more cumulative science in this important area.

The appeal of probability

The probability of a proposition or a future event, according to the Bayesian or personalist view, is a measure of a person's degree of belief in it, given the information currently known to that person. The notion of probability may be derived from a set of simple axioms of rational decision-making under uncertainty, which form the basis of decision theory (Savage, 1954). The force of these axioms, and hence of the laws of probability derived from them, arises from the fact that a people who violate them and are willing to act on "incoherent" probabilities (for example, which do not satisfy Bayes' rule) are liable to demonstrable loss. Notably, an opponent could always design a "Dutch book", that is a combination of bets that they would be willing to accept, according to their professed beliefs, but which, in sum, would result in a guaranteed loss (de Finetti, 1974).

One advantage of being embedded in a theory of decision making is that it provides an *operational definition* for the probability of an event, in terms of the person's willingness to take bets based on the outcome of the event. Secondly, in combination with a utility model of preferences, it provides a clear, axiomatically-based approach for making decisions under uncertainty (Holtzman & Breese, 1985). Thirdly, it provides well-defined ways of using empirical data (Spiegelhalter, 1986), and evaluating the accuracy, resolution and calibration of UISs (Lichtenstein, Fischhoff & Phillips, 1982). No non-probabilistic measure of uncertainty offers these advantages. It has also been shown that for any reasonable scoring rule (which rewards a decision maker based on his or her probability assessments and the actual outcomes or truth of the propositions), any scalar measure of uncertainty is either worse than probability (produces a expected lower score) or is equivalent to it (Lindley, 1982).

Probabilistic UISs

A set of m propositions, $\{A_1, A_2, \ldots A_m\}$, each of which may be true or false, gives rise to 2^m *different possible elementary events*, each being a particular combination of proposition values. e.g. $(A_1 \ \& \ \sim A_2 \ \& \ \ldots \ A_m)$. A complete joint probability distribution over these propositions specifies a probability for each event, and so requires specification of 2^m-1 parameters. The exponential complexity of this *complete* representation clearly rules it out as a viable approach for practical systems and so simplifying assumptions are essential. In all practical UISs, the evidential relationships are modelled as an *inference network*, in which each proposition (or variable) is directly related to only a few others. Each such link is represented by a *rule*, which provides evidence about a consequent proposition, C, based on the degree of belief in some logical combination of its *antecedents*, A_i, for example, $(A_1 \ \& \ \sim A_2) \dashrightarrow C_3$. A "strength" (one or more numbers) is associated with each rule, whose probabilistic interpretation varies according to the scheme. Each such UIS needs to provide functions for propagating the uncertainty measures through logical conjunction, disjunction, negation, and generalized *modus ponens*, as well as a function for combining the evidence from multiple rules that bear on one consequent. The best known schemes are Certainty Factors (CFs) developed for Mycin (Shortliffe & Buchanan, 1975) and the scheme used in Prospector (Duda et al, 1976), from which many variants have been derived. Both of these were originally intended as approximations to Bayesian inference.

Neither system is completely consistent with a complete probabilistic scheme, and the implied probability distributions are incoherent. Kim and Pearl have devised an ingenious scheme for representing and propagating probabilistic information over a kind of inference net they term *Bayes' networks* (Kim & Pearl, 1983). This can maintain global coherence over the network,

using only an efficient local updating mechanism. These are similar in spirit to the *influence diagrams* developed for decision analysis (Shachter, 1985).

An alternative approach to dealing with a partially specified probability distribution is to estimate the full distribution using the *Maximum Entropy Principle*. This minimizes the additional information assumed in filling out the distribution, consistent with the specified constraints (usually marginal and conditional probabilities). While this approach has several desirable properties (Shore and Johnson, 1980), computation of maximum entropy distributions is, in general, prohibitively expensive with more than a few propositions, despite attempts to improve algorithms (Cheeseman, 1983). However, many popular probabilistic updating schemes, including , conditional independence assumption, Jeffrey's rule, and odds-ratio updating, are actually special cases of Maximum Entropy, and the related Minimum Cross Entropy update (Wise, 1986).

Objections to probability

Despite its attractions, probability has been under sustained attack as a viable scheme for representing uncertainty in AI, ever since McCarthy and Hayes dismissed it as "epistemologically inadequate". Among the criticisms have been the following:

1. Probability requires vast amounts of data or unreasonable numbers of expert judgments.
2. It can't express ignorance, vagueness or "second-order uncertainty".
3. It doesn't distinguish reasons for and against, or identify sources of uncertainty.
4. The inference process is hard to explain.
5. It can't express linguistic imprecision.
6. It requires unrealistic independence assumptions.
7. It is computationally intractable.
8. It is not how humans reason.
9. It doesn't make much difference what method you use anyway.

Several recent articles have assembled similar lists of objections, overlapping with the first five or six listed here; they have provided eloquent rebuttals (Spiegelhalter, 1986, Pearl, 1985a, Cheeseman, 1985). Below is an extremely brief summary of their conclusions, without attempt at explanation. The interested reader is referred to the original articles. The main focus of this article will be objections six to nine and some related issues of the *heuristic adequacy* of probabilistic schemes.

Summary of rebuttals

In evaluating the criticisms and rebuttals, it is important to distinguish claims about probabilistic inference in general from claims about specific quasi-probabilistic UISs incorporating various heuristic assumptions. These rebuttals have been primarily in defense of the theoretical possibilities of probability rather than particular UISs. Failure to keep in mind this distinction has sometimes resulted in misunderstanding and fruitless argument.

The belief that probabilistic representations require vast amounts of data seems to derive from frequentist interpretations of probability, and does not apply to the Bayesian or subjectivist interpretations usually advocated. Inordinate quantities of subjective judgments should not be necessary either, since humans are subject to analagous limitations to other UISs, and our intuitive knowledge of probabilistic dependencies is represented by relatively sparse networks (Bayes networks), where most variables are not directly dependent. The question of whether or not two variables are directly probabilistically dependent is a qualitative judgment which is relatively easy to make (Pearl, 1985a).

Ignorance, vagueness or second order uncertainty may be represented by a range of probabilities, or by a predictive distribution over a probability, expressing how the prior probability might change after consulting a specified information source (Cheeseman, 1985, Spiegelhalter, 1986). Although it is often sufficient to represent each probability by its

mean value, unless decisions about gathering new information are being contemplated.

It is true that a single probability by itself doesn't distinguish the sources, type and effect of the pieces of evidence on which it is based, but it is certainly possible to retrieve and clearly express this information in probabilistic schemes (Pearl, 1985a). For example, the *evidence weight* (log likelihood ratio) provides a convenient additive measure of the relative importance of each piece of evidence for a conclusion. The weights of supporting evidence can be added to the prior weight, and weights of disconfirming evidence subtracted in a sort of "ledger sheet" to arrive at the total final weight (Spiegelhalter, 1986). Evidence weights are also useful in explaining probabilistic reasoning. As long as the underlying inference network is sparse, as Pearl argues it will be, the inference process should be explainable in simple, comprehensible steps (Pearl, 1985a).

The advantage often claimed for Fuzzy Set Theory over probability is that the former can model linguistic imprecision, whereas probabilities are only defined for unambiguously specified ("crisp") events or propositions (Bonissone, 1982). Indeed probabilists have generally not addressed the issue of linguistic imprecision, aside from studies of the correspondence between probability phrases and numbers (Beyth-Marom, 1982). There is plenty of experimental evidence that probabilistic inference is not a very good model for human linguistic reasoning (Kahneman, Slovic & Tversky, 1982). But there has been little experimental investigation of claims that alternative UISs offer better models. A study comparing human judgment to Fuzzy Set Theory found that subjects' judgment of the "plausibility" of the intersection of two fuzzy sets was better modelled by the multiplication of "plausibilities", analogous to the probabilistic rule for the intersection of two independent events, rather than by the minimum plausibility rule of Fuzzy Set Theory (Oden, 1977). A problem in such studies of non-probabilistic schemes is setting up a convincing comparison when the measure of uncertainty has no operational definition.

Assumptions about dependence

While probability can in theory cope perfectly with non-independent sources of evidence, most actual UISs cannot. Consider the following:

> **Chernobyl example:** *The first radio news bulletin you hear on the accident at the Chernobyl nuclear power plant reports that the release of radioactive materials may have already killed several thousand people. Initially you place small credence in this, but as you start to hear similar reports from other radio and TV stations, and in the newspapers, you believe it more strongly. A couple of days later, you realize that the news reports were all based on the same wire-service report based on a single unconfirmed telephone interview from Moscow. Consequently, you greatly reduce your degree of belief again.*



Figure 1: Inference network for Chernobyl example

This illustrates how multiple, independent supporting sources of evidence increase the confirmation of a hypothesis, but the confirmation is reduced if they are correlated. Most of us seem quite capable of handling this kind of intuitive reasoning in practice, even if we don't have the terminology to describe it. However none of the better known UISs are actually capable of distinguishing between independent and correlated sources of evidence. They each make various arbitrary fixed assumptions about the presence or absence of dependence. So all are inherently incapable of performing this normal commonsense reasoning. For example, the Fuzzy Set operators for *and* and *or*, are equivalent in effect to probabilistic rules assuming *subsumption* among antecedents, i.e. where the least likely proposition logically implies the more likely one(s). This is equivalent to assuming the *maximum possible* correlation between input propositions. Prospector and

Mycin CFs use similar rules for *and* and *or*. On the other hand, Prospector, and Bayes networks assume *conditional independence* when combining evidence from different rules, as in the Chernobyl example. Figure 2 shows *Pr(A&B)* as a function of *Pr(A)* given *Pr(B)=0.6*, assuming minimum overlap between *A* and *B* (Minc), independence (Ind), or maximum overlap (MaxC), which is the Fuzzy Set assumption. It illustrates the range of results possible from alternative assumptions about correlation.

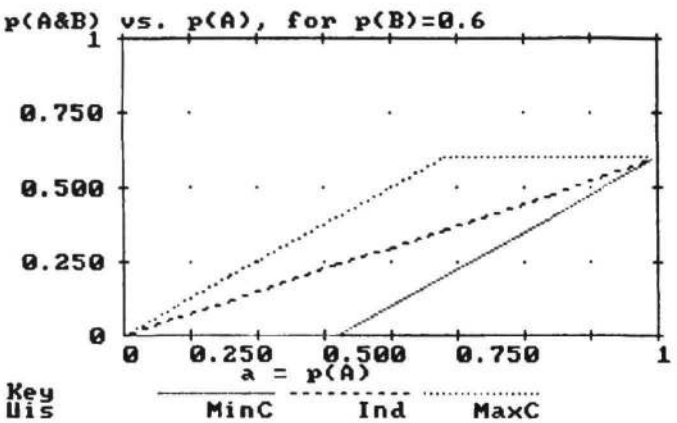


Figure 2: The effect of assumptions about correlation, (Wise & Henrion, 1986)

It has sometimes been claimed as an advantage of some non-probabilistic UISs, including Fuzzy Set Theory (Bonissone, 1986), that they avoid having to make *any* assumptions about dependencies. But in fact, as we have seen, the Fuzzy Set combination functions are equivalent, at least in effect, to specific probabilistic assumptions. It is true that non-probabilistic languages for uncertainty do not provide a general framework for modelling correlated evidence, since they do not provide a well-defined language for expressing the ideas. But to claim that they can therefore avoid making unsupported assumptions about correlations is akin to claiming that a new settler in Alaska can deal with the winter precipitation by adopting the language of an equatorial tribe with no word for snow.

Unfortunately to deal completely with probabilistic dependences is inherently complex (exponential in the number of evidence sources), and no inference network system which represents uncertainty in each proposition by one or a few parameters can deal with its full complexity. An alternative approach to is to represent uncertainty by a range of two probabilities and to compute both the largest and smallest probabilities compatible with the ranges of the antecedents. This does avoid making any specific, unsupported assumptions , although there is a danger of ending up with vacuous results (probability limits of 0 and 1).

The original Bayes' Network scheme of Kim and Pearl is restricted to Chow Trees i.e. singly connected graphs, so that conditional independence between convergent sources of evidence can be preserved (Kim & Pearl, 1983). However Pearl has suggested a method of removing the cycles, either by conditioning on variables in the cycle, or by adding extra nodes (hidden variables) that allows restructuring the probabilistic dependencies to avoid cycles (Pearl, 1985b).

Another approach is to represent the uncertainty in each proposition by a sample of truth values representing a random sample of possible worlds. These can be combined and propagating using the usual mechanisms of deterministic logic. Correlations due to multiple paths in the inference network or dependencies specified between inputs are handled correctly without special mechanisms. This *incidence calculus* (Bundy, 1986) or *logic sampling* (Henrion, 1986) involves a form of Monte Carlo simulation. Its accuracy depends on the sample size chosen. This approach can be reasonably efficient (it is linear in the network size) and seems promising, but its full potential remains to be explored.

Diagnostic and predictive inference

Diagnostic inference involves reasoning from observable manifestations to hypotheses about what may be causing them, for example reasoning from symptoms to diseases. *Predictive or causal inference* involves reasoning from causes (or causal influences, such as genetic or environmental factors that might increase susceptibility to a disease) to possible manifestations (Tversky & Kahneman,

1982). Consider the following:

> **The sneeze example:** *Suppose you find yourself sneezing unexpectedly in the house of an acquaintance. It might either be due to an incipient cold or your allergy to cats. You then observe animal paw marks, which increases your judged probability of a cat in the vicinity (diagnostic inference), which, in turn, increases the probability that you are having an allergic reaction (predictive inference). This also explains away the sneezing, and so decreases the probability you are getting a cold.*

Notice that this reasoning involves a mixture of both diagnostic and predictive inference. Having rules that allow reasoning backwards and forwards like this creates a danger of vicious circles, where, say, the probability of a cat would increase the probability of the allergic reaction, and vice versa. To avoid this, it seems necessary to keep a record of the sources of different uncertain evidence for each variable, so that you can avoid possible double counting. Pearl's scheme for Bayes' Networks keeps the flows of diagnostic and predictive evidence separate to avoid such cycling, and only combines them to calculate the aggregate degree of belief in each node (Pearl, 1985b). Figure 3 shows a Bayes Network representation of propositions mentioned.

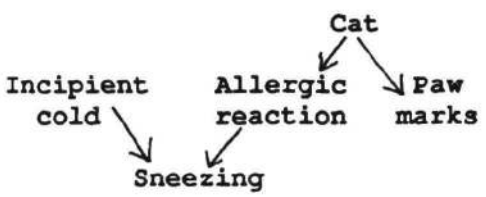


Figure 3: Bayes' network for sneeze example

Independent evidence for the allergy helps to explain the sneezing and so reduces the probability of a cold. Alternatively, the observation of a mild fever might increase the probability that it was a cold and so decrease the probability of the allergy. Thus the presence of sneezing induces a negative correlation between the cold and the allergy, which would otherwise be independent. This kind of reasoning, which we may term *intercausal*, is a natural consequence of the simple logical relation that sneezing can be caused either by a cold or an allergy. Pearl's propagation scheme for Bayes Networks models this correctly, but other rule-based schemes have a very hard time with it.

Consider the following rule for medical diagnostic inference which performs intercausal reasoning (Clancey, 1983):

> *If the patient has a petechial rash and does not have leukemia, then neisseria may be present.*

This reflects the medical fact that a petechial rash can be caused either by neisseria or by leukemia, and so the rash is evidence for neisseria unless it has been explained by leukemia.

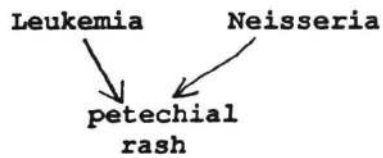


Figure 4: Bayes' network for rash example

If the system was also intended to help diagnose leukemia, it would need an additional rule:

> *If the patient has a petechial rash and does not have neisseria, then leukemia may be present.*

With both rules the inference network would contain a cycle (from leukemia to neisseria and back again), which is liable to cause updating difficulties, at least sensitivity to the sequence in which evidence arrives. The underlying problem is that these rules embody general knowledge about inference under uncertainty as well as specific medical knowledge.

It would be far better to be able to specify the essential medical knowledge in causal form, that "a petechial rash can be caused either by neisseria or by leukemia" (with specified conditional probabilities if the relationships are uncertain). The UIS should then be able use this to make the uncertain diagnostic inferences implied by either rule, or even reason predictively from the diseases to the symptom, according to the demands of the situation. Pearl's scheme can do this effectively and consistently, while maintaining a clear separation between the inference methods and domain knowledge. Clancey in a critique of Mycin has emphasized the desirability of separating the representation of inference strategy from domain knowledge (Clancey, 1983). But it does not appear that schemes, like Mycin, or Prospector representing knowledge primarily as diagnostic rules, rather than probabilistic causal relations, are capable of this.

Should UISs emulate humans?

This objection to probability has been forcefully stated by Paul Cohen:

> *"[it is] puzzling that AI retains models of reasoning under uncertainty that are derived from normative theories..., because the assumptions of the normative approaches are frequently violated, and because the probabilistic interpretation -- and numerical representation -- of uncertainty summarizes and fails to discriminate among reasons for believing and disbelieving. ... models of humans as perfect processors of information are not only inaccurate, but also unlikely to lead to efficient and intelligent reasoning." (Cohen, 1985), p.9*

Cohen here advocates the strategy, which appears to have been successful elsewhere in AI research, of adopting heuristic approaches based on human intuitive reasoning, rather than theoretically optimal, but computationally intractable schemes. Cognitive psychologists have indeed provided us with ample evidence that human inference under uncertainty is not accurately modelled by Bayesian decision theory (Kahneman, Slovic & Tversky, 1982). But there is little experimental evidence that proposed non-probabilistic UISs are better models. Very likely there is considerable variation between tasks and between individuals. It is an important and challenging task for cognitive psychologists to build better models of judgment under uncertainty, but it seems quixotic for those primarily interested in developing better expert systems to seriously attempt to emulate human judgment. That is not to say that evidence about human reasoning including introspection may not give us excellent ideas for devising new and better UISs, but the criterion for judging their usefulness should be the quality of their performance, rather than how well they simulate human thought processes.

One feature of human judgment observed by psychologists has been termed the *representativeness heuristic*: When asked the probability that object A belongs to class B, people typically evaluate it by the degree to which A is representative of B, that is by the degree to which A resembles B (Kahneman, Slovic & Tversky, 1982). This leads to judgments which are insensitive to the prior probability of A, and contrary to Bayes' rule. Cohen and colleagues explicitly adopt the representativeness heuristic for representing uncertainty in the classification system, GRANT, (Cohen *et al*, 1985), which deliberately ignores prior information. Other UISs also explicitly exclude prior probabilities, including Mycin (Buchanan & Shortliffe, 1984). The rationale has been that prior probabilities are too hard to estimate, and it is better to avoid them. However, for both GRANT and Mycin (and its derivatives), ignoring priors is functionally equivalent to assuming equal priors (for the probabilities of agencies funding a proposal, or the probabilities of disease organisms). Occasionally such flat priors may be appropriate, but more often it means ignoring important information about differing frequencies.

The following example points up the dangers of this approach:

> **Blood test example:** *James is engaged to be married, and takes the routine pre-marital blood test required by the state. To his horror, the test comes back positive for syphilis. His physician explains to him that the test is very reliable, having an false positive rate of 1%, and so the chance he has the disease is 99%. Aghast, James wonders what to tell his fiancee.*

Most physicians will give the same advice as James' one does. Like other people, they are poor intuitive Bayesians (Kahneman, Slovic & Tversky, 1982) and tend to ignore prior or base-rate information. Using the representativeness heuristic, the chance that James has VD is judged by the degree to which he (having a positive blood test) is representative of people with VD. In this, and many similar cases, the heuristic leads to a conclusion that is badly wrong:

> *Fortunately, James' fiancee, Alice, is not only understanding, but a Bayesian statistician. She finds out from the physician that the prevalence of syphilis among men from James' background is about 1 in 10,000. Based on this, she concludes that the probability he actually has the disease is about 1%, and decides to go ahead with the wedding.*

For those of us not so lucky as to be marrying a Bayesian, would we rather consult a physician or an expert system modelled on normal human judgment, or would we prefer one based on normative Bayesian principles?

Does it matter which you use?

Even if one accepts the arguments that probability is epistemologically adequate to represent uncertainty, it is clear that only approximations to it are computationally tractable for real systems. Despite the theoretical differences between systems, does it really make much difference to the conclusions of a rule-based expert system which scheme you use? There has been a common perception in the AI community that the performance of systems is relatively insensitive to the choice of UIS; that the important differences are to do with qualitative knowledge rather than quantitative uncertain inference. This may be true, at least for some domains. But so far, belief in this insensitivity seems to have been based primarily on ideology, since there has been little systematic analysis or experimental evidence published.

One early piece of evidence was a comparison of Mycin's method for combining evidence from different rules with a probabilistic model (Shortliffe & Buchanan, 1975). This showed a pronounced tendency to under-respond. On average, strong aggregate evidence for or against a conclusion was computed to be as about half as strong (half the CF) as it should be. This in itself, may not have mattered much, since Mycin used relative CFs for making decisions. But in 25% of the cases the system responded in the wrong direction. Confirming evidence actually reduced the CF or vice versa. The developers of Mycin suggest that Certainty Factors are satisfactory for the initial application domain (selecting antibiotic therapy), but that "We would need to perform additional experiments to determine the breadth of the model's applicability", (Buchanan & Shortliffe, 1984) p. 700. However CFs and related UISs are now being used for many other applications, apparently without the benefit of such experiments.

Recently there have been a few comparative studies of UISs. Tong and colleagues have compared 12 variants of the Fuzzy Set Rules for *and*, *or*, and *modus ponens* combinations in terms of their performance in a fixed rule base (Tong, 1985). They found that the performance of all rules with smooth response (i.e. not discontinuous) did reasonably well in their example. Vaughan and colleagues have done a comparison of the Prospector scheme with odds-ratio updating (Vaughan, 1986) for a systematic range of single rules. They found that Prospector did well in many cases, but that there are some situations in which it performs poorly. Wise has argued that the appropriate standard for comparison is a system using Maximum Entropy to fit a complete prior to specified input probabilities and rule strengths, and Minimum Cross Entropy for updating it (Wise & Henrion, 1986, Wise, 1986). This ME/MXE approach is actually a generalization of the odds-ratio approach used by Vaughan *et al.* Wise has performed comparisons of six UISs, including CFs, Fuzzy Set Theory, and a probabilistic scheme with Conditional Independence, against the ME/MXE scheme, for individual rules, and small assemblies of 2 or 3 rules, 30 cases in all, each with all input probabilities systematically varied (Wise, 1986). For purposes of comparison, the degree of membership of a Fuzzy Set was equated to probability. The performance of the UISs varied considerably over the different situations. All worked well in at least some cases, and none worked well in all cases. There were some situations in which some UISs were worse than random guessing.

It is not hard to construct examples in which CFs (and other widely-used UISs) produce results that disagree badly with a complete probabilistic analysis, even having the wrong qualitative sensitivities. Experienced knowledge engineers may be aware of at least some of the problems inherent in the UIS they use, and may know how to modify rule-sets to mitigate the undesirable behavior, at least for some anticipated situations. However, some of the problems are quite subtle, even though their effects can be severe. In any case it seems dangerous to rely on the ability of the knowledge engineer to "program around" such problems, particularly given our sketchy understanding of what all the problems are.

Conclusions

Probability has often been criticized as epistemologically inadequate for representing uncertainty in AI, but many of these criticisms have stemmed from incomplete understanding of probabilistic inference. In this paper, I have focussed on a number of important advantages that probabilistic representations have over other proposed measures of uncertainty, which have not loomed so large in the debate hitherto. Personal probability has an unambiguous operational definition, and it is embedded in a rational theory of decision-making under uncertainty -- we know what it means, and we know how to make decisions using it. Probabilistic inference is epistemologically adequate to perform three important kinds of reasoning that humans are capable of: (a) taking into account non-independence between sources of evidence, (b) engaging in mixed diagnostic and predictive inference, and (c) inter-causal inference, between alternative causes of an event, as in "explaining away". Non-probabilistic rule-based UISs may be able to simulate these in particular cases, at least qualitatively, but only by confounding general knowledge about uncertain inference with the domain specific knowledge in the rules.

These types of inference are important, and we can learn a great deal from studying human reasoning. But it is not necessarily desirable that a UIS should duplicate *all* features of human judgment under uncertainty, including such strategies as the representativeness heuristic that can lead to severe biases, as in the blood test example. Where cognitive limitations cause human judgments to diverge from the results of normative theory, surely it is better to use the latter when expert systems are advising on important decisions, as in medical or defense applications.

If one accepts the arguments for the epistemological adequacy, or even superiority, of probability, serious questions may still be raised about its heuristic adequacy -- can practical, computationally efficient implementations be built? The Bayes' Network approach seems very promising, but work still remains to be done to deal conveniently and generally with multiply connected networks (i.e. dependent sources of evidence). Monte Carlo logic sampling seems to offer possibilities here, both as a practical implementation and as an intellectual link to deterministic logic. Although several probabilistic methods for dealing with second-order uncertainty, distinguishing the effect of different sources of evidence, and explaining probabilistic reasoning have been suggested, there remains considerable work to be done to develop implementations and experimental study of their acceptance and usefulness to system builders and users. Whatever the theoretical merits of probabilistic representations, the AI community has a venerable tradition of pragmatism, and many will understandably remain unconvinced until these more sophisticated probabilistic schemes have demonstrated success in large scale applications. On the other hand, disturbing evidence is emerging about the performance of the most popular UISs, and complacency would be inappropriate as they are applied to new tasks with major potential consequences. There is an urgent need for more rigorous experimental evaluations of UISs for a range of realistic rule-bases to find out under what circumstances they can be relied on, and when they may be seriously wrong.

Acknowledgements

These ideas have benefitted from discussions with Ben Wise and Chris Elsaesser. This work was supported in part by the National Science Foundation under grants PRA-8413097 and IST-8603493.

References

Beyth-Marom, R. How probable is probable? A numerical taxonomy translation of verbal probability expressions. *J. of Forecasting*, 1982, *1*, 257-69.

Bonissone, P.P. A Fuzzy Sets based linguistic approach: Theory and applications. In M.M. Gupta & E. Sanchez (Eds.), *Approximate Reasoning in Decision Analysis*. North-Holland, 1982.

Bonissone, P. P. Plausible Reasoning: Coping with Uncertainty in Expert Systems. In *Encyclopedia of AI*. John Wiley & Sons, 1986.

Buchanan, B. G. & Shortliffe, E. H. *Rule-based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project.* Addison-Wesley, Reading, Mass, 1984.

Bundy, A. Incidence Calculus: A mechanism for probabilistic reasoning. In Kanal, L.N. & Lemmer, J. (Eds.), *Uncertainty in AI.* North-Holland: Amsterdam, 1986.

Cheeseman, P. A methods of computing generalized Bayesian probability values for expert systems. In *Proceedings of 8th International Joint Conference on AI.* Karlsruhe, W. Germany: , 1983.

Cheeseman, P. In Defense of Probability. In *Proceedings of 9th International Joint Conference on AI.* Los Angeles, Ca: , 1985.

Clancey, W.J. The Epistemology of a Rule-Based Expert System - A Framework for Explanation. *AI*, 1983, *20*, 215-251.

Cohen, Paul R. *Heuristic Reasoning about Uncertainty: An AI Approach.* Pitman: Boston, 1985.

Cohen, P., Davis, A., Day, D., Greenberg, M., Kjeldson, R., Lander, S., & Loiselle, C. Representativeness and Uncertainty in Classification Systems. *The AI Magazine*, Fall 1985, pp. 139-149.

de Finetti, Bruno. *Theory of probability.* New York: Wiley, 1974.

Doyle, J. *The Ins and Outs of Reason Maintenance* (Tech. Rep. CMU-CS-83-126). Carnegie Mellon University, 1983.

Duda, R. O., Hart, P. E., and Nilsson, N. J. *Subjective Bayesian Methods for Rule-Based Inference Systems* (Tech. Rep. SRI Technical Note 124). SRI International, 1976.

Fox, J. Knowledge, decision making and uncertainty. In W. Gale (Ed.), *AI and Statistics.* Addison-Wesley: Reading, MA, 1986.

Gale, W. (ed.). *AI and Statistics.* Addison-Wesley: Reading, MA, 1986. (In press).

Henrion, M. *Propagation of Uncertainty by Logic Sampling in Bayes' Networks* (Tech. Rep.). Department of Engineering and Public Policy, Carnegie Mellon University June 1986.

S. Holtzman & J. Breese. Exact reasoning about uncertainty. In *Workshop on Uncertainty and Probability in AI.* AAAI 1985.

Kahneman, D.; Slovic, P. & Tversky, A. *Judgment under Uncertainty: Heuristics and Biases.* Cambridge: Cambridge University Press, 1982.

Kanal, L.N. & Lemmer, J. (eds.). *Uncertainty in AI.* North-Holland: Amsterdam, 1986.

Kim, J. & Pearl, J. A Computational Model for Combined Causal and Diagnostic Reasoning in Inference Systems. In *Proceedings of IJCAI-83.* , 1983.

Lichtenstein,S., Fischhoff, B. & Phillips, L.D. Calibration of probabilities: The state of the art to 1980. In Kahneman, D.; Slovic, P. & Tversky, A. (Eds.), *Judgment under Uncertainty: Heuristics and Biases.* Cambridge: Cambridge University Press, 1982.

Lindley, D.V. Scoring rules and the inevitability of probability. *international Statistical Review*, 1982(50), pp. 1-16. (with commentary).

Oden, G.C. Integration of Fuzzy Logical Information. *J. Experimental Psychology: Human Perception and Performance*, 1977, *3*(4), 565-75.

Pearl, J. *How to do with probabilities what people say you can't* (Tech. Rep. CSD-850031). Cognitive Systems Laboratory, Computer Science Department, University of California, Los Angeles, CA, 1985.

Pearl, J. *Fusion, Propagation, and Structuring in Bayesian Networks* (Tech. Rep. CSD-850022). Cognitive Systems Laboratory, Computer Science Department, University of California, Los Angeles, CA, 1985.

Savage, LJ. *The Foundations of Statistics.* New York: John Wiley & Sons, 1954.

R.D. Shachter. Intelligent Probabilistic Inference. In *Workshop on Uncertainty and Probability in AI.* AAAI, 1985.

Shafer, G. *A Mathematical Theory of Evidence.*

Princeton University Press, 1976.

Shore, J. E., and Johnson, R. W. An Axiomatic Derivation of the the Maximum Entropy Principle. *IEEE Transactions on Information Theory*, January 1980, pp. .

E. H. Shortliffe & B. G. Buchanan. A model of Inexact Reasoning in Medicine. *Mathematical Biosciences*, 1975(23), pp. 351-379.

Spiegelhalter, D. J. A statistical view of uncertainty in expert systems. In W. Gale (Ed.), *AI and Statistics*. Addison-Wesley: Reading, MA, 1986. (in press).

Tong, R. M., Shapiro, D. G. Experimental Investigations of Uncertainty in a Rule-based System for Information Retrieval. *International Journal of Man-Machine Studies*, 1985.

Tversky, A. & Kahneman, D. Causal Schemata in Judgments under Uncertainty. In Kahneman, D.; Slovic, P. & Tversky, A. (Eds.), *Judgment under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press, 1982.

Vaughan, D. S., Perrin, B. M., Yadrick, R. M., Holden, P. D., Kempf, K. G. An Odds Ratio Based Inference Engine. In Kanal, L.N. & Lemmer, J. (Eds.), *Uncertainty in AI*. North-Holland: Amsterdam, 1986.

Wise, B.P. *An experimental comparison of uncertain inference systems*. Doctoral dissertation, Carnegie Mellon University, Department of Engineering and Public Policy, 1986.

Wise, B.P. & Henrion, M. A Framework for Comparing Uncertain Inference Systems to Probability. In Kanal, L.N. & Lemmer, J. (Eds.), *Uncertainty in AI*. North-Holland: Amsterdam, 1986.

Zadeh, L. A. Making Computers Think Like People. *IEEE Spectrum*, August 1984, pp. .

Specificity of Expertise in Clinical Reasoning

Guy-Marie Joseph & Vimla L. Patel

McGill University

An important development in recent research in problem solving has been the development of new theories and more rigorous methods with which to formulate detailed hypotheses about the processes that constitute expert performance in various domains. Domains in which such research can be found include chess (Chase & Simon, 1973), physics (Larkin et al., 1980), and medicine (Elstein et al., 1978; Patel & Groen, 1986).

Studies of expertise have led to the hypothesis that the dominant component of expert performance is related to the expert's ability to represent problems successfully, and that this results from the expert's having a rich and well organized knowledge base in which patterns of features in the problem are associated with concepts at varying levels of generality. This enables experts to recognize the salient features of the problem, to interpret information relevant to its solution (Johnson et al., 1981), and search for the methods and operations to be used in solving the problem (Greeno & Simon, 1984) all with greater efficiency than novices. Moreover, expert problem solvers tend to use forward chaining strategy when solving a simple or familiar problem (Simon & Simon, 1978; Chi et al., 1981; Glaser, 1985; Larkin et al., 1980). This forward-chaining strategy changes, however, to a very sophisticated means- ends analysis (Greeno & Simon, 1985; Chi et al., 1981) or backward chaining (Bhaskar & Simon, 1977) when solving a difficult or unfamiliar problem. These studies have provided evidence suggesting hypotheses about the kinds of processes and knowledge structures experts use while working on familiar problems or problems within their domain of expertise.

These research suggests that expertise is determined by subjects' (i) prior knowledge of a particular domain, (ii) their knowledge of general problem solving methods, and (iii) their knowledge of any domain-specific heuristics. Testing experts in problem solving outside their

domain of expertise should thus make it possible to contrast their performance in the two domains and thereby begin to assess the relative contributions of domain-specific knowledge vs. general problem solving ability. The need for making this contrast, and the difficulty of doing so, has been pointed out most recently by Greeno & Simon (1984).

Studies of experts working on problems outside their specialty area have found that experts' problem-solving skills diminish notably in such unfamiliar domains (Kassirer et al., 1982). Miller (1975) found that experts approach non-specialty problems "more parsimoniously", and that they resort to strategies commonly associated with novice performance. Means-ends analysis was found to be the primary strategy used when the problem solver has little domain-specific knowledge (Greeno and Simon,1984). Ernst and Newell (1969) have shown that with unfamiliar problems, experts rely more on problem-solving strategies rather than on domain-specific knowledge. Boshuizen and Claessen (1977) identified some of the problems involved in such studies. Another important characteristics of expert performance is the difference between the solution strategies used by experts and novices. The distinction is that experts use a "forwar-chaining" strategy, whereas novices use a "backward-chaining strategy".

Subjects' problem spaces are determined, in general, by the search methods that the subjects use, the features of the problem's solution, and their knowledge of the problem domain. In situations where expert problem solvers do not have special training in a problem domain, they must construct a problem space that includes representations of the strategies and constraints necessary to solve the problem. If subjects have special training or experience in the problem domain, their prior knowledge includes general characteristics of the problem space, and their representations of the individual problems are based on this knowledge (Greeno and Simon, 1984). These results and observations lead us to believe that experts working within and outside of their specialities will use very different strategies and/or whatever domain knowledge they have available.

Attempts to assess the relative contributions of domain knowledge vs. problem solving strategies require specific information about both how the subject's representation of the problem

changes over time and the domain knowledge necessary to construct and modify these representations. Neither of the two most widely used methods, however, seem to be able to provide this kind of information (Patel et al., 1986). "Think-aloud" protocols (e.g. Greeno & Simon, 1984; Kassirer et al., 1982) provide rich, complex data that are approximately concurrent with the subject's reasoning and therefore provide information about the subject's changing representation of the problem. However, protocol analysis methods (e.g., Ericsson & Simon, 1984) have been limited in their success at providing more than global information about subjects' processing. One of the major concerns raised by protocol analysis is the lack of formal, objective methods for analyzing complex verbal material. Methods of propositional analysis have been proposed to meet this concern (e.g., Kintsch, 1974; Frederiksen, 1975) and have most often been applied to stimulus texts and recall protocols. Although recall protocols provide quite detailed data about representation in memory and inferences based on domain knowledge, they are very limited in the amount of information they can provide about how representations of a problem change over time.

Recently, a number of researchers have addressed these problems by combining the two methods, i.e., performing propositional analyses of think-aloud or on-line protocols (e.g. Kintsch & Greeno, 1985). This strategy has proven useful for studying strategies of reading comprehension (Scardamalia & Bereiter, 1984) and writing (Breuleux & Bracewell, 1986). As a further refinement, this study prompts subjects to think aloud at regular intervals. This on-line procedure allows finer control of the stimuli the subject is to respond to, and may provide more detailed information about the subject's changing representation of the problem as a function of specific portions of the input text.

In the present study we are primarily interested in testing the extent to which expertise is dependent on domain knowledge, rather than the interaction of domain knowledge and problem solving processes. If expertise is primarily dependent on domain knowledge, then when working on problems outside their domain of expertise, experts should resort to strategies (e.g. backward chaining) that are commonly associated with novices. However, if expertise is more a function of

general problem solving methods, then there should be few differences in expert performance across the boundaries of their domain of expertise.

METHOD

Subjects.

Nine senior physicians associated with the Faculty of Medicine at McGill University volunteered as subjects for the study. Although all subjects had a general knowledge of medicine, the two groups differed in their specific knowledge of endocrinology: there was a high domain-knowledge (HDK) group consisting of four endocrinologists and a low domain-knowledge (LDK) group consisting of five cardiologists.

Materials.

A description of a clinical case in endocrinology based on a real patient was selected as stimulus material (Appendix 1). The case was described in the following order; the patient's history (Segments 1 to 4), findings obtained from examination of bodily systems (Segments 4 to 22), and results from laboratory tests (Segments 23 and 24).

The diagnosis of the case has three subcomponents varying in specificity. The first subcomponent is a slow progressive decrease in thyroid function: chronic hypothyroidism. The second subcomponent is a change from a chronic to an acute form of hypothyroidism, precipitated by a prescribed cough medicine and leading to a state of myxedema pre-coma. The third subcomponent is an autoimmune condition called Hashimoto's thyroiditis as the cause of the hypothyroidism. The complete diagnosis, then, is: Hashimoto's thyroiditis leading to hypothyroidism, precipitated to myxedema pre-coma by previous medication.

Procedure.

Subjects were tested individually at their hospital offices. They were first given a short practice session to familiarize themselves with the procedure. The clinical case was presented on the screen of an Apple Macintosh one sentence at a time, using software developed for this purpose by the first author. Subjects initiated the presentation of each sentence by pressing the mouse

button. For each sentence presented, subjects were asked to verbalize their thoughts about the contribution of the sentence towards the generation of a diagnosis. As soon as the subject pressed the mouse button again, the next sentence was presented, replacing the previous one. Subjects were not allowed to look back to any prior part of the text. After presentation of the entire case, subjects were asked to summarize the case and to provide their diagnosis(es). Subjects' verbalizations were recorded and later transcribed verbatim.

ANALYSES

Protocol Analysis.

The information in each text segment of the original stimulus text was initially categorized as either critical, relevant, or nonrelevant, based on the judgements of two expert endocrinologists. Cues were rated as "critical" if they were deemed necessary and sufficient for the generation of the accurate diagnosis. Relevant cues were considered necessary but not sufficient for the generation of the accurate diagnosis. Nonrelevant cues included the remainder of the information, the omission of which should not have affected the generation of an accurate diagnosis. The inter-rater reliability was **.92** for the critical cues and **.85** for the relevant cues.

Each subject's protocol was transcribed so that each segment from the original text was followed by the physician's comments on that segment. Subjects' comments on each segment were then coded for the presence of the different types of cues, and links generated from the cues to hypothetical causes. If the cue was only repeated and no link was suggested, then it was coded as absent (see example 1). If the cue was repeated and a link was suggested, then the cue was coded as accessed and a link was coded as hypothesized (see example 2).

(1) ***Segment #9***: Her body temperature was 36 deg. C.
Subject: Her body temperature was 36 deg. C., that does not tell me very much.

In the preceding segment (see Appendix 2, Subject 2), there are no cues coded as accessed and no links generated from it.

(2) ***Segment #3***: She complained of feeling tired all the time, had a loss appetite, a 30 lb. weight gain and constipation.

> ***Subject***: She had a loss of appetite, a 30 lb. weight gain, and constipation. OK... Right now we are wondering whether *she has got hypothyroidism* when you are looking at this.

In this situation there are three cues accessed (underlined), and three links generated from these cues (linking statement in *italics*) to one of the subcomponents of the diagnosis (hypothyroidism). This method is described in more detail in Joseph & Patel (1986).

Accuracy of diagnoses.

The diagnoses of all the subjects were analyzed for the presence or absence of the three subcomponents with the assistance of an expert endocrinologist who did not know which subjects belonged to which group.

Frame analyses.

One representative subject from each group was analysed in more detail (Subjects 2 and 3). Frame analyses (cf. Patel & Groen, 1986) based on methods of propositional analysis were performed on the relations that the subjects generated between cues and their hypothetical causes, and generated semantic networks corresponding to each hypothesized relation. Subjects' frames were then compared to determine their strategies for using the cues provided to generate diagnostic hypotheses, as well as differences in the problem representations they generated as the information was presented.

RESULTS AND DISCUSSION

Protocol analysis: *Cues selected and links generated.*

The results of the analysis of cues selected and links generated indicate that the main difference between the two groups is the number of links generated from the critical cues (Figure 1). High domain knowledge HDK subjects generated more links from the critical cues selected whereas low domain knowledge LDK subjects generated fewer links from the critical cues selected. Both groups generated fewer links from relevant cues. However, the HDK group selected more relevant cues and generated more links from these cues than did the LDK group.

There was also greater variability between groups in the relationship between critical cues selected and links generated, than for the links generated from relevant cues.

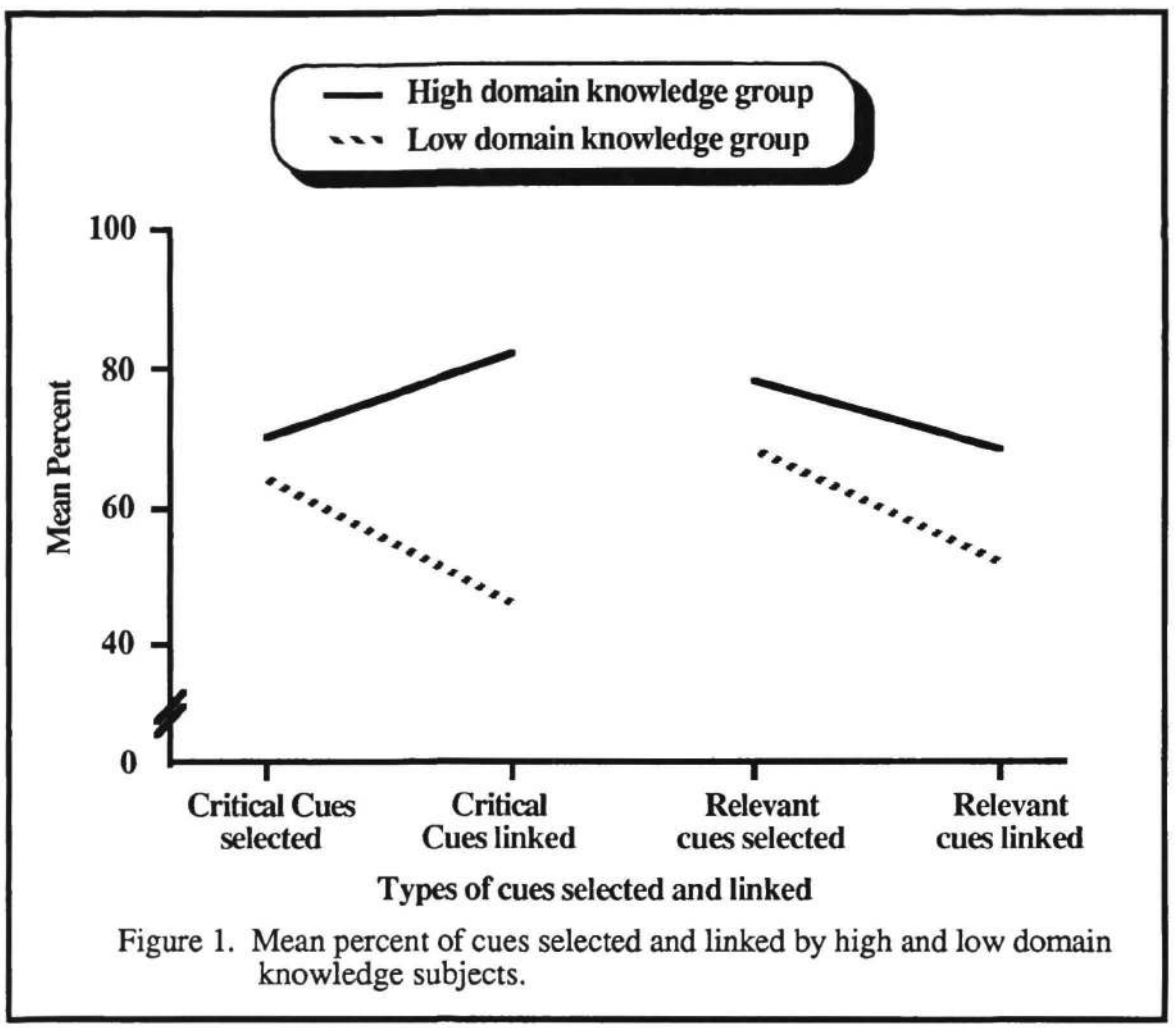


Figure 1. Mean percent of cues selected and linked by high and low domain knowledge subjects.

Accuracy of diagnoses:

All the subjects in this group generated the general subcomponent of hypothyrodism in their diagnostic statement (Figure 2). Sixty percent generated the myxedema subcomponent, but they all omitted the pre-comatose aspect of the myxedema. Finally, only twenty percent of the subjects generated the Hashimoto subcomponent as part of their diagnostic statement. In addition, subjects generated differential diagnoses, an ordered list of the most likely diagnostic possibilities of the causes of the patient's disorder. The most frequent differentials listed were: Adrenal failure, pituitary tumor, Hypophyseal disorder, and pericardial effusions.

Thus far, the use of surface measures to evaluate the performances of experts working on problems within and outside their domain of expertise has not yielded meaningful information about the role of domain knowledge and problem solving strategies in the process of clinical reasoning. The main difference in the analysis of accuracy of the diagnosis that LDK experts are more general in determining their diagnostic statement. Their uncertainty thus explains why they generate a very general list of differentials for the causes of the patient's disorder.

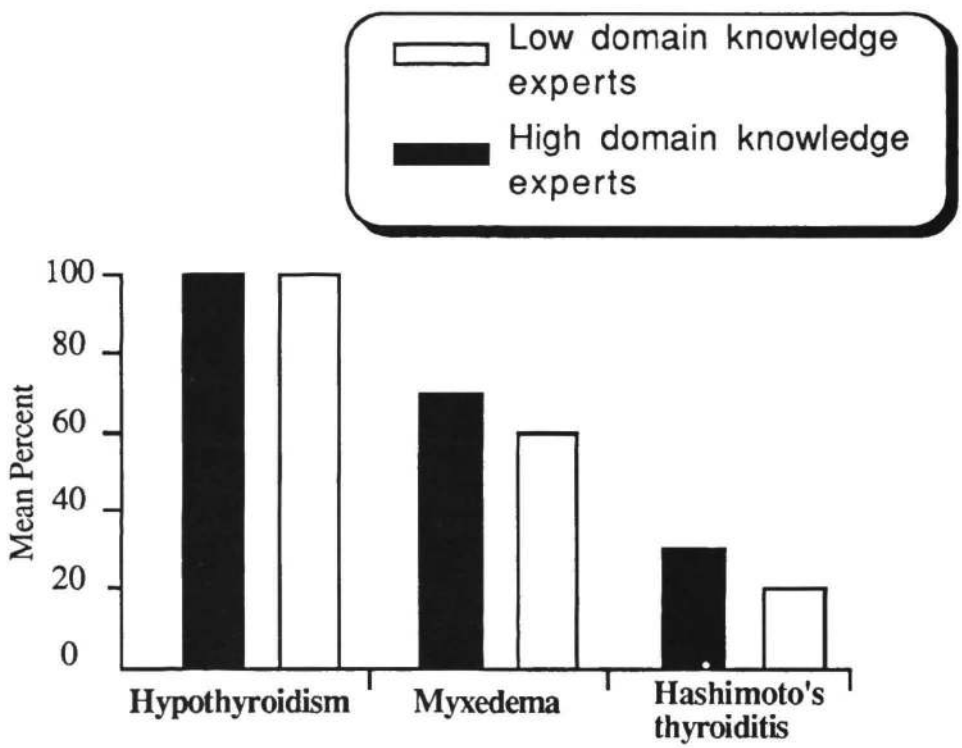


Figure 2. Mean percent of subcomponents of the acurate diagnosis by high and low domain knowledge subjects.

The latter finding is consistent with studies of comprehension of clinical cases in which no differences were found for experts' recall of relevant and non-relevant information in "atypical" cases (Coughlin & Patel, 1986). Therefore, more detailed and structural analysis of, the heuristic

strategies and prior knowledge that are needed to generate a representation of the problem is expected to yield process differences between the two groups. The results of this analysis is provided in the next section.

Analysis of the online protocol of a HDK subject: The analysis of the structural representations of the online protocol of an HDK subject is presented segment by segment in Figure 3. Segments in which there were no responses were omitted. The first important finding of this analysis was that although this subject identified the three subcomponents of the diagnosis, he did not report them when asked to provide his diagnosis. His diagnostic hypothesis started with a "forward chaining" strategy, followed by a "backward chaining" strategy used to elaborate on the underlying causal mechanisms of the relationship between the cues and the hypothesis (Figure 3, Segments 3 and 4). After the generation of the three diagnostic subcomponents (Segment 7), the processing of the cues followed a model of "pattern recognition" in order to either confirm the diagnostic hypothesis or to rule out the two hypotheses that were generated at the beginning of the case.

Analysis of the online protocol of a LDK subject: The analysis of the structural representation of the online protocol of a LDK subject is provided in Figure 4. Contrary to the representations of the HDK subject, this subject provided only the general subcomponent of the diagnosis, hypothyrodism at Segment 4. The other two components were not mentioned with the exception of Segment 9, where the subject mentions the "advanced hypothyrodism". The subject generated many diagnostic hypotheses (e.g. psychiatric disorder, chronic renal disease, anemia) without ruling them out as possibilities, and seemed to generate the hypotheses using a form of "backward chaining" (Segments 4 and 5). Finally, this subject generated very few causal explanations (e.g. causal links) and instead used a process that was more cue driven. That is, evaluated individual cues as they were presented. In some instances he ruled out a possible hypothesis but reinstated the hypothesis at the presentation of the next cue (Segments 10 and 11).

The results of the following study would tend to support the hypothesis that when working on problems in which experts lack specific domain knowledge, they rely on general knowledge in

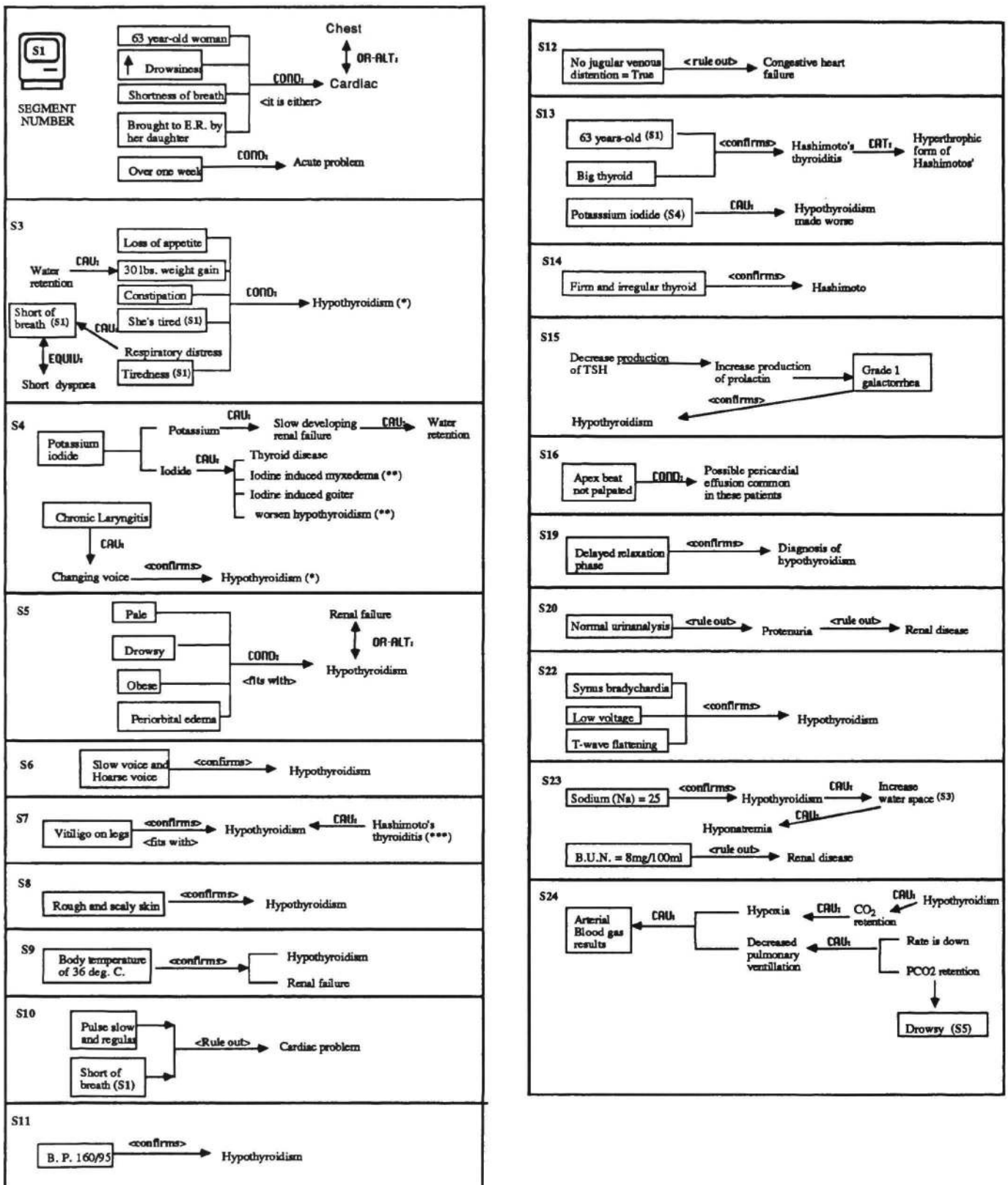


Figure 3. Structural representation of the online protocol segment by segment of a high domain knowledge subject. (S#2)

(*): Indication of the generation of the first subcomponent of the diagnosis for the first time
(**) Indication of the generation of the second subcomponent of the diagnosis for the first time
(***) Indication of the generation of the third subcomponent of the diagnosis for the first time

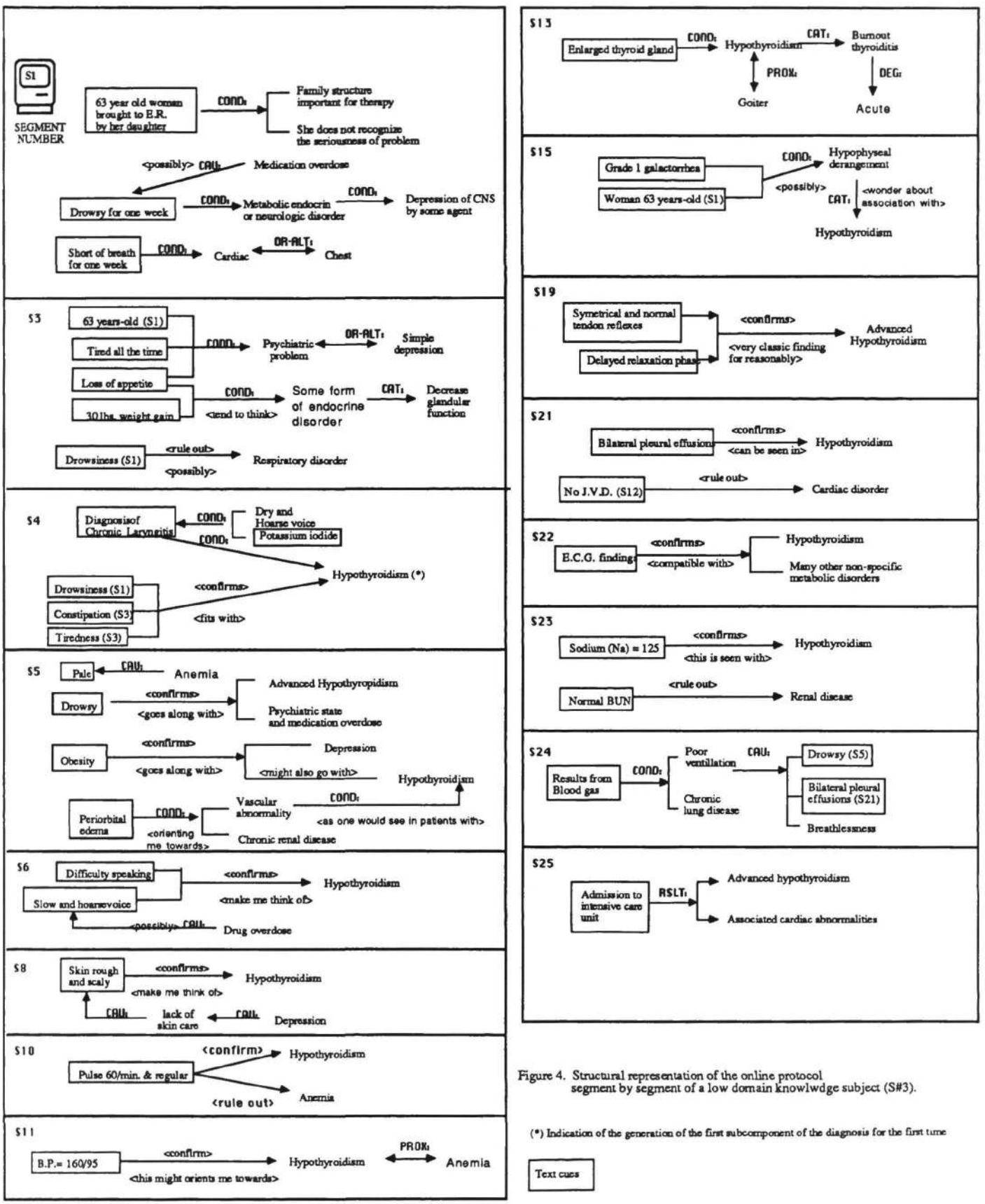


Figure 4. Structural representation of the online protocol segment by segment of a low domain knowlwdge subject (S#3).

(*) Indication of the generation of the first subcomponent of the diagnosis for the first time

Text cues

the problem domain to solve the problem. However, the use of general knowledge directly affects the overall representation of the problem as well as the types of strategies that are used to generate the solution.

The analysis of the two representations and strategies generated by the two subjects revealed information about process differences between HDK and LDK experts. The HDK subject used forward chaining to generate hypotheses and used backward chaining strategy to explain the relationship between the cues and the hypotheses, in providing the underlying pathophysiological mechanisms of the patient's disease. This finding is consistent with the earlier finding that HDK experts generated more links from critical cues than the LDK expert. The HDK subject's representation of the problem was also more elaborated with more causal links. He also provided fewer diagnostic hypotheses and ruled out more alternative hypotheses. These finding led to the use of the term "mechanismic" to describe the process of the HDK expert.

The LDK expert used more backward chaining strategies. He generated more diagnostic hypotheses and his problem representation was much less elaborated and detailed than the HDK expert. The hypotheses generated by the LDK expert were very general and were related to the general subcomponent of hypothyroidism. He evaluated and classified the cues in terms of their relationship to the condition of hypothyroidism. We therefore use the term "classificatory" to describe the process of the LDK expert.

The methods and analysis techniques used in this study reflect the complex nature of the relationship between domain knowledge and problem solving strategies, as well as the problem that researchers must face in the assessment of such complex phenomena. The use of the online methods seems very promising in the study of cognitive processes, especially to study the types of heuristic strategies that used to solve problems, strategiy shifts, and the types of knowledge required to do so.

ACKNOWLEDGEMENTS

We wish to thank Drs. Y. Patel, S. Magder and D. St. Vil for their assistance with the medical aspects of the study. We particularly want to acknowledge the support of the physicians who generously voluntereed their time. We wish to thank M. Dillinger, Dr. R Bracewell, and A. Breuleux, and the graduate students for their help at the various stages of the study. This work was supported in part by grants from the Josiah Macy Jr. Foundation (No. B8520002) and the Natural Sciences and Engineering Council of Canada (A2598) to Vimla Patel.

REFERENCES

Bereiter, C., & Bird, M. (1985). Use of thinking aloud in identification and teaching of reading comprehension strategies. Cognition & Instruction, 2, 131-156.

Bhaskar, R., & Simon, H. A. (1977). Problem solving in semantically rich domains: An example from engineering thermodynamics. Cognitive Science, 1, 193-215.

Boshuizen, H. P. A., & Classen, H. F. A. (1982). Problems of research into medical-problem solving. Medical Education, 16, 81-87.

Breuleux, A. & Bracewell, R. J. (1986). A discourse analysis approach to the analysis of think aloud protocols. McGill working papers in cognitive sciences #7. Montreal, McGill University.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. Cognitive Psychology, 1, 55-81.

Chi, M. T. H., Glaser, R., & Rees, E. (1981). Expertise in problem solving. In R. J. Sternberg Advances in the psychology of human intelligence. LEA.

Coughlin, L. D., & Patel, V. L. (1986). Text comprehension and the effect of expertise in the domain of medicine. Paper presented at the Annual Conference of the American Education Research Association, San Francisco, CA.

Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1978). Medical problem solving: An analysis of clinical reasoning. Cambridge, MA: Harvard University Press.

Ericsson, K. A., & Simon, H. A. (1984). Protocol analysis: Verbal reports as data. Cambridge, MA: MIT press.

Ernst, G. W., & Newell, A. (1969). GPS: A case study in generality and problem solving. New York: Academic Press.

Frederiksen, C. H. (1975). Representing logical and semanticstructure of knowledge acquired from discourse. Cognitive Psychology, 7, 371-458.

Glaser, R. (1985). Thoughts on expertise. Technical report #8, LRDC, University of Pittsburgh, Pittsburgh, PA.

Greeno, J. G., & Simon, H. A. (1986). Problem solving and reasoning. In R. C. Atkinson, R. J. Herrnstein, G. Lindzey, & R. D. Luce (Eds.), Stevens' handbook of experimental psychology (rev. ed.). New York: Wiley.

Groen, G. J., & Patel, V. L. (1985). Medical problem solving and cognitive psychology: Some questionable assumptions. Medical Education 19, 95-100.

Hayes-Roth, B., & Hayes-Roth, F. (1979). A cognitive model of planning. Cognitive Science, 3, 275-310.

Johnson, P. E., Duran, A. S., Hassebrock, F., Moller, J., Prietula, M.,Feltovich, P. J., & Swanson, D. B. (1981). Expertise and error in diagnostic reasoning. Cognitive Science, 5, 235-283.

Joseph, G-M., & Patel, V. L. (1986). Online analysis of expertise in medical problem solving. Technical Report No. 11, Center for Medical Education, Faculty of medicine, McGill University.

Kassirer, J. P., Kuipers, B. J., & Gorry, G. A. (1982). Toward a theory of clinical expertise. The American Journal of Medicine 73, 251-259.

Kintsch, W. (1974). The representation of meaning in memory. Hillsdale, NJ: LEA.

Kintsch, W., & Greeno, J. G. (1985). Understanding and solving word arithmetic problems. Psychological Review, 92, 109-129.

Larkin, J., McDermott, J., Simon, D. P., & Simon, H. A. (1980). Expert and novice performance in solving physics problems. Science, 208, 1335-1342.

Miller, P. B. (1975). Strategy choice in medical diagnosis, Project MAC. Report no. TR-153, Massachussets Institute of Technology.

Patel, V. L., Frederiksen, C. H., & Groen, G. J. (1983). Relationship between comprehension and medical problem solving. In the essence of clinical competence: Psychological studies of expert reasoning in medicine. Symposium conducted at the Research in Medical Education Conference of the Association of American Medical Colleges, Washington, DC.

Patel, V. L., & Frederiksen, C. H. (1984). Cognitive processes in comprehension and knowledge acquisition by medical students and physicians. In H. Schmidt & H.G. DeVolder (Eds.), Tutorials in Problem Based Learning. Assen, Holland: Van Gorcum.

Patel, V. L., & Groen, G. J. (1986). Knowledge based solution strategies in medical reasoning. Cognitive Science, 10, 91-116.

Patel, V. L., Arocha, J. F., & Groen, G. J. (1986). Strategy selection and degree of expertise in medical reasoning. Paper in this volume.

Scardamalia, M., & Bereiter, C. (1984). Development of strategies in text processing. In H. Mandl, N. Stein, & T. Trabasso (Eds.), Learning and comprehension of text. Hillsdale, NJ: LEA.

Simon, D. P., & Simon, H. P. (1978). Individual differences in solving physics problems. In R. Siegler (Ed.), Children's thinking: What develops? Hillsdale, N.J.: Lawrance Earlbaum.

REFLEXIBILITY IN PROBLEM SOLVING: THE SOCIAL CONTEXT OF EXPERTISE

James A. Levin
University of Illinois

Margaret M. Riel
University of California, San Diego

Moshe Cohen
Hebrew University of Jerusalem, Israel

Michael Goeller
Marcia Boruta
University of California, San Diego

Naomi Miyake
Aoyama Gakuin Woman's College, Tokyo Japan

ABSTRACT

What are the factors that cause a problem solver to become blocked? And what are the factors that allow a person to become unblocked? These are the motivating questions for a set of studies we conducted of individual and joint problem solving. By constructing an isomorph of the classic "water jar" problems (Luchins, 1942) as a dynamic graphic microworld, we were able to identify several factors involved in producing blocked states. By comparing the behavior of individuals tackling the "missionaries and cannibals" problem to pairs of people solving this problem, we have been able to identify ways in which problem solvers operating in a social context are able to overcome problem solving blocks that are difficult for individuals. These studies point to the importance of "reflection" (evaluation of problem solving results) for flexible problem solving. These results may also account for the difficulty in showing learning in "discovery learning" uses of computers, such as the use of Logo, since such uses also often do not encourage students to reflect on the outcome of their problem solving.

When cognitive theorists and educators examine problem solving, the focus is generally on the initial steps in the problem solving process: problem definition, alternative paths possible to the solution, and the possible problems that arise when people "fail" to discover the relatively easy solution because of negative transfer from other problems or domains.

There is, however, an important part of the problem solving process that is less often described in cognitive research and often is missing from instruction in problem solving. This is the reflection or evaluation of the solution that was found. Was it the best possible solution? The only solution? How was it discovered? Could it be repeated? What justification can the problem solver offer

for his or her move?

We often solve problems by looking for the most immediate solution with little concern for other ways that we could have solved the problem. If the problem is never likely to reoccur this method may be appropriate. However, for problems that reoccur, skilled problem solvers will be those who have a deeper understanding of the fit between the problem and the problem solution. We will discuss here several experiments in two different problem solving situations which point to the vital role that reflection play in problem solving.

The Missionaries and Cannibals Problem

One way to get problem solvers to reflect on their problem solving strategies is to put them in a social situation in which they must convince one another that a given move or strategy is the best one.

"There are three missionaries and three cannibals on one side of the river and your task is to get them across the river using a two person boat without ever letting the cannibals outnumber the missionaries on a side of the river." This is a commonly used problem that involves a sequence of eleven steps from initial state to conclusion (Ernst & Newell, 1969; Reed, Ernst, & Banerji, 1974; Jeffries, Polson, Razran, & Atwood, 1977; Levin & Hutchins, 1981). The sequence is difficult to see immediately and finding the solution often involves illegal moves and repetition of moves that do not move the problem solver any closer to the solution.

We had subjects solve this problem in two different conditions. The first is the more standard situation in which a single person manipulates objects that represent the problem while "talking aloud" about the steps of problem solving. The verbal protocols often related what the problem solver was doing, but not why. The plans and strategies remained difficult to infer from the verbal reports.

The second condition was to ask two people to work together to solve the problem. When people solve problems in social settings, they often discuss their plans or reflect on the success of previous actions (Miyake 1986; Suchman, 1985; Martin, 1983). The need for coordinated action and division of labor often leads to shifting responsibility for monitoring and evaluating each action taken by the group.

We had 10 individuals and 10 pairs of subjects solve the problem. All sessions were audiotaped with an experimenter taking notes about things that will not be available from the tape. The notes and the audiotape are used to construct a transcript of the verbal interaction and a problem solving transcript. The problem

solving transcripts list all moves either made or considered from the initial state through the eleven steps necessary for the solution of the problem. We used this information to compare the individual sessions with that of the pairs.

Comparison of individual and joint problem solving. There is a very different pattern when a pair of subjects solve the missionaries and cannibals problem than when individual subjects work alone. The single subjects on the average took twice as long, (13.30 minutes for single subjects and 7.09 for pairs), and made more moves (27 to 17) with a higher percent of both illegal (1.2% to .08%) and of repeat (40.2% to 22.4%) moves. The pairs were also better at detecting their own illegal moves or errors while with single subjects it was often necessary for the experimenter to indicate that an illegal move was made.

The single subject's approach to solving the problem was to begin by moving the pieces directly with no evaluation of possible moves. Their verbal reports most often described what they were doing with less description of their plans for solving the problem. Single subjects rarely proposed a move, considered it and then carried it out. Instead the planning was done "on the fly", often with the subjects expressing a sense of frustration about solving the problem. When they ran into problems, they were less likely to begin the problem over. Consequently, when they did solve the problem they had no clear memory of the solution path, because of all the incorrect or backward moves along the way.

Pairs of subjects, on the other hand, were able to solve the problem much faster than they were likely to do it alone. The pairs talked about their moves as did the single subjects, but the nature of the talk was different. The pair's talk was concerned with which of a number of contemplated moves should be made. This type of talk (negotiation and planning) seemed to be productive in finding solutions to problems. The need to justify a move often led to reflection on a given move and an analysis on how it was likely to bring the problem solvers closer to the goal state. The second person also served as a monitor (Miyake, 1986), noting illegal moves and the lack of progress of a given approach. In the cases where pairs of subjects found themselves having difficulty at a particular step in the problem, they were more likely to start over from the beginning the whole sequence of steps. This contrasted to the behavior of the single subjects who would continue to look for a move that would lead them to the end. Thus when the pairs found a solution, they executed it from start to finish with few irrelevant moves. The relative ease at which two subjects solved the problem compared to that of the single subjects suggested that the interaction between the subjects was an important resource for problem solving.

Water Jars to Charged Particles in Zapworld

The second set of experiments reported here is based on the Luchins Water Jar Experiment (Luchins, 1942). In this experiment subjects solve a set of problems using one particular procedure, then find it difficult to give up using that procedure on other problems, even when it is less efficient or even ineffective. What role does reflection play in helping subjects flexibly move beyond a conventional solution to find a better solution?

We implemented an isomorph of the water jar experiments in InterLISP on a Xerox 1108 computer. In this version, which we called "Zapworld", the subject is shown a number of moving objects each with a certain amount of "charge." The goal is to accumulate a specified amount of charge by gaining charge from charged objects (by touching them with a mouse-driven pointer and pushing a button) and by losing charge to uncharged objects.

The problem set contains a sequence of 12 problems. The first 2 are example problems and the next five problems can all be solved by using a particular lengthy procedure. The next two problems (called critical 1 and 2) can be solved by the same lengthy procedure or by a shorter "direct-method" procedure. The 10th or "direct-method problem" can only be solved by the shorter procedure. Then two more problems (critical 3 and 4) were given in which either the long or direct method procedure could be used.

In the classic experiments by Luchins, only 19% of the subjects saw and utilized the direct method for solving the first set of critical problems. His subjects took much longer to solve the direct method problem and only 39% made the shift to the direct method of solution for the last two critical problems.

When subjects worked these same problems in Zapworld we found a surprising result. The subjects did not get blocked on the "critical problems" in the same way as the findings of Luchins would predict. When faced with the first set of critical problems, 63% shifted to the direct method immediately. After the "direct-method" problem 85% shifted to the direct method for the last two critical problems.

We began to use alternate problem isomorphs to understand this result. We used computer printouts of the problem to create a pencil and paper version of the task. Subjects were blocked in this format in a way that was similar to that found by Luchins although the finding was not quite as strong as he reported. When they reached the first set of critical problems 41% shifted to the direct method and 55% shifted to the direct method for the last two problems.

We also compared the strategies that were used by the subjects on paper and on the computer as they worked the first 7 problems.

Solving the problem on paper requires the subject to carry out the arithmetic involved in the problem (which is done automatically in the computer version). So subjects developed and applied a lengthy problem solving procedure that worked for the initial problems. Once this procedure was developed, they continued to use it until they found that it did not work. They were more likely to use the longer procedure over the direct method even after a problem that indicated that the procedure might not work in all cases. The paper and pencil subjects seemed to approach the problem in exactly the way one would predict from the original Luchins data.

The computer implementation of the problems weakened the learning and the automatic application of previously successful strategies. The implementation of the problem on the computer with the mouse pointer enabled the students to play with the problem and not necessarily attend to the computations that are the only strategy available to the person working with paper. The behavior of the subjects was similar to that we had seen in the single subjects' approach to the missionaries and cannibals problem. They could try things out by interacting with the problem, using visual cues to suggest the next action. It was more like what is frequently called the "discovery learning" approach to problem solving. The students tried a number of different strategies with each problem and one subject even discovered a new and original solution to the "direct-method" problem when working on the computer. Another subject was about to use the strategy that had been successful in the past but he made an error in discharging to the wrong particle. The new state created on the computer by this error suggested the direct method of solving the problem.

Since the work with missionaries and cannibals had pointed to the role of evaluation or reflection in problem solving, we decided to try a change in the procedure that would encourage the subjects to reflect on their problem solving approach. The Zapworld procedures were modified so that after each solved problem, the subject had to record how they had solved the problem. Once subjects were asked to record how they had solved the problems, they did appear to develop the same procedure as the subjects who worked on paper and this was confirmed by their performance on the the first 2 critical problems. On this problems, 70% of the subjects continued to use the long procedure. After the "direct-method" problem, we found another surprise. Almost all the subjects, (90%) shifted to the short method for the "direct-method" problem and the critical problems that followed it.

This finding suggests that reflection on the problem solving strategies resulted in a clearer development of a problem solving procedure that is applied for efficiency. But once that strategy becomes less effective, the subjects were able to shift to the direct method with no difficulty.

We were able to constrain the computer environment somewhat by making it a requirement that the subject record the steps that he or she took to find the solution after solving the problem. When forced to record the steps to solution, the problem solvers began to search for the mathematical procedure that could be used to describe the solution. The performance on the first two critical problems indicates that the subjects were more likely to experience the initial problem of being blocked but as soon as a problem required a direct approach, they were the most likely to try this new direct approach on the next set of critical problems.

Implications for Computer-Based Problem Solving Instruction

In both these problem domains we found that reflection on the moves that were necessary for finding the solution led to better problem solving. In the Missionaries and Cannibals, the subjects who worked cooperatively were placed in a situation in which they had to negotiate their moves. The cooperative condition made is necessary for the subjects to explain why a given move was likely to bring them closer to the solution of the problem.

In the Water Jars isomorph, Zapworld, we made it a condition of the task that the subjects stop and explain how they had arrived at the solution of a problem. This reflection on the solution helped focus the attention of the subjects on a productive strategy. Unlike the subjects in the classic Luchins experiments or the subjects that did the same task on paper, the subjects who reflected on the problem solution were able to shift to a new "direct-method" solution procedure with no difficulties.

These findings suggest why problem solving environments on the computer often allow students to do what looks like sophisticated problem solving, but these same students then fail to transfer these problem solving skills to problems in other domains. Research on the effect of Logo on students' problem solving skills has shown little transfer (Papert, Watt, diSessa & Weir, 1979; Pea & Kurland, 1984). If we could arrange for students to reflect on what they have accomplished in these domains, then we might see more flexibility in their application of these skills to other problem domains. Our results suggest that interaction with the computer in such settings might be more effective if there is a reflective stage in which students review what they have done on the computer and why. It may be that some of this takes place naturally when subjects work together on a problem.

This research has focussed on the important step of reflection on problem solving. It suggests that teaching students to analyze what they have done will help them develop flexibility in using a new approach when blocked.

References

Ernst, G. W., & Newell, A. (1969). GPS: A case study in generality and problem solving. New York: Academic Press.

Hutchins, E. L., & Levin, J. A. (1981). Point of view in problem solving. (Report CHIP 105). La Jolla, CA: Center for Human Information Processing.

Jeffries, R., Polson, P. G., Razran, L., & Atwood, M. (1977). A process model for missionaries-cannibals and other river-crossing problems. Cognitive Psychology 9, 412-440.

Luchins, A. S. (1942). Mechanization in problem solving: The effect of Einstellung. Psychological Monograph, Whole issues 54.

Martin, L. (1983). Children's problem solving as inter-individual outcome. Unpublished PhD Dissertation, La Jolla, CA: University of California, San Diego.

Miyake, N. (1986). Constructive interaction and the iterative process of understanding. Cognitive Science 10.

Papert, S., Watt, D., diSessa, A., & Weir, S. Final report of the Brookline Logo Project. Part II: Project summary and data analysis. Cambridge MA: Massachusetts Institute of Technology, Artificial Intelligence Laboratory Memo No 545.

Pea, R. D., & Kurland, D. M. (1984). Logo programming and the development of planning skills. (Technical Report No. 16). New York: Bank Street College of Education, Center for Children and Technology.

Reed, S. K., Ernst, G. W., & Banerji, R. (1974). The role of analogy in transfer between similar problem states. Cognitive Psychology 6, 436-450.

Suchman, L. Plans and situated actions: The problem of human-machine interaction. Palo Alto: Xerox Palo Alto Research Center Technical Report.

This research has been supported by the Personnel and Training Research Program, Office of Naval Research under Contract No. N00014-85-K0266. Thanks for Chris Pribe for his programming efforts.

Intelligent Tutoring Systems for Scientific Inquiry Skills[1]

Valerie Shute and Jeffrey Bonar

Intelligent Tutoring Systems
Learning Research and Development Center
University of Pittsburgh

Abstract

We describe the initial prototypes of several intelligent tutoring systems designed to build students' scientific inquiry skills. These inquiry skills are taught in the context of acquiring knowledge of principles from a microworld that models a specific domain. We have implemented microworlds for microeconomics, electricity, and light refraction. All of the systems are highly interactive; students can pose questions, conduct experiments by manipulating domain specific factors, and record results. Using protocol studies of expert and non-expert learners using these microworlds we identify important inquiry strategies. We have represented these strategies formally, allowing the microworld to detect effective and ineffective inquiry strategies. We conclude with a description of a partially implemented "inquiry coach". This coach will be incorporated into the microworlds and teach the inquiry strategies in the context of the specific microworld domain knowledge.

1. Introduction

How can we help students develop better scientific investigative behaviors? In this paper we describe the key features of several intelligent, interrogatable microworlds under development. These microworlds are designed to support students in learning scientific inquiry skills, as well as in learning the particular domain modeled in the microworld. Our systems make online inquiry tools available to students and generally foster student interrogation. In particular, our systems coach the student in the skills of

---

[1]This work was supported by the Learning Research and Development Center, supported in part as a research center by funds from the National Institute of Education (NIE), Department of Education. The opinions expressed do not necessarily reflect the position or policy of NIE and no official endorsement should be inferred.

conducting systematic investigations.

We and others in our lab have developed laboratory microworlds for the scientific domains of microeconomics, light refraction, and electrical circuits.[2] In each microworld the student is given the goal of learning the regularities and rules of the domain from observation and discovery. Initially, the student manipulates the environment and observes the results. As new knowledge is acquired, students can form hypotheses and conduct experiments to test those hypotheses. As we describe in detail later, students use the microworlds either to explore, experiment, or do exercises.

Work in our microworld environments is safe and immediate, without the typical complications of laboratory activities in the real world. For example, the microworlds are designed to eliminate much of the extraneous information and phenomena which hinder the discovery of regularities. The microworlds and their underlying simulations allow for instant feedback about the effects of varying relevant variables.

In the process of doing or learning science an individual makes observations, generates hypotheses, tests principles and laws, and predicts from theories. We are concerned with taking widely differing individual exploration strategies, and coaching our students to be more systematic and effective. Students differ on many dimensions and microworlds must be flexible enough to accommodate these differences. For example, prior related knowledge and systematicity of interrogative skills represent two important characteristics underlying successful learning in a microworld environment. Below, we describe work in progress on a exploration skills intelligent tutor.

Our paper is organized as follows. First, we describe a prototype intelligent interrogatable discovery environment: the economics microworld called *Smithtown*. Second, we describe the inquiry tools provided in our microworlds. Students use an interactive notebook, tabular displays, and graphical displays. Students also can propose hypotheses and examine a history of their explorations. Third, we describe the specific inquiry behaviors (skillful and otherwise) found in our protocols of novice and expert learners using *Smithtown*. Fourth, we describe a partially implemented intelligent tutor for the inquiry skills described, focussing our discussion on the diagnostician and the coaching component. Finally, a plan for evaluating the effectiveness of this research approach is discussed.

---

[2]The microeconomics microworld was developed by Valerie Shute and Jamie Schultz. The refraction microworld was developed by Peter Reimann [Reimann, 1986]. The electrical circuits microworld was developed by Jeff Bonar, Joyce Ivill, Cindy Cosic, Leslie Wheeler, Gary Strohm, and Paul Resnick. All these systems have been developed using LOOPS on Xerox Interlisp-D workstations.

2. Overview of the System

Our microworld for microeconomics is a hypothetical town called *Smithtown*. When the student sits down at the computer, s/he is introduced to this simulated town.[3] A series of menus pop up for the student to select from (see Figure 2.1). The first menu contains the markets currently available in *Smithtown*. The student uses this menu to select a good or service to investigate. The second menu is the 'planning menu', containing all possible variables that either may be changed by the student, or that change as a function of something else. The student must state the variables they are interested in investigating. This assists the system in understanding and classifying student activities. Third, a menu with relevant economic indicators for *Smithtown* appears. These indicators include average income, population, weather, consumer preference index, and number of suppliers. Each of these indicators has a system supplied default value (e.g., population = 10,000). The current value for each of these indicators is shown on the screen.

After the student examines and/or modifies the indicators, s/he sees the 'prediction menu'. If desired, the student may use this menu to state the variables and relationships when predicting the outcome of an event. Finally, the student is presented with a 'Things to Do' menu where s/he can see the effect of market manipulations on price, quantity demanded, quantity supplied, surplus, or shortage. Using this information the student may:

1. Adjust the market price or have the computer make a price adjustment,

2. Use the inquiry tools to assist in the investigation (e.g., make a notebook entry of the market data), or

3. Select an experimental framework.

The experimental frameworks let the student manipulate the market in various systematic ways and observe the effects. For example, a student might want to generalize a concept across goods. This would allow him/her to see how widely a concept applies. The framework to accomplish this generalization is: *Change the good, keep the same independent variables.*

As students interact with new subject-matter situations, they compare their observations with their current beliefs and theories. Consequently, these beliefs may be rejected, accepted, modified, or replaced. In the course of this developing knowledge students ask questions, make predictions, make inferences, and generate hypotheses about why certain events occur with systematic regularity. In *Smithtown*, the results of this student interaction with the environment are immediate, dynamic, and recordable. We do this with a set of inquiry tools, described in detail in the following section.

---

[3]Before actually interacting with the system, the student receives a short guidebook to Smithtown, outlining the purpose, set up, and terminology of the system.

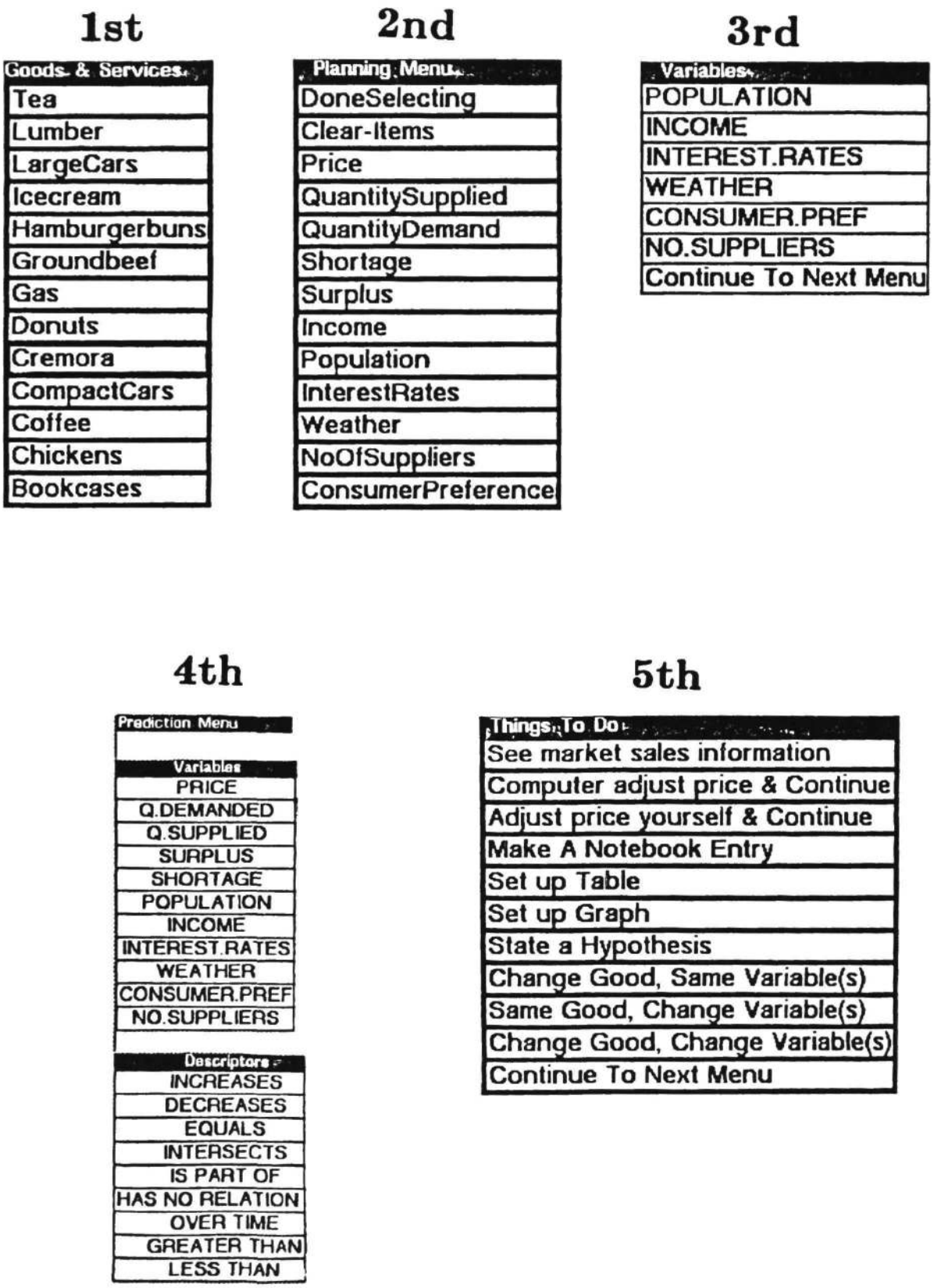


Figure 2-1: FLOW OF MENUS

3. Tools for Systematic Investigations

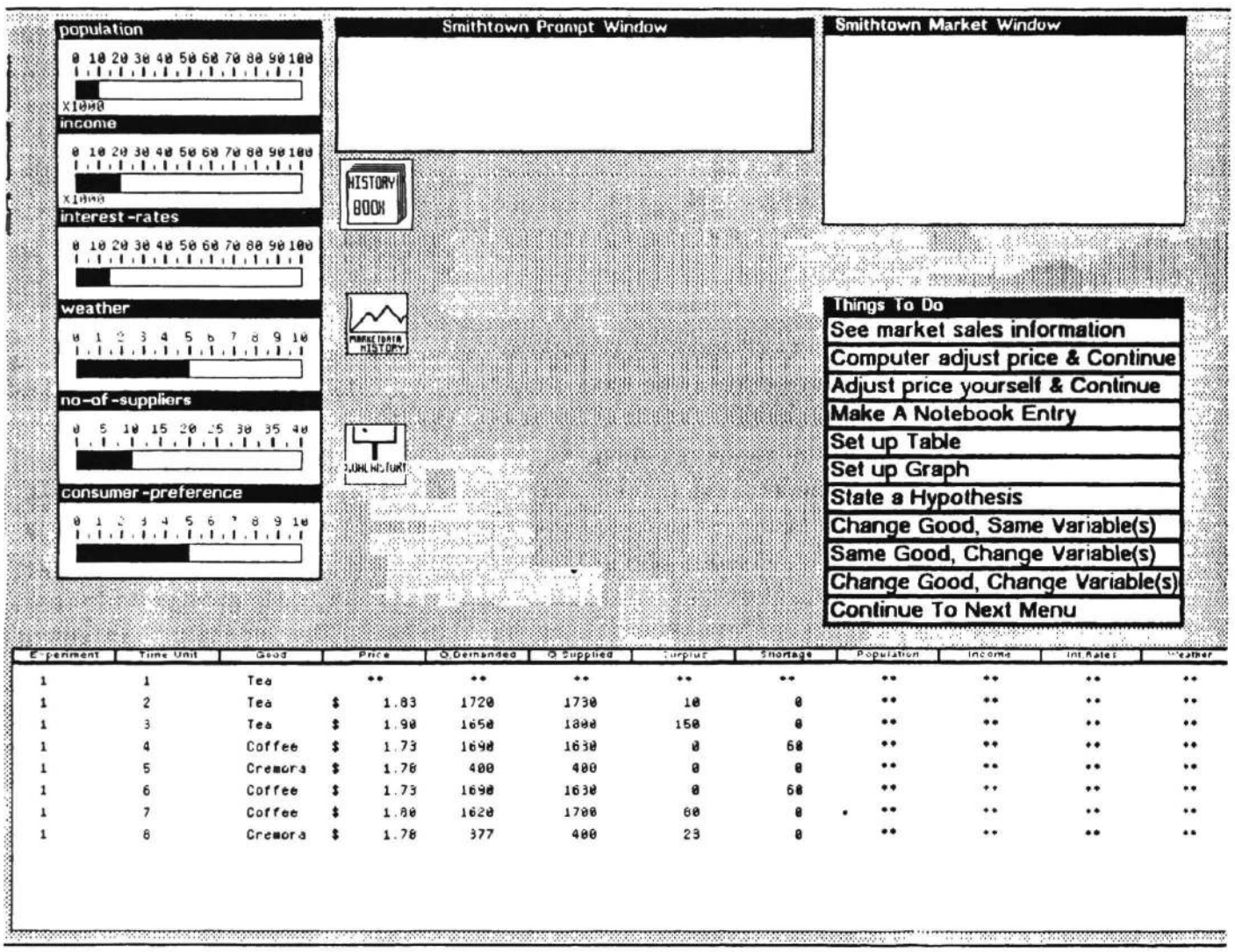
We have several online tools for scientific investigations in the microworld environments. These include: a *Notebook* for collecting data and observations, a *Table* to organize data from the notebook, a *Graph* utility to plot data, a *Hypothesis menu* to compose relationships among variables, and three *History windows* that allow the students to see a chronological listing of behaviors, data, and concepts learned so far. Each of these will be discussed in turn.

An example of the Notebook is shown in Figure 3.1. The students select variables to record, and current values are automatically put into the Notebook. Once they have collected data in the Notebook, they can elect to isolate some of the variables and put them together into a Table. Tables provide sorting tools for reordering the entries (see Figure 3.2) This is an important tool for simplifying and making sense of raw data in the Notebook. The Graph utility allows a student to plot data collected from their explorations/experiments. This provides an alternative way of viewing relations between variables. An example of a Graph is shown in Figure 3.3.

The Hypothesis menu (Figure 3.4) allows students to make inductions or generalizations of relationships from the data they have collected and organized. There are actually three interconnected menus comprising the Hypothesis menu. First, the *connector menu* includes the items: if, then, as, when, resulting in, and, the. Next, the *variable menu* contains the economic indicator variables used by the system: income, population, quantity demanded, demand, quantity supplied, supply, market price, surplus, shortage, and so on. Finally, the *descriptor menu* describes the types of change: decreases, increases, equals, intersects, is part of, has no relation to, is greater than, and is less than. As students choose words from these menus, the emerging statement appears in the Hypothesis Statement Window.

A pattern matcher analyzes key words from the input and checks whether this matches stored relationships for each targeted concept. If so, the system flags that concept as having been conditionally learned. Otherwise, the student is informed that the statement is not understood.

Three history windows are included in the system. As students continue to interact with the microworld, histories accumulate summarizing the various actions resulting from different explorations and experiments. This summary is maintained in the Student History Window. The Market Data Window keeps a record of all variables and associated values that the student has manipulated. Finally, there is the Goal History Window. This provides a chronological representation of what the student has successfully learned in terms of concepts targeted by the system.



Experiment	Time Unit	Good	Price	Q Demanded	Q Supplied	Surplus	Shortage	Population	Income	Int.Rates	Weather
1	1	Tea	**	**	**	**	**	**	**	**	**
1	2	Tea	$ 1.83	1720	1730	10	0	**	**	**	**
1	3	Tea	$ 1.90	1650	1800	150	0	**	**	**	**
1	4	Coffee	$ 1.73	1690	1630	0	60	**	**	**	**
1	5	Cremora	$ 1.70	400	400	0	0	**	**	**	**
1	6	Coffee	$ 1.73	1690	1630	0	60	**	**	**	**
1	7	Coffee	$ 1.80	1620	1700	80	0	**	**	**	**
1	8	Cremora	$ 1.70	377	400	23	0	**	**	**	**

Figure 3-1: NOTEBOOK WITH DATA RECORDED

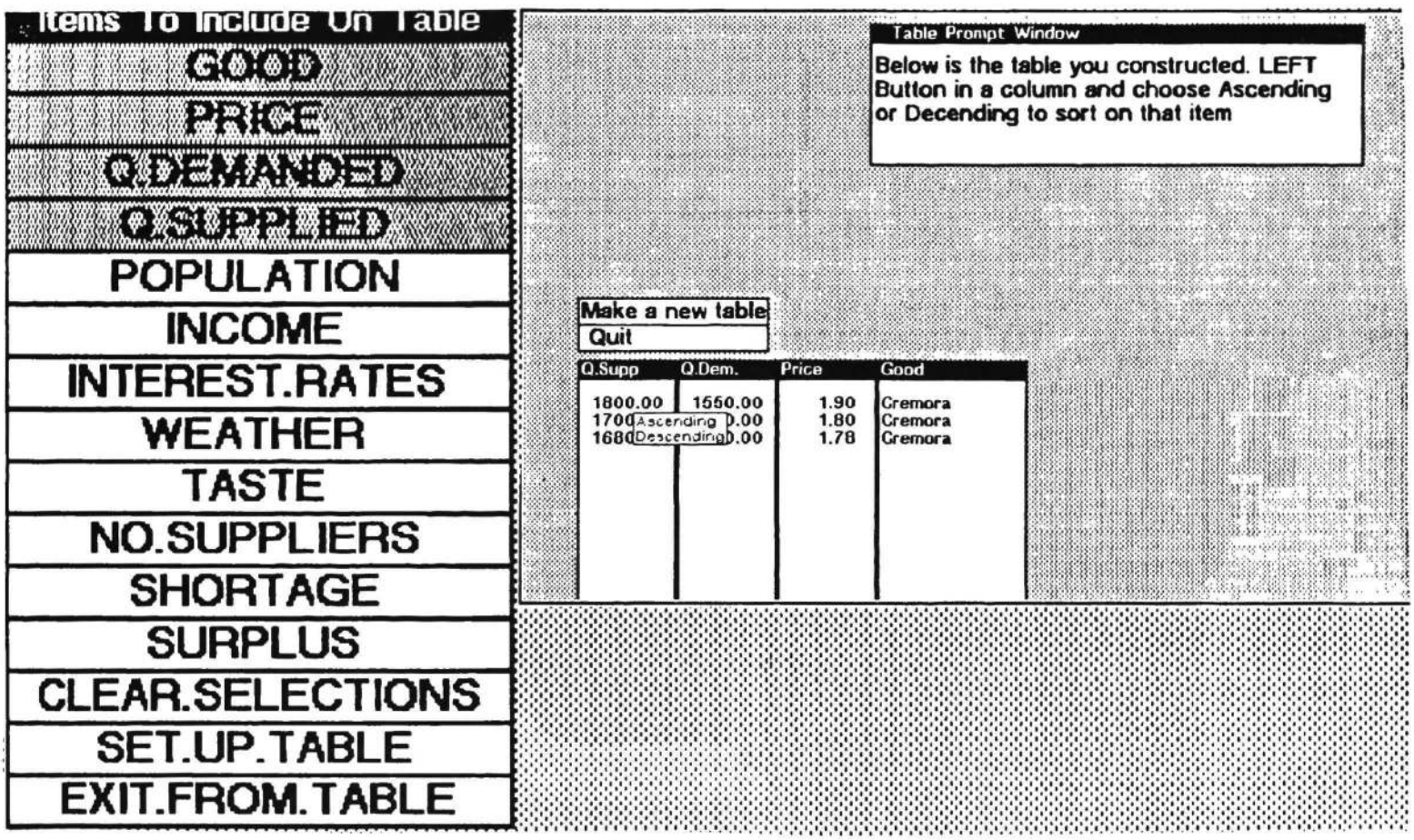


Figure 3-2: TABLE WITH FOUR VARIABLES REPRESENTED

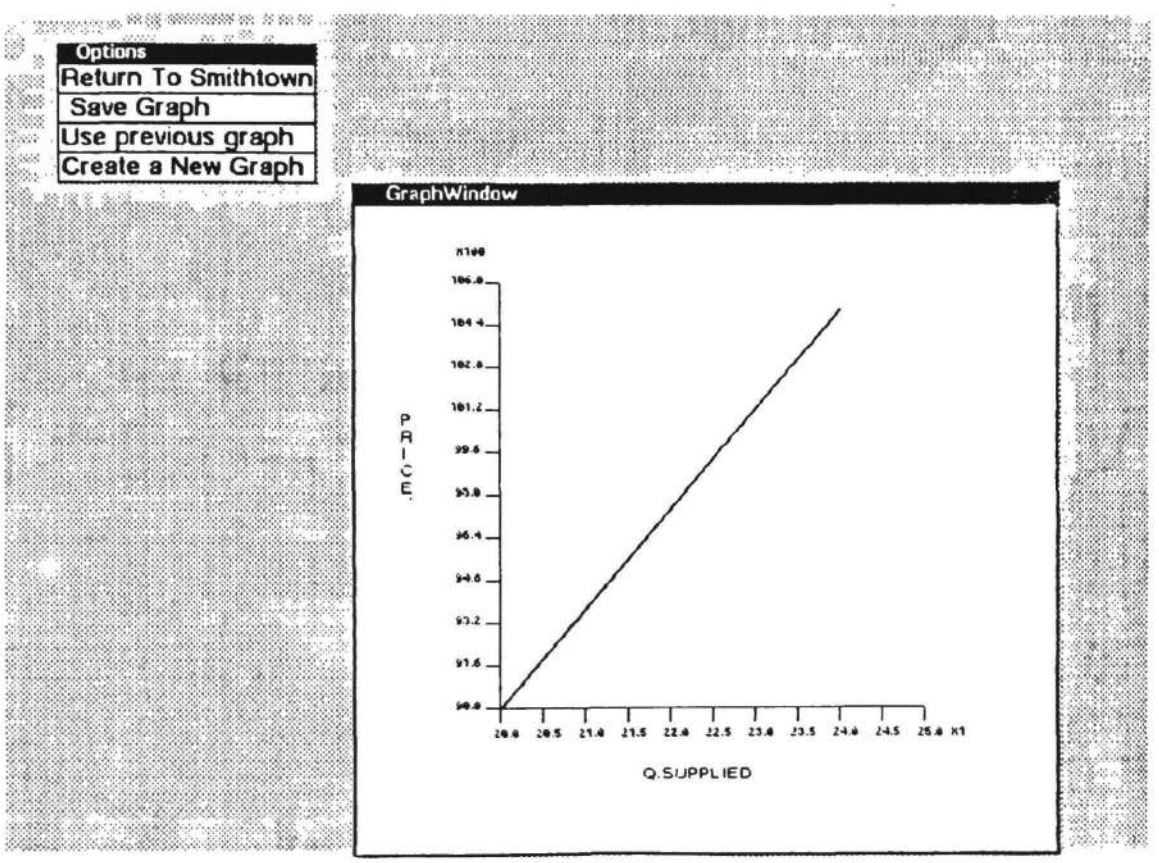


Figure 3-3: GRAPH WITH PRICE AND QUANTITY SUPPLIED PLOTTED

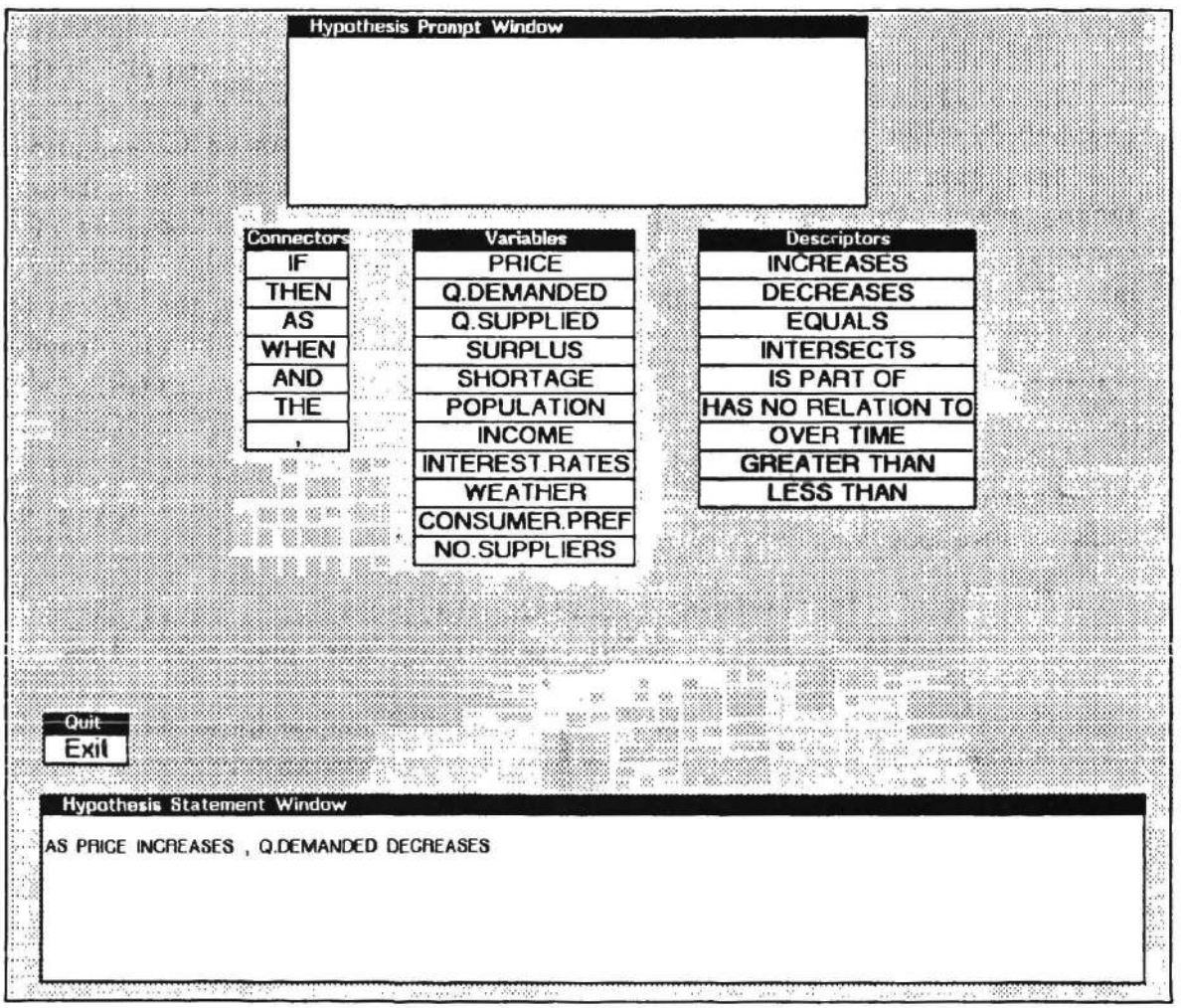


Figure 3-4: MENUS FOR HYPOTHESIS GENERATION

4. Investigative Behaviors

The scientific inquiry tools provide students with means for collecting, organizing, and understanding data from their investigations. Moreover, there are several types of systematic investigations recognized by the system. These include *explorations*: obtaining information from the microworld in order to refine and complete developing hypotheses about the microeconomic concepts; *experiments*: a series of student actions conducted to confirm or differentiate hypotheses; and *exercises*: tests on a previously confirmed hypothesis, perhaps to see the extent or limitations of its application. Experiments carry a specific prediction while explorations do not.

The distinction into the type of investigation being conducted allows the system to determine whether tutorial intervention will occur. That is, a student is considered in *exploration* mode if s/he has no prediction of outcomes from an investigation. Ignoring the 'prediction menu' would result in this classification. There is no assistance from the inquiry coach while in this mode, unless the student remains too long in it (e.g., 20 consecutive actions without making a prediction). If a student predicts an experimental outcome, we consider her/him in *experimental* mode and allow the system to intervene with the student. Finally, if the student is replicating the results from a previous experiment, s/he is classified as being in *exercise* mode and no intervention will occur.

Using the tools provided, there are various dimensions on which student investigative performance can be evaluated. The following is a listing of scientific behaviors we look for:

- *Baseline Data Collection*: For an initial exploration, is data collected from the market in equilibrium, before any variables have been altered?

- *Baseline Data Entry*: For an initial exploration, is data entered into the notebook from the market in equilibrium, before any variables have been altered?

- *Thorough Data Entry*: Does the student make a notebook entry every time a variable has been changed?

- *Relevant Data Collection and Entry*: Does the student enter only those variables that are being actively manipulated or change as a result of a manipulation (i.e., no superfluous information or incomplete recordings)?

- *Generalize a Concept Across All Goods*: Did the student try to generalize an economics principle as it holds for all goods? For example, having learned the law of demand operating in the donut market, does the student attempt to generalize to other goods?

- *Generalize a Concept Across Related Goods*: Did the student try to generalize an economics principle across specifically related goods, i.e.,

investigating substitute/complementary relationships or component-of relations?

- *Sufficiently Large Change to Variables*: Are the changes made to the variable(s) large enough to detect market effects if there are any?

- *Sufficiently Small Change to Variables*: Are the changes made to the variable(s) small enough to discriminate and refine subtle patterns in the data?

- *Number of Variables Changed*: Did the student change only one variable at a time for comparison and/or recording? This is related to what a student already knows; over time, a student can progressively handle more variables.

- *Isolating Variables in the Table*: Did the student put only a few variables into the table to reduce and make sense of the data?

- *Sorting on Relevant Variables*: Was the sorting option used on relevant variables in the table (e.g., if price was systematically varied, then it should be the sort key.)

- *Plotting Variables*: Did the student use the graph utility to plot potentially meaningful relationships between variables?

- *Saving Graphs*: Was the graph of a significant relationship saved for later comparisons?

- *Superimposing Graphs*: Was there an attempt made to superimpose two graphs to see relationships between functions, like supply and demand, or two demand curves in parallel?

- *Hypothesis Specification*: Were any hypotheses stated as a result of the observed systematicities in the data, either abstracted from the Table or from the Graph?

- *Complexity of Hypotheses*: Is there an increase over time in the chaining of variables when the student generates hypotheses from the menu? As knowledge increases, the number of variables strung together should go from two to more, in progressively more complex relationships.

- *Passively Monitoring the Market-- Computer Price Change*: Did the student allow the computer to make adjustments to the market price of a good and see the ensuing repercussions? For explorations, monitoring is probably the best strategy to follow.

- *Actively Manipulating the Market-- User Price Change*: Did the student more actively make adjustments to the market price of a good and see the market changes? For experiments, manipulating the price oneself is probably more effective.

- *Failure Driven Behaviors*: If a student conducts an experiment testing some specified belief/prediction, the results can lead to a number of different actions. If the experiment is confirmed, there may follow a generalization attempt, or an exercise, both of which test the limitations of the idea. If the experiment is disconfirmed, the student may do one of the following:

    - Re-do the experiment with parameter changes.

    - Ignore the results and go on to something new, or

    - Try a new hypothesis that fits the observed data.

5. An Intelligent Tutor for Inquiry Skills

We are currently implementing an intelligent tutor for the inquiry skills discussed above. Our tutor for *Smithtown* is organized to provide a range of guidance that can be gradually increased or decreased, depending on the characteristics of the learner's performance. At one end of the range is a purely discovery environment. As long as the student is progressing, the microworld will remain a discovery environment. "Progress" is defined as (a) demonstrating appropriate investigative behaviors, and (b) learning the domain concepts at a regular rate. At the other end of the range is a directive environment that explicitly assists the student in using an interrogative skill that is deemed problematic for him/her. For example, the system might instruct a student to enter data in the notebook from the market before altering any variables.

The system is composed of four main components: the knowledge base, the diagnostician, the student model, and the coach. The *knowledge base* includes the targeted elements to be learned, such as the "law of demand", from economics domain knowledge, and "generalization of a concept", from scientific skill knowledge. The *diagnostician* is a set of software *critics* that monitor the student's success in applying the inquiry behaviors and learning the domain concepts. The *student model* is the updated representation of the student's evolving knowledge base, and the *coach* instructs the student based on information provided to it by the diagnostician regarding deficient skills.

Tutor knowledge about the domain concepts is kept in the *knowledge base.* Using the knowledge base, the tutor models whether a student is progressing in the acquisition of domain knowledge. Although a student may be quite efficient in interrogating the environment and performing experiments, s/he may extract only a subset of the relevant concepts for instruction in the particular domain. To accomplish this assessment, the

instructional domains for each microworld have been broken down into key concepts that are loosely organized in a bottom-up manner, from simpler to more complex ideas (see Figure 5.1) (see [Bonar, Cunningham, Schultz, 1986] for a description of this approach).

Although the microworld need not constrain the student to learn the concepts in any prescribed order, the conceptual hierarchy provides the basis for a model of student knowledge. Each concept is associated with a rule or relationship among economic variables. For example, the law of demand relates *price* and *quantity demanded* in an inverse relationship. When a student uses the hypothesis menu and generates a valid statement about the underlying variable relationships for a particular concept, then the system flags the concept as having been learned. This is called an *overlay student model* ([Carr and Goldstein, 1977]). The student is then provided with a congratulatory statement by the system, including the proper name of the concept. For instance, the system may respond to a student having just specified the law of demand (i.e., as price increases, quantity demanded decreases), with *"Congratulations, you have just discovered what economists refer to as the Law of Demand. Please note that the converse is also true. That is, as price decreases, quantity demanded increases. Now please conduct an experiment that illustrates this phenomenon."* This request for an experiment guarantees that the student possesses not only declarative knowledge about the concept in question, but also procedural knowledge about how to construct an experiment to demonstrate it.

The *diagnostician* evaluates a student's interrogation of the system. It determines whether a student is proceeding in a systematic, efficient manner. To do so, a comparison is made between the student's actual behavior and optimal behavior (much as was done with the WEST tutor [Burton and Brown, 1982]). For example, a student might consistently alter multiple variables at the beginning of each experiment. This obscures any conclusions that might be drawn from the resulting market conditions. This problem area would then be noted by comparing the student's behavior to the ideal behavior where only one variable is changed.

Each scientific behavior is translated into a *critic* consisting of specific student actions or conditions to be met. A match is attempted and the student model is updated accordingly. For example, the behavior: "Sufficiently Large Change to Variables" addresses the questions: *Are the changes made to the variables sufficiently large enough to detect any market effects if there are any?*

The critic, translated into English, is shown in Figure 5-2.

In conjunction with each critic are actions where the student might go astray. These define the *buggy versions*. In Critic 6, the asterisk shows the line where students

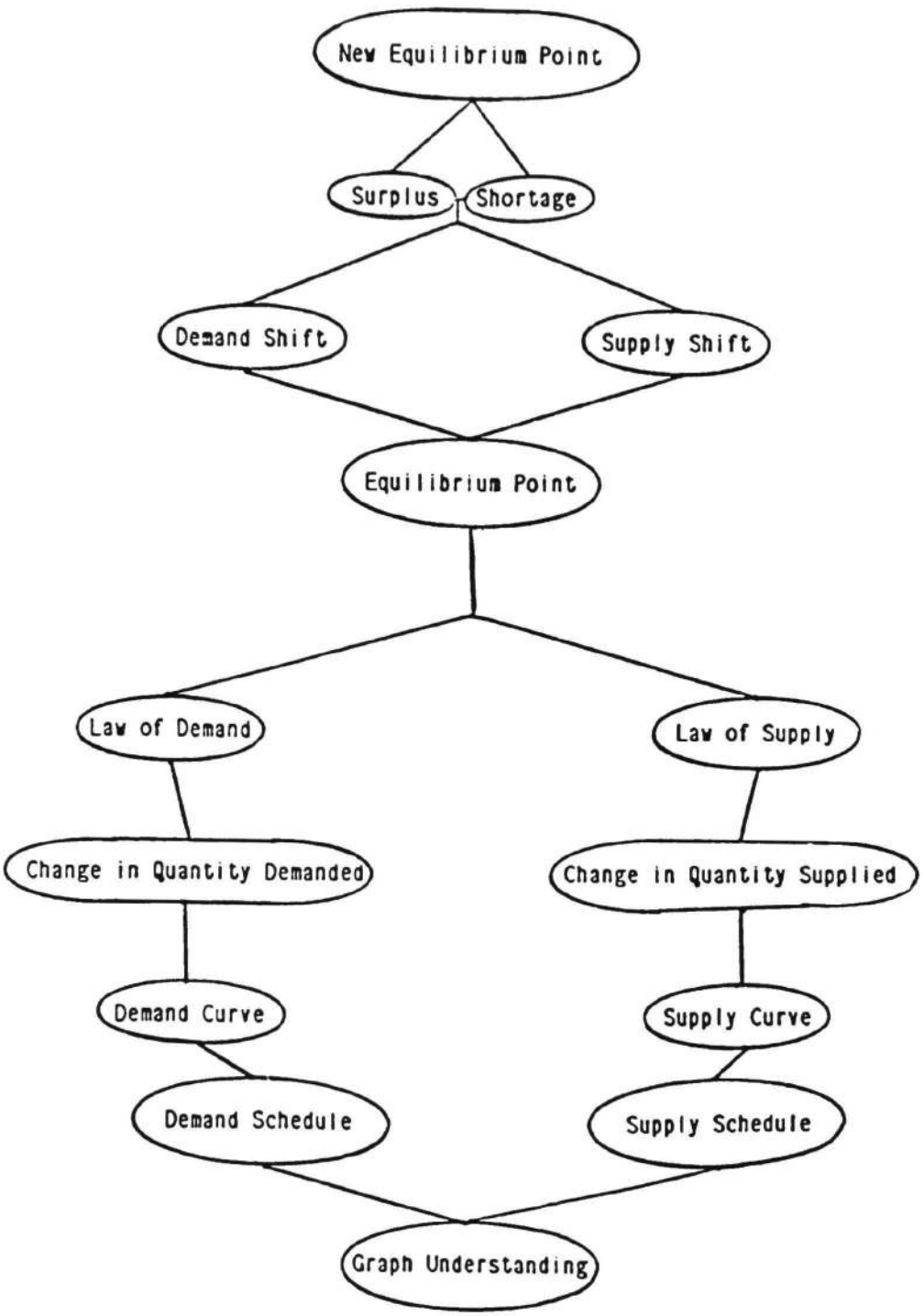


Figure 5-1: HIERARCHY OF SUPPLY AND DEMAND CONCEPTS

Critic 6

IF: For any experiment, AND
for any time number, AND
for any good selected, AND
from variable-menu there is a list of variables changed, AND
* the degree of change is at least 10% of the total range
(maximum minus the minimum values)

THEN: The behavior for 'sufficiently large change to variables' has been demonstrated.

ELSE: Alert the coach for helping the student, depending on the value of the Critic6-counter, which provides information on the number of times the student has erred on this skill.

Figure 5-2: A Critic For "SUFFICIENTLY LARGE CHANGE TO VARIABLES"

Explanation 1 (vague):

If you make a change, you should try to stack the deck. In other words, do something regarding the size of the change so that you'll be able to actually see any effects in the market place if there really are any.

Explanation 2 (analog mapping):

If you entertained the notion that having a lot of hats made a person more self-confident, you would probably compare the self-confidence of a group of people with no hats (or a small number of them) to those having many hats, where "many" was some large enough number to see if your hypothesis was correct. Apply the same logic here.

Explanation 3 (explicit):

You increased the <variable name> by only <amount>. Double it now and check out how the market changes for the values of quantity demanded and quantity supplied.

Figure 5-3: LEVELS OF EXPLICITNESS FOR REMEDIATING CRITIC 6

might err.[3]

The results of the diagnosis are passed to the coach. If there is sufficient evidence of inappropriate student strategies or floundering, the coach will interrupt the student. This coach does not discuss the general inquiry skill directly. Instead, it addresses the issue in the context of the student's current investigation in the microworld. Intervention will occur after the coach receives several corroborations of deficient performance and if the student is in *experiment mode*. "Deficiency" is defined as either not using a behavior when it was appropriate, or using a buggy version of that behavior. A "batting average" is computed for each behavior consisting of the number of times the skill was used divided by the number of times it should have been used. If that number is less than some threshold (such as $< 50\%$), this will prompt a response by the coach. Also, if there are several deficient student behaviors, the tutor can address each, one at a time. This involves a hierarchy of coaching where, for example, the behavior for *Baseline Data Collection* takes precedence over *Save a Graph*, and so on.

To encourage an independence of thinking, the tutorial assistance will progress in terms of explicitness in a 3-part range: (1) Vague initially, with abstract directions and examples for the student, (2) Clearer, with analogies provided to the student, and (3) Very clear, with explicit instructions for conducting a particular experiment. For the example given above, the buggy version of this critic would address the size of the change to the variable. Figure 5-3 shows the three responses.

The preliminary principles calling for interrupting a student's activity are:

- Student weakness must be apparent. What is an "apparent" level will initially be an arbitrary threshhold ratio, to be refined after experimentation.
- Feedback should not be always critical, but also appreciative of any good aspect of a student's behavior.
- Any instruction or guidance should be spaced. That is, at least a gap of two or three events between interruptions.

6. Evaluation of the System

To evaluate whether the system is effective in enhancing students' learning scientific inquiry skills, we will need to look at transfer effects. That is, a student going through one microworld environment learning microeconomic principles should consequently transfer their interrogative skills in the investigation of a new domain, say, geometric optics.

---

[3]Performance data is kept as a proportion of actual use of behavior versus when the behavior should have been used.

Studies are planned to assess the degree to which students apply inductive reasoning skills within and across subject-matter domains. In the context of the various microworlds, a series of near and far transfer situations are being designed, drawing on recent literature on analogical reasoning, cognitive mapping, and rule assessment procedures (e.g., [Cheng & Holyoak, 1986, Gentner, D., 1983, Gick and Holyoak, 1983, Siegler and Klahr, 1982]). In addition, recent information will be considered from artificial intelligence research on learning systems that attempt to improve their own performance in inferential reasoning and problem solving ([Michalski, Carbonell and Mitchell, 1983]).

We are in the process of running subjects (college freshmen) on the system. None of the subjects will have any formal economics training. The major focus will be on how this group interacts with the system, analyzed on two levels: (1) Whether students become more facile across experimental sessions[4] in using the scientific tools/behaviors in the microworld, and (2) If the students actually learned any of the targeted microeconomic concepts. The first question will be answered by protocol analyses of subjects' justifications of actions, as well as detailed histories of student actions kept by the microworld. The second question will be answered from an analysis of the difference scores between pre- and post-tests.[5] Performance with regard to the solution of the complex scenarios will allow us to see individual differences in knowledge representation for economic phenomena. This will compare economic system understanding before and after interaction with the microworld.

7. Concluding Remarks

Research on scientific inquiry learning will contribute to the teaching of subject-matter content with interrogation and inferential skills. Our microworlds provide simulated experimental environments and allow for student interrogation of scientific phenomena by making observations, organizing the data obtained, formulating explanatory hypotheses, and testing experimental predictions. The microworlds incorporate a set of tools which support a student's inquiry process by making these scientific processes visible and manipulatable. Our microworlds allow student patterns of scientific reasoning to be monitored, assessed, and coached. This research should give us information for new methods in science instruction and contribute to knowledge of how to teach higher-order problem-solving skills.

---

[4]Note: the design involves four experimental sessions with the tutor, two hours each session. These occur every other day.

[5]Two types of objective tests have been developed for this purpose: a) multiple choice and short answer test of economic concepts, and b) complex scenarios requiring solution of 'what would happen if...' questions.

8. Acknowledgements

The authors wish to thank Jamie Schultz for his excellent work in developing the *Smithtown* system. Paul Resnick developed an earlier version of the system. Robert Glaser, Peter Reimann, and Kalyani Raghavan have contributed to the research reported here.

9. Bibliography

Bonar, Jeffrey G., Cunningham, Robert, and Schultz, Jamie. (1986, September). An Object Oriented Architecture for Intelligent Tutoring. In *Proceedings of the ACM Conference on Object Oriented Programming Systems, Languages, and Applications.*

Burton, Richard R., and Brown, John Seely. (1982). An Investigation of Computer Coaching for Informal Learning Activities. In Sleeman, Derek and Brown, John Seely (Eds.), *Intelligent Tutoring Systems* (pp. 79-98). London: Academic Press.

Carr, Brian and Goldstein, Ira P. (1977). *Overlays: A Theory of Modelling for Computer Aided Instruction* (AI Memo 406). MIT AI Lab.

Cheng, P.W., & Holyoak, K. J. (1986). Pragmatic reasoning schemas. *Cognitive Psychology,* , .

Gentner, D. (1983). A theoretical framework for analogy. *Cognitive Psychology, 7,* 155-170.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15,* 1-38.

Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (1983). *Machine learning: An artificial intelligence approach.* Palo Alto, CA.: Tioga.

Reimann, Peter. (1986). *REFRACT: A Microworld For Geometrical Optics.* Unpublished manuscript. Unpublished Technical Report.

Siegler, R. S., & Klahr, D. (1982). When do children learn? The relationship between existing knowledge and the acquisition of new knowledge. In R. Glaser (Ed.), *Advances in instructional psychology* (Vol. 2) (pp. 121-211). Hillsdale, NJ: Erlbaum.

Focus and learning in program design

Robert S. Rist

Cognitive Science Program
Department of Psychology,
Yale University.

ABSTRACT

The construction and retrieval of plan schemas was studied using a programming task. Novice programmers were asked to write programs and their problem solving strategies analysed from the manner in which they expanded a program goal. Four programs were used in the study, comprising two sets of problem isomorphs. These isomorphs required the same programming plans for solution, but differed in their cover stories. Plan development within a problem solving episode and across episodes was thus easily tracked.

Development of a plan to achieve a problem goal showed a focus of attention on the current goal to the exclusion of concerns about other parts of the program. A solution was not found by a process of reasoning and top-down design. Rather, one goal was selected for expansion and in the process of this expansion new goals were discovered and solved to the extent required by the current context. The program was built up from solving individual goals, not down from the problem specification. The development of plans to achieve three of these goals, the selection, simple (non-looping) sum and read loop goals, is discussed in some detail to show the relation between problem solving and schema formation. Evidence of top-down design using schemas appeared only late in the study, suggesting that its use in teaching should be delayed until the novice can 'speak the language' of schema retrieval.

INTRODUCTION

A generative model of program construction is presented that tracks the development of program plans in novice programmers. The outstanding feature is a focus on the current problem under consideration, the current goal. This forces a local view of the program, but creates the pieces that are required to implement the goal. It also creates many bugs (Spohrer, Soloway and Pope, 1985) when these pieces need to be later combined. Novice behavior is best described at this local level to capture the relative isolation of knowledge that excludes consideration of facts outside the current context. Systematic, top-down development is not useful for initial problem solving, since the pieces of mentioned knowledge do not yet exist. It is achieved by novice programmers only after prolonged and laborious effort.

The focus during initial problem solving is on how to create plans to achieve the program goals. One goal at a time is selected and a plan constructed within the context of this current goal. A focal idea is expanded and tested to see if it satisfies the goal (Kant, 1985; Sussman, 1975). In the process of expansion, new goals are discovered and implemented to the extent required by the current context; the program is grown from these seeds rather than designed. Once a plan has been formed, it can be modified or moved to satisfy other constraints, such as data flow, control flow or efficiency. Once many such plans have been constructed, the focus of attention can shift from local details to comparing plans and plan segments. Planning, which uses plans as stable conceptual chunks, replaces local problem solving. Schema retrieval and tailoring replaces schema construction.

The development of three plans is described in detail to show a local focus that is gradually expanded. The simple (non-looping) *sum* plan adds a set of values to find their total; its development illustrates the selection of objects and their combination. The *selection* plan shows the change from focus on a single goal at a time to a combination of related goals. This plan

divides the input into a set of categories so that the correct calculations can be made for each category. The *read* plan shows the creation of a sequence of goals from the focal goal. The need to read a set of data items creates the new goal of looping; the use of a WHILE statement for looping creates the need to deal with loop termination. Effects of focus apart from any particular plan schema were seen in general problem solving methods, where new information was integrated (incorrectly) in terms of the focal goal.

DESCRIPTION OF THE STUDY

Design.

A protocol study was used to analyze the program construction strategies of novice programmers. Subjects were students beginning an introductory course in Pascal at Yale University. The study began in week three of the course and subjects were tested as soon as possible after learning a programming construct, such as IF-THEN-ELSE or WHILE, before they had the opportunity to use the construct in assignments. Eight novices began the study, but one dropped out after the second week. Each remaining subject wrote four programs, two sets of program isomorphs that required the same plans but used different cover stories.

The first isomorph required selection and sum plans (Soloway and Ehrlich, 1984); it was presented in weeks three and five of the course, with a read loop added for selection in week five. The second isomorph was presented in weeks four and six. It required looping, running total, running count and maximum plans; the selection plan was used to count the input values in a set of categories. The same plans (read, sum, select) were used for all programs in the study, allowing any shift from plan construction to retrieval to be easily seen. The data base consists of 30 programs using the same plans in different contexts, about 50 hours of protocol data.

Analysis.

The basic structure of plan generation was shown by the changes a plan underwent in development. The plans were identified using the Bug Catalogue developed at Yale (Spohrer, Pope, Lipman, Sack, Freiman, Littman, Johnson and Soloway, 1985). The catalogue lists several hundred bugs taken from over 200 programs, grouped according to the goals and plans in which the bugs occurred. Within this goal/plan context, it classifies a bug in terms of what is wrong with the plan (malformed, missing, spurious, etc.), a descriptive taxonomy. The bugs have been described here in terms of a process model of their conceptual, developmental structure, to show how they arise during problem solving.

The different stages through which a plan evolved was seen within a single program. They were also seen across programs, where the plan was re-created or retrieved when needed. The same development strategies were seen in all subjects, but the first form of the emerging plan differed markedly across subjects. Thus all forms of the plan were not seen in all subjects. The analysis presented is a composite of stages based on partial overlap between subjects.

PLAN DEVELOPMENT

When a new problem is first encountered, there is no existing schema to guide solution or attention. The problem solving is guided by the problem information given (objects and values) and the simple operators (verbal, logical or mathematical) that can be used to combine this information. The initial focus is on identification of these objects and operators, then on their combination and finally on factors outside the locus of attention. The solution is driven from the most basic pieces outward. The current plan is tested to see if it achieves the goal and a solution accreted around it, with multiple cycles of identification, expansion and testing if necessary.

The problem fragment that required these pieces is shown below, in the form of the first

isomorph, the first problem in the study. The second isomorph presented it as a problem of calculating Welfare benefits for unwed mothers, where the amount of additional benefits decreased as the number of children increased:

> An electric company charges its customers by the kilowatt hour (kwh) for electricity used. The cost per kilowatt hour decreases as a customer uses more electricity according to the following rate schedule:
>
> ```
> 9 cents per kwh for the first 350 kwh
> 5 cents per kwh for the next 275 kwh
> 4 cents per kwh for the next 225 kwh
> 3 cents per kwh for all kwh over 850 kwh
> ```

The sum plan.

Two main variants of the sum plan were seen, based on development of the schema from a focus and then on the retrieval of this schema and the expansion of its slots. The initial problem solving effort is centered on the rate that must be calculated for *this* category, on the rate for the selected range. Subsequent expansion of the schema treats the new rate as a formula to be added to the existing framework. Six stages were seen in the development of the sum plan. The first four of these may be termed plan construction, since they involve a large amount of problem solving by the novice. Each stage reflects the problem solving that has occurred, from a very simple analysis to one that integrates all of the information required for a correct solution. The fifth may be viewed as a retrieved solution that still retains the form of its construction. The final (expert) form is the complete schema, divorced from its ontogeny.

1. *Straight rate*: the surface form of the description creates the final code. The focus is on calculating the charge for this category. At this stage, the object has been identified (kwh), the existence of separate rates and categories noticed and the rate has been attached to the object. The rate is then directly calculated as

   $$cost1 := kwh * .09$$
   $$cost2 := kwh * .05$$

2. *New + old simple*: the surface form of the sum is calculated. In this stage, the term 'next' has been parsed and it is realised that the different rates must be added to get the final charge. The focus is on the rate for the category and so the new rate is coded first, then the other pieces without further analysis as

   $$cost1 := kwh * .09$$
   $$cost2 := kwh * .05 + kwh * .09$$

3. *New + old local*: the correct form for the new rate is calculated. In this stage, the subject has realised that the rate changes over boundaries and only the amount of kwh over the boundary should be included in the new rate calculation. The focus is still on the new charge and so the insight is not extended to both charges:

   $$cost1 := kwh * .09$$
   $$cost2 := (kwh - 350) * .05 + kwh * .09$$

4. *New + old constructed*: the correct form of the schema is constructed. In this stage, the boundary insight is applied to all the individual rates, but the subject has explicitly coded one of the previous stages or has to devote considerable energy and time to constructing the solution. A particularly strong block to the discovery of the correct solution is that the old calculation does not involve the program object, kwh. The object has to be removed and a constant inserted in its place:

*cost1 := kwh * .09*
*cost2 := (kwh - 950) * .05 + 950 * .09*

5. *New + old retrieved*: the schema is retrieved, but still shows the new part focus. In this stage, the parts appear without apparent effort and the correct solution is mentioned as it is coded:

*cost1 := kwh * .09*
*cost2 := (kwh - 950) * .05 + 950 * .09*

6. *Old + new retrieved*: the *sum* schema is retrieved and expanded in the 'logical' order. This final stage shows the problem solving automated and attached to the components within the overall schema. The schema is retrieved and control given to the components in the order old plus new, where the old part repeats through the number of steps required:

*cost1 := kwh * .09*
*cost2 := 950 * .09 + (kwh - 950) * .05*

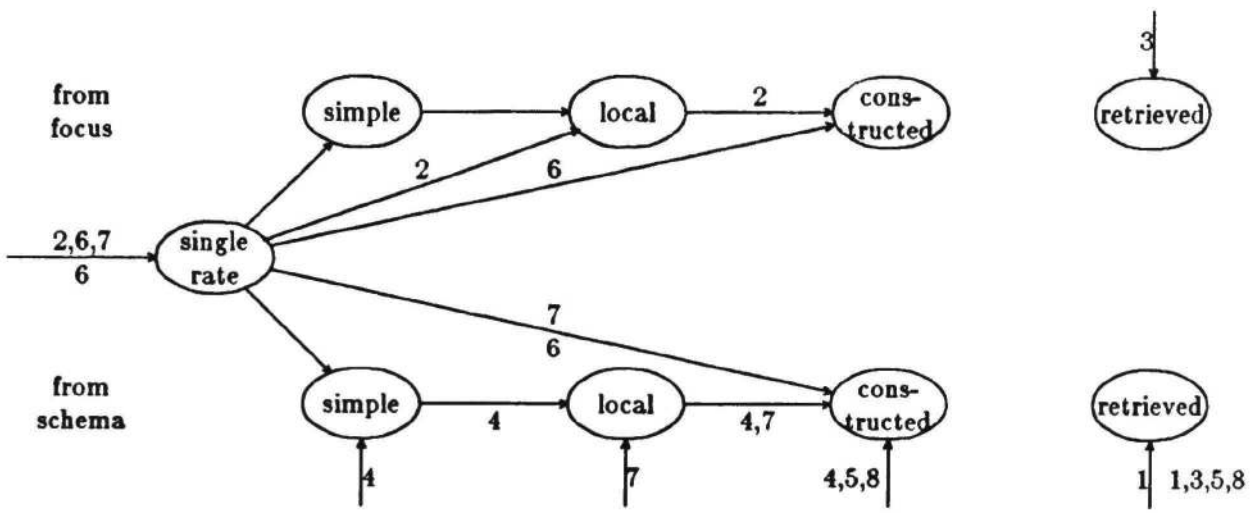


Figure 1: Development of the sum plan by subject

The expertise of the subjects in the study differed greatly. One subject retrieved the final *sum* plan without effort to solve the problem in the first isomorph. Another started at stage one both times. The transitions seen in the data are shown in Figure 1. Here, the development for the first presentation of the problem is shown above or to the left of the links. The behavior of the subject on the second isomorph is shown below or to the right of the link. Subject solutions within a problem solving episode developed according to the taxonomy described. Solutions across the episodes reflect the change from schema formation to schema retrieval and expansion of the slots.

The selection plan.

The selection plan provided the context for sum calculation by selecting when each of the new rates was applicable. The two plans may be treated in relative isolation, however, since all six stages of the sum plan occurred in the context of a successful selection. The declarative information used for the two plans has considerable overlap, but the use of boundaries in selection was not extended to the sum calculation. Generally, local focus mitigates against such

systematic comparisons, since the relevance of a particular item of data is determined by search from the focal goal. Schema application, with its description of required slots and slot fillers, is not yet available for use.

The select plan shows a development from separate goals each using IF to test the defined, exclusive ranges, to the complete ELSE-IF form where the structure of the ELSE-IF replaces the tests used in separate IFs. The change may be viewed as the progressive replacement of explicit boundary tests by the implicit testing of the ELSE-IF construct. Initially, each case is treated locally, in isolation from the other ranges. Over the course of development, the concept of exclusive, abutting boundary conditions and their use in the ELSE-IF statement emerges. This creates a different organisation of the problem information, from one based on separate goals to one based on filling the newly-defined slots of the ELSE-IF, embedding the boundary information and testing within the language construct. Several subsidiary plans are involved with the select plan and also show a developmental shift. These are the choice of what to use as a selector (one or two ranges), the form of the test within the selector and the combiner (OR or AND) when two tests are used. These will be discussed after the select plan.

Else-if development

Five stages of the select plan were seen for the problem described here, that of continuous, abutting ranges. The general form of the development may be seen as an optimization that requires less code for each stage. More informatively, it may be seen as the use of the structure of the ELSE statement to replace the explicit selection tests required by a local formulation, the embedding of the explicit goals into a structure where they are achieved implicitly.

1. *IF tests*: the selection is done using separate IF tests for each case. The case is then expanded within the context of the IF and treated in isolation from the other pieces. The condition in the test, the selector, may be one or both boundary tests.
2. *IF to ELSE-IF test*: the subject begins by using separate tests, but at some point, usually the second IF, links the separate tests using ELSE statements. The ELSE functions purely as a connector between IFs and has no other functionality.
3. *ELSE-IF with separate tests*: an ELSE-IF structure is used from the outset, but both boundary tests are still included as the selector. The decision to use the ELSE-IF as the operator was local and had no impact on subsequent (local) goal expansion.
4. *ELSE-IF with last test*: an ELSE-IF structure is used, but the subject realises that the structure can be used to test one of the conditions and only one boundary need be specified in each if statement. This insight is applied to all the tests, so the last test also has an explicit (unnecessary) boundary.
5. *ELSE-IF*: the final, expert form, with one boundary tested in all cases except the last, where it is implicitly true.

The initial goal of the subject is to select out the different ranges so that the charge can be computed for each category. A selection operator is retrieved (IF or ELSE-IF) and each of the categories is analyzed within the context of this operator. The definition of the category is inserted in the condition slot of the operator. This definition usually is conceived as a range with two boundaries, which is most directly coded as a test for more than the lower bound and less than the upper bound, creating the repeated IF structure of the first stage.

In the second stage, the IF schema has been formed and its use now becomes an issue. The set of exclusive alternatives cues the ELSE structure, which is now seen simply as an IF connector, and the separate IF statements are linked by this (syntactic) connector. In the third stage, the analysis is made before the code is written and the structure laid out for all the cases; each case is then expanded separately. In the fourth stage, the subject realises that only one test is required inside an IF, and so each IF test, including the last, is coded with a single test. In the last stage, the realization is fully applied and the optimized ELSE-IF structure created.

Selector development

In solving the selection problem, the first step is to discover the different categories. The simplest form of analysis focuses on the conditions that separate one category from another, such as 'under 350', 'over 350', 'over 625' and so on. These can then be directly coded using the known comparison operators. In the second stage, the subject realizes that both conditions are required and codes them both as the test condition.

Combiner development

When two selector tests are defined, they need to be connected with a boolean operator, AND or OR. The correct operator is AND, yet very often subjects chose to use OR and corrected themselves later. The OR operator is used correctly in the context of independent goals, such as enumeration (it's one or the other). The locality heuristic assumes that everything is independent until proven otherwise, when AND is substituted as the correct connector.

Test development

The initial concern in defining the boundaries for selection is to state which values enclose the category. A local focus on each of the categories created a definition in terms of equality, by specifying the lowest and highest possible values. In the electric bill, the equality focus identified the value pairs (0, 350), (351, 625), (626, 850) and over 850. These were then coded using >= and <= tests. The second stage saw the replacement of these 'add-one' values and >= tests with tests for >. This returned the boundary values to their form in the problem description and may indeed have created the initial insight that eventually led to the ELSE-IF optimization, that one plan's ceiling is another plan's floor.

The read plan.

The focal flavor of development is particularly noticeable in this plan. The goal of reading the data spawned the goal of looping, which was coded. The loop was then ignored, since the goal seemed to be achieved by writing the loop code. The rest of the goals were implemented and attention only returned to the loop code when the program was executed. The implications of using an input-controlled loop (Soloway and Ehrlich, 1984), with an end value of 99999, were discovered at this point. The removal of this value started a new round of problem solving. The difference between a bug and an extension of the design is the time at which the bug is found, not the process of discovery or solution. If the goal is currently focal, then an addition is part of the new, improved design. If the goal is past, and in the normal case if all the goals are past and the program considered complete, then the addition is a patch to solve a bug. The problem solving strategy and local focus is the same in either case.

Two patterns in the read plan development are especially interesting. The first is that novices would often come to the end of the WHILE loop and reconstruct the goal of terminating the loop at that point. The local focus means that attention is directed toward achieving the current goal and when it is achieved, or in this case when a sequence of goals was achieved, the context is reconstructed from the external memory provided by the program code. The reconstructed goal is then implemented in isolation from the program and indeed the problem, since substitute solutions such as keeping a count were generated as solutions. These contradicted the problem specification, which stated that 'end of input is signalled by a value of 99999'.

The second pattern emerged in debugging the WHILE loop. Two types of solutions were seen, arising from different conceptualizations of the bug. The first conceptualization was that the value *in the calculation* was wrong and the value had to be fixed, leading to a backout strategy. The second conceptualization was that the bug was *caused by* the read. This produced

two types of fixes. The first was to protect the calculation from the read *value* by adding code, such as a guard. The second was to protect the calculation from the *last* read by moving the read to the bottom of the loop and using the WHILE test as an implicit guard. These plans show increasing distance from the bug symptom to the final solution.

The stages in development seen to achieve the read goal were

1. *single read*: this is the simple prompt (WRITELN/READ) plan.
2. *read then loop*: the subject coded the input prompt and read and then mentioned the loop goal, inserting code before the read statement.
3. *loop then read*: the loop statement preceded the read. The need for a loop was noticed when reading the program and the read schema was coded.
4. *different plan*: local plans were spawned to solve the reconstructed termination goal. Two inappropriate plans were seen in the data. The first used an input value as a counter to control the loop. The second asked the operator to input a character variable, 'Y' or 'N', to control the loop. This plan had been taught in the course and was retrieved when a loop terminator was required.
5. *the backout plan*: subtract 99999 from the sum and 1 from the counters. The bug is that the variables contain the wrong value, so the simplest solution is to fix the values. The backout could be done directly where needed, after the loop and before the output, or be slightly improved and done when initializing the variables.
6. *selective backout*: if the value is 99999, make it zero. The bug is that the wrong value has been added in, so the solution is to make the value innocuous and convert it to zero. This fix saves the sum, but all the counters still need to be backed out.
7. *guard*: only add if the value is not 99999. The bug conceptualization is that the wrong value has been added and the patch is to prevent the end value from being included in the calculations. This saves all the running totals, but creates two tests for the end value, once in the WHILE and once in the guarding IF.
8. *read at end*: read the data at the end of the loop. The conceptualization of the bug is that one too many data have been read in, so the read at the end stops the extra read by using the control machinery to guard the calculations. It is an uncommon solution, because it requires a system viewpoint and then a selective backout.
9. *double read*: read before the loop and at the bottom of the loop. It is an extremely difficult solution because it requires a system view of the loop and two separate reads to achieve the single goal. The use of two pieces of code violates the implicit locality heuristic, that a goal is achieved by one piece of code. Subjects added code with abandon, but seldom changed existing code to achieve a new goal. This plan was commonly used for validation, to loop until a good value was read, but the different problem solving context seems to have prevented transfer.

The solutions tried by each subject are shown in Figure 2; links are labelled with the number of the subject who showed that particular transition. The plan was necessary in three of the problems; the first problem did not involve looping. The data from these three problems may be read from top to bottom of the subject numbers. The place marker '-' is used to indicate that no subject made a transition on a particular problem. Some subjects ran out of time on initial attempts and did not encounter the 99999 problem, so some early paths are not complete. Thus on the first occasion, subjects 3 through 7 coded locally from the read to the loop goal. Subjects 3 and 4 then used a guard, subject 5 ran out of time, subject 6 used a retrieved, inappropriate solution and subject 7 used backout then guard. It is interesting to note that none of the plan patches were schematised, being abandoned or redeveloped from the focus each time.

Program development.

The effect of focus has been demonstrated within individual plan segments. Two more general examples of local development will be presented here, showing the use of a local

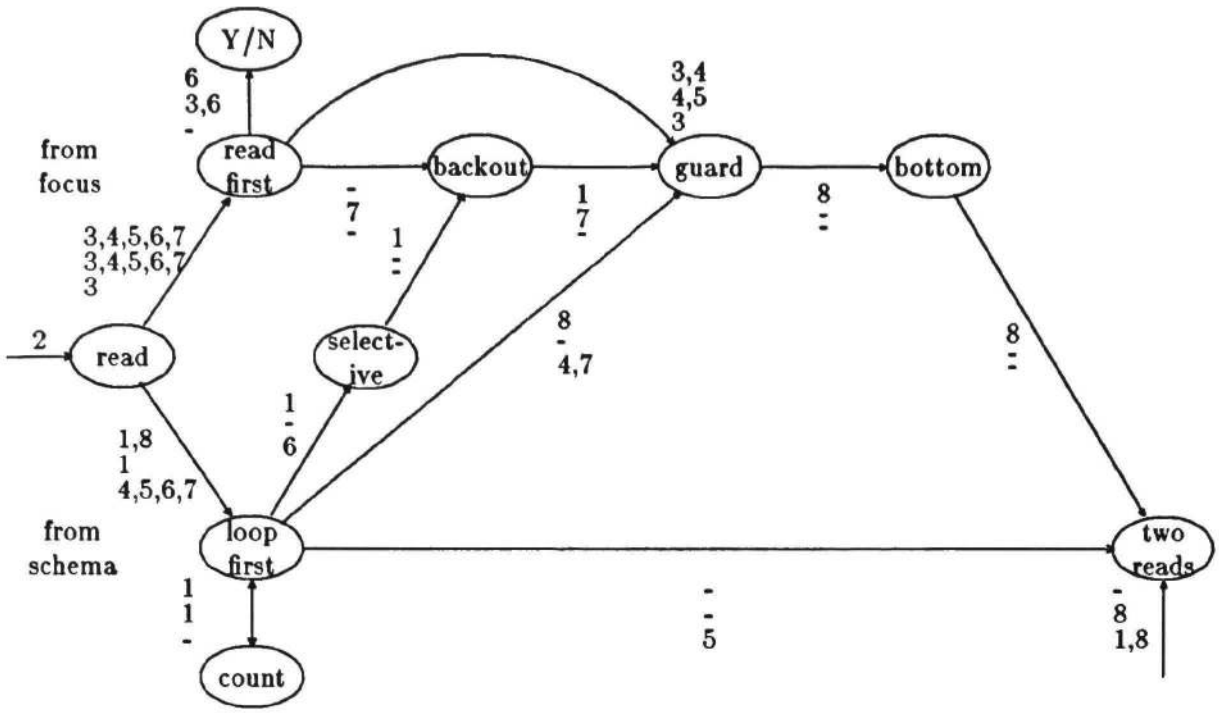


Figure 2: Development of the read plan by subject

expansion strategy divorced from any simple, identifiable schema. The first reflects the development of knowledge about the structure of plans in a program. The second reflects the integration of information by a focal goal as a general problem solving heuristic.

A plan may be viewed as consisting of several segments, each of which fulfills a required role. These roles are general plan pre- and post- requisite descriptions, such as input, process and output (IPO) or initialise, calculate and use. The theoretical value of such a description is demonstrated in Rich's (1981) plan calculus; evidence for their use by novices and experts is presented in Rist (1986). The development of a modular program structure may be viewed as the hierarchical organisation of plans and plan segments that implement the 'correct' ordering at the appropriate level. A simple loop program, for example, shows a goal centered structure (IPO) at the most abstract level. Plan segments are then organised by each of these roles (III, PPP, OOO), spreading elements of a plan across the program in a role centered structure.

In understanding the problem, novices identified the goals and then the input objects for these goals. The goals and their associated inputs create the set of transforms (IPO) that is used to build the program. In the electricity program there were two goals, to find the cost (C) from the number of kilowatt hours used and the billing area (A) from the customer number. These goals are independent and their role organisations thus show the reasoning process unconstrained by issues of data flow. The correct, systematic placement of roles and objects must be learned through experience. The types of role patterns for the two isomorphs are shown in Table 1.

The initial problem solving organisation comes from the role structure. As was seen for the loop, the first heuristic is to read in the data, then process it. The input role is especially useful in problem solving in programming, since the input defines the objects required, is simple, always independent and presents few problems. At each role transition (I -> P, P -> O), an object must be selected for the new role. The simplest selection is to use the goal that was last used, producing the object order ACCAAC shown on the first line. The first stage of learning and planning is to link the process and output roles for the same object, creating an implicit goal link

Table 1: Development of role organisation

occurrence				pattern		
subjects		numbers				
elec	welfare	elec	welfare	independent		loop
3	4	1	1	I_aI_c P_cP_a O_aO_c		
6		1		I_aI_c P_cP_a O_cO_a	-->	I_cI_a P_cP_a O_cO_a
2,4,5,7,8	3	5	1	I_a $I_cP_cO_c$ P_aO_a		
1	1,5,6,7,8	1	5	$I_cP_cO_c$ $I_aP_aO_a$		

(CACA). This link is made explicit in the the next stage, where process and output are organised underneath the objects (CCAA). The goal organisation is then extended to all the roles and the goals are treated independently, producing the order CCCAAA used in the second isomorph. A similar development was seen for the validate (V) role. It was initially considered at the end of the program for all input objects (role centered, IIPPOOVV) and coded in the program at various positions. It was then considered at the end of each independent goal (IPOV, IPOV) and finally the role was embedded in the normal plan structure (IVPO, IVPO).

The first role reorganisation (line 2) creates the concept of systematic nesting, here nesting under roles. Complete role nesting is derived from this pattern by extending it to the input role, with the transition ACCACA -> CACACA. This organisation is correct for plans inside loops, where the main processing is delimited by the loop boundaries. The expert selects which pattern is needed from the type of plan; the novice begins by using local expansion and develops these patterns and selection rules through debugging.

The incorrect placement of plan roles inside the loop is shown in Table 2. Programs two and four used several plans, such as *max* and *average* that are not discussed in this paper; the table indicates the total number of incorrect placements for all four plans. Only those roles that provided a choice for placement are included in the table, such as final calculations for average and percentage and program output; these could be placed inside or after the loop. Their location is a sensitive indicator that demonstrates the lack of developed discourse rules.

Table 2: Local plan expansion

Program	Subject number						
	1	3	4	5	6	7	8
	IPO	IPO	IPO	IPO	IPO	IPO	IPO
2		14				22	
3		1			1	1	1
4		24				34	

The Welfare problem (problem 3) demonstrates local problem solving in the interpretation and organisation of problem data. It was the isomorph of the electric bill problem, requiring subjects to calculate the benefits an unwed mother would receive in a year, given the number of children in the family. At this point, three weeks into the study, it was expected that the selection and sum plans would have been schematised and could be retrieved and used. This was supported; subjects applied the plans and inserted values in the plan schema slots without apparent effort. The interesting point was that over half (four out of seven) of the subjects retrieved the plan, expanded it and coded benefits for each of the categories including the category of no children. A value of children less than zero was an error, a value of one or greater received benefits according to the rate schedule given and a value of zero received zero benefits. From a global perspective it is difficult to see how an unwed mother could have zero children.

IMPLICATIONS

The shift from a local and concrete to a system-oriented and abstract view in design marks and defines expertise in a domain. That shift has been explained here as arising from the problems encountered by a novice, the problem solving techniques used to overcome them and the use of the products, plan schemas, in solving further problems. The change from goal expansion to modular planning has been traced to the interaction between plan flow and control and data flow via the creation and automation of plans and plan segments.

It is interesting to compare the results of this study to the theoretical models of skill acquisition discussed by Sussman (1975), Brown and VanLehn (1980) and Hayes-Roth (1983). The general method of Sussman, that of attempting a solution and creating new goals and patches as required, was the design strategy used by novice subjects. The solution of impasses has been explained by a method of adopting a focal object, operator or schema and developing a solution around it. The problem solving heuristics used by Hayes-Roth can now be placed in a developmental sequence, from the method of noticing a bug and applying a local patch to examining the underlying causality of the bug and fixing the cause instead of the symptom.

The detailed analysis of the information involved in coding a computer program offers the hope of a realistic simulation of human problem solving behavior. The use of schema as knowledge organisers echoes Schank's (1982) MOP organisation and extends it by demonstrating the ontogeny of such theoretical constructs in an experimental task. The use of these MOPs in understanding involves many of the same issues as the use of plan schema in program design. Study of their development provides a strong answer to the question of what is inside a schema.

I wish to express my appreciation to Jim Spohrer and Dana Kay for their trenchant criticisms of early versions of the paper. The study was funded by grant number 66430 from IBM.

REFERENCES

Brown, J. S. & VanLehn, K. (1980). Repair theory: A generative theory of bugs in procedural skills. *Cognitive Science*, **4**, 379-426.

Hayes-Roth, F. (1983). Using proofs and refutations to learn from experience. In R. S. Michalski, J. C. Carbonell and T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach*. Palo Alto, CA: Tioga.

Kant, E. (1985). Understanding and automating algorithm design. *IEEE Transactions on Software Engineering*, **SE-11**, 1361-1374.

Rich, C. (1981). *Inspection methods in programming.* (Tech. Rep. AI-TR-604). Boston: MIT, MIT AI Lab.

Rist, R. S. (1986). Plans in programming: Definition, demonstration and development. In E. Soloway and S. S. Iyengar (Eds.), *Empirical studies of programmers*. New York: Ablex.

Schank, R. C. (1982). *Dynamic memory*. Cambridge, MA: Cambridge University Press.

Soloway, E. & Ehrlich, K. (1984). Empirical studies of programming knowledge. *IEEE Transactions on Software Engineering*, **SE-10**, 595-609.

Spohrer, J. C., Pope, E., Lipman, M., Sack, W., Freiman, S., Littman, D., Johnson, L. & Soloway, E. (1985). *Bug catalogue: II, III, IV.* (Tech. Rep. 386). New Haven: Yale University, Department of Computer Science.

Spohrer, J. C., Soloway, E. and Pope, E. (1985). A goal/plan analysis of buggy Pascal programs. *Human-Computer Interaction*, **1**, 163-207.

Sussman, G. J. (1975). *A computer model of skill acquisition*. New York: American Elsevier.

Towards Completely Integrated Parsing and Inferencing*

Christopher K. Riesbeck and Charles E. Martin
Yale University

ABSTRACT

Our goal is the complete integration of natural language understanding with the rest of cognition. Two mechanisms that we have developed and implemented to achieve this goal are: (1) the Direct Memory Access Parsing (DMAP) algorithm, based on the notion of lexically-guided memory search and concept refinement, and (2) an inference-triggering process based on the notion of concept refinement failures. Together, these two mechanisms form a tightly-integrated system of parsing and inferencing, with no artificial boundaries between them.

Introduction

Our goal is the complete integration of natural language understanding with the rest of the cognitive system. The benefits of full integration are obvious: such a parser could take full advantage of whatever knowledge was present in memory (or, at least, could take as much advantage as any other memory process could), and other memory processes could make full and immediate use of linguistic input, without waiting for a final interpretation to be formed.

This paper describes two mechanisms that we have developed and implemented to achieve this goal. First, our parsing algorithm is a process of lexically-guided memory search. That is, patterns of words and concepts guide a general memory search process towards relevant memory structures, like lighthouses guiding a ship into harbor. We call this *direct memory access parsing* (DMAP). Second, our inference processes are triggered by *specialization failure structures*, generated by the parser, to record problems in building new instances of memory structures. Together, these two mechanisms form a tightly-integrated system of parsing and inferencing, with no artificial boundaries between them.

Our memory structures are frame-like objects called *Memory Organization Packets* (MOPs), organized into the standard part-whole (*packaging*) and class-subclass (*abstraction*) hierarchies [Schank 1982]. We integrate parsing knowledge into memory by attaching linguistic templates to these memory structures, in a manner reminscent of the Teachable Language Comprehender [Quillian 1969]. These templates, called *concept sequences*, are patterns of words and concepts. Attached to Milton Friedman, for example, is the lexical phrase "Milton Friedman". Attached to the general concept of a communication event is the concept sequence "[actor of communication] says [object of communication]." Any memory structure can have one or more concept sequences, plus every structure implicitly inherits the sequences attached to abstractions of that structure.

The dictionary in DMAP, which we call the *concept lexicon*, is simply a set of pointers from words and concepts to the concept sequences they appear in. For example, "Milton" has a pointer to the sequence "Milton Friedman" which is attached to the person concept MILTON-FRIEDMAN, and the concept HUMAN has a pointer to the sequence "[actor of communication] says [object of communication]." Obviously, "Milton" might point to every person the system knows named Milton, and HUMAN might point to every action described with a sequence involving a human. Our current concept lexicon only has a pointer from a word or concept to a concept sequence if the word or concept appears at the beginning of the sequence. Even so, the DMAP model depends on the use of

*This work was funded in part by the Air Force Office of Scientific Research under contract F49620-82-K-0010.

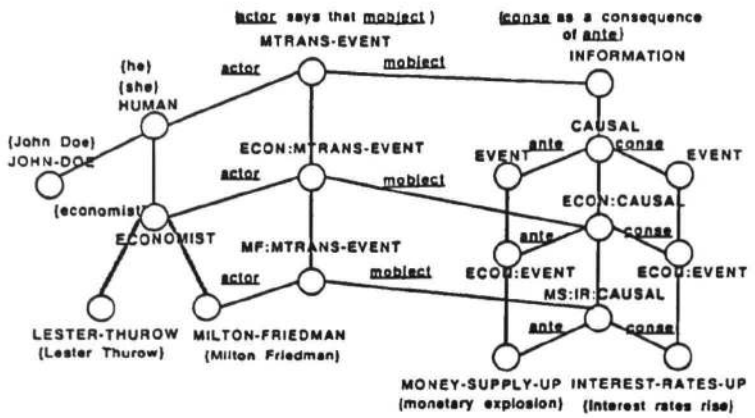


Figure 1: A portion of the DMAP memory.

parallel activation and intersection to resolve the basic combinatorial explosion, as is presumed in a number of other recent models [Small et al. 1982] [Hahn and Reimer 1983] [Granger et al. 1984] [Waltz and Pollack 1984][Charniak unpb].

When DMAP reads a sentence, such as "Milton Friedman says that high interest rates are the consequence of the monetary explosion," the concept lexicon leads it to concept sequences attached to memory structures, such as "economist claims economic causal connection," (ECON:MTRANS-EVENT [1] in Figure 1), and "rise in money supply causes rise in interest rates" (MS:IR:CAUSAL). Filling out these sequences activates the associated memory structures.

Because the parsing process, as described in the next section, pushes activation down to the most specific memory structures available, exactly which memory structures the parser settles on depends on which ones are already in memory. If this claim of Friedman's has been seen before, then seeing it again, as originally stated, or paraphrased, will guide the parser to the previously built memory structure MF:MTRANS-EVENT.

If MF:MTRANS-EVENT is not already in memory, then the parser is stops at a more general level, such as ECON:MTRANS-EVENT. When the parser cannot finding a more specific structure, either because there are none, or because the ones that exist do not match the input, it activates a specialization failure structure. One such structure is "actor exception", which means that the input event partially matches some existing memory structure, but the actors are different. For example, if the system has already seen the Milton Friedman sentence, and now reads "John Doe blames the large increase in money supply for the rise in interest rates," it will find the Milton Friedman structure, but be unable to specialize to it because of the mismatch in actors.

Specialization failure structures, like other memory structures, are organized by part-whole and class-subclass relationships. The responses attached to these failure structures are *reconciliation* processes that propose resolutions to the failure. Resolutions are proposed by activating potential explanation patterns (XPs) [Schank 1986].[2] A routine domain-specific XP for explaining why two economists say the same thing is "they belong to the same economic camp." The XP may or may not be applicable, depending on what the system already knows, i.e., what other memory structures it has that package and abstract Thurow and Friedman.

In the rest of this paper, we first discuss the details of direct memory access parsing. We will then describe two classes of specialization failure structures, and how they're activated during the parsing process.

Direct Memory Access Parsing

This section describes the current implementation of the DMAP interpreter. The interpreter uses a marker-passing architecture to identify relevant memory structures from the input text and the expectations in memory. Two kinds of markers are used in the system: *activation* markers, which cap-

[1] MTRANS is our primitive marker for communication events [Schank 1975].

[2] Specialization failure structures are similar in spirit to the Exception MOPs proposed in [Riesbeck 1981].

ture information about the input text and the current selection of relevant memory structures, and *prediction* markers, which indicate which memory structures may be expected to become relevant.[3]

The structure of memory

Figure 1 is a simplified portion of the DMAP interpreter's memory to represent the communicative act of the following text.

> Milton Friedman: Interest rates will rise as an inevitable consequence of the monetary explosion we've experienced over the past year.[4]

The central structure to this portion of memory is the MOP MF:MTRANS-EVENT, which packages MILTON-FRIEDMAN and MS:IR:CAUSAL via the actor and mobject roles. The concept sequence {actor says that mobject} is associated with the more general MTRANS-EVENT MOP. This concept sequence is the linguistic template used to recognize the memory structure MTRANS-EVENT. It is a pattern of words (e.g., "says") and concepts in the form of packaging relationships from which concepts are referenced (e.g., actor references HUMAN from MTRANS-EVENT, but MILTON-FRIEDMAN from MF:MTRANS-EVENT).[5]

Concept activation

Memory structures are activated by placing activation markers on them. Activation markers are created in two situations.

- *System input:* when an input word is read by the interpreter, an activation marker is created and placed on the associated lexical item in memory.
- *Concept sequence recognition:* when every element of a concept sequence has been activated, an activation marker is created and placed on the associated memory structure.

Activation markers are passed up the class-subclass abstraction hierarchy from their associated structures. This is a recursive process; all structures which receive an activation marker continue to pass it on to their own abstractions. When a memory structure receives an activation marker, that memory structure is said to have been *activated*; the activation marker contains a pointer to the originally activated memory structure.

For example, an activation marker associated with the memory structure MONEY-SUPPLY-UP will be passed to ECON-EVENT, which in turn passes the marker to EVENT. All of these structures are activated, while the activation marker keeps a pointer to MONEY-SUPPLY-UP.

Concept prediction

The *concept lexicon* indexes concept sequences under the memory structures referenced by the first elements of the sequences. For example, the MILTON-FRIEDMAN structure's concept sequence {Milton Friedman} is indexed under the lexical item "Milton", and the MTRANS-EVENT MOP's concept sequence {actor says that mobject} is indexed under the structure HUMAN, since HUMAN is referenced by the actor role of MTRANS-EVENT. These indicies are not shown on the diagram.

Prediction markers represent concept sequences which are in the process of being recognized. Whenever a memory structure is activated, prediction markers are created for all the concept sequences indexed by that memory structure through the concept lexicon. A prediction marker is a structured object which records its associated *concept sequence*; the *current element* of the concept sequence (either a packaging role or a lexical item); the *reference structure* which will be activated if the concept sequence is recognized—initially this is the structure associated with the concept sequence; and the *target* of the prediction (either a memory structure or a lexical item).

A prediction marker is always located at the target memory structure, which is derived from the reference structure and the current element. A prediction marker is passed through memory whenever its reference structure or current element is changed; this takes place through two concurrent processes.

[3] The markers of the system are structured objects which would be unacceptable in current connectionist theories [Feldman and Ballard 1982] [Waltz and Pollack 1984].

[4] *The New York Times*, August 4, 1983.

[5] The EVENT and ECON:EVENT MOPs are shown twice in the diagram to make the packaging relationships clear. There is only one structure for each in memory.

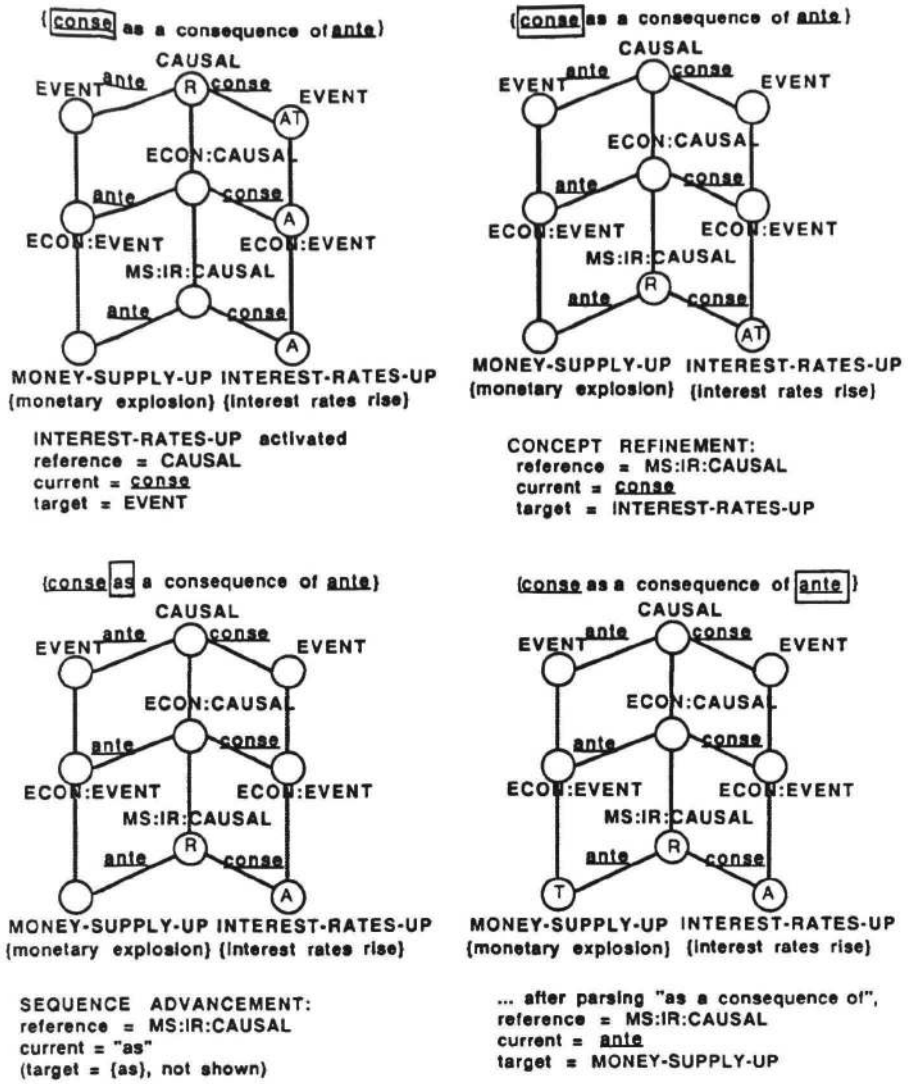


Figure 2: How prediction markers are passed.

- *Concept refinement:* This occurs when the target of a prediction marker whose current element is a packaging role is activated. Since the reference structure will generally package an abstraction of the originally activated structure, the reference structure can be replaced by a specialization which explicitly packages the original activation.

- *Prediction satisfaction:* when the target of a prediction marker is activated, the current element of the concept sequence is satisfied by that activation. The current element is replaced by the next element in the concept sequence. (If there are no further elements, then the concept sequence has been recognized, and the reference structure is activated.)

For example, the concept sequence {conse as a consequence of ante} is associated with the structure CAUSAL; if INTEREST-RATES-UP is activated, the subsequent activation of EVENT will index this concept sequence through the concept lexicon, and a new prediction marker will be produced. Concept refinement will change the reference structure from CAUSAL to MS:IR:CAUSAL (the most specific memory structure which packages the activated memory structures), and prediction satisfaction will change the current element to the "as" lexical item. The new target is this lexical item, and the prediction marker is passed to this memory structure. After reading the phrase "as a consequence of", the current element will become the mobject role, which ref-

erences MONEY-SUPPLY-UP from the reference structure MS:IR:CAUSAL. MONEY-SUPPLY-UP becomes the new target. This "trace" of the parse is depicted in Figure 2.

Specialization Failure Structures

We've now seen how the parsing process works when the net result is simple recognition. A key event that the DMAP interpreter watches for is when an activation marker meets a prediction marker. This causes a specialization process to occur, which spreads the prediction markers down the abstraction hierarchy. If the prediction markers can be pushed all the way down to the level of the input activation markers, then the parsing process has reconciled the input completely with some existing memory structure. Thus, given the memory in Figure 1, "Milton Friedman says that high interest rates are a consequence of the monetary explosion" activates MF:MTRANS-EVENT and that node packages nodes with input activation markers.

If MF:MTRANS-EVENT did not already exist in memory, then the parser would activate the specialization failure structure MISSING-SPECIALIZATION, which represents the situation where the parser can't specialize a memory structure because no specialization of that structure exists. In this case, the parser can't specialize ECON:MTRANS-EVENT, even though Milton Friedman and his argument are more specialized than ECONOMIST and ECON:CAUSAL.

The MISSING-SPECIALIZATION structure packages the general structure and the specialized fillers, e.g., ECON:MTRANS-EVENT and the specialized fillers MILTON-FRIEDMAN and MS:IR:CAUSAL. The reconciliation strategy attached to MISSING-SPECIALIZATION is simple: instantiate a new specialization of the existing memory structure to package the specific items. We call this reconciliation ROTE-MEMORY because it simply adds the items seen to memory. If MF:MTRANS-EVENT was not already in memory, then ROTE-MEMORY would create it.

ROTE-MEMORY is invoked only in the simple situation where you know things of a certain type can occur, and one of them does. The input matches completely a general pattern and there is no more specific version of the pattern to compare the input with.

Of more interest is the case where the input *partially* matches some existing structure. For example, suppose that MF:MTRANS-EVENT is in memory and the parser reads "John Doe blames the large increase in the money supply for the rise in interest rates." When this is parsed, the parser is unable to specialize from ECON:MTRANS-EVENT to MF:MTRANS-EVENT because of the mismatch between the actor of the input, i.e., John Doe, and the actor of MF:MTRANS-EVENT, i.e., Milton Friedman.

This specialization failure is called ECONOMIC-MTRANS:ACTOR-EXCEPTION. It packages together: a *reference structure* that couldn't be specialized, a *potential reference structure* (a specialization of the reference structure) that the parser could not reach because the *activated structure* which satisfied the actor role of the prediction marker is not an abstraction of the *anomalous feature structure* packaged by the actor role of the potential reference structure. In the John Doe example these would be ECON:MTRANS-EVENT, MF:MTRANS-EVENT, JOHN-DOE, and MILTON-FRIEDMAN, respectively.

The general situation of two people saying the same thing can be explained in many ways, but we are interested here in explanations specific to economists and their public pronouncements about the economy. That is why we focus on ECONOMIC-MTRANS:ACTOR-EXCEPTION, rather than the more general MTRANS:ACTOR-EXCEPTION.[6] One reconciliation strategy for ECONOMIC-MTRANS:ACTOR-EXCEPTION is CREATE-ECONOMIST-CAMP. This strategy takes the following actions.

- Create a new *economist camp* memory structure which is a specialization of the least abstract generalization of the activated and anomalous feature structures, and make these features specializations of it.
- Create a new *camp mtrans event MOP* which is a specialization of the reference structure; the new mop packages the economist camp

[6]Since these failure structures are organized in the regular hierarchical memory format, strategies at the level of MTRANS:ACTOR-EXCEPTION are accessible if more specific strategies are ineffective in resolving the anomaly.

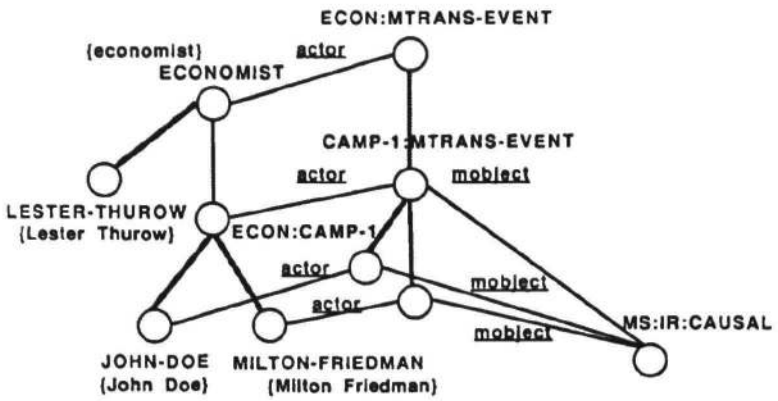


Figure 3: Building new structures.

structure and the other activations of the input.

- Add the potential reference structure and the structure created by the ROTE-MEMORY strategy as specializations of the new camp mtrans event MOP.

In this example, the strategy creates as a specialization of ECONOMIST an economist camp MOP (call it ECON:CAMP-1) and makes JOHN-DOE and MILTON-FRIEDMAN specializations of it. It packages this new MOP by a new camp mtrans MOP, CAMP-1:MTRANS-EVENT, which is a specialization of ECON:MTRANS-EVENT, and packages MF:MTRANS-EVENT and the new John Doe mtrans event created by ROTE-MEMORY as specializations of it. CAMP-1:MTRANS-EVENT has the interpretation that the new economic camp claims that that increased money supply is causing high interest rates. Figure 3 shows the net result.

A topic for future research is how the system might learn a name for ECON:CAMP-1, such as "monetarist," while reading text.

CREATE-ECONOMIST-CAMP would apply if we knew nothing else about Friedman and Doe except these two statements. Suppose however that the concept of "monetarist" is already in memory, Friedman and Doe are instances of it, Lester Thurow is known to be in another camp, and the parser reads "Lester Thurow blames the rise in interest rates on the increased money supply." The appropriate failure structure in this case is ECONOMIC-MTRANS:CAMP-EXCEPTION, which is a specialization of ECONOMIC-MTRANS:ACTOR-EXCEPTION that represents a failure to specialize because of a camp mismatch, rather than an actor mismatch.

ECONOMIC-MTRANS:CAMP-EXCEPTION, is an anomaly requiring more creativity in explanation than our simple routine XP strategies are intended to account for. Possible explanations are that Thurow is changing camps, or he's lying, or the monetarist position is becoming more acceptable to non-monetarists. Coming up with explanations at this level of complexity is under investigation [Schank 1986], but not in the DMAP framework.

Causal Chains

Although we focussed above on specialization failure structures and related strategies for adding economic communication events to memory, the same approach is being used to direct inferencing to build causal chains as well.

Building causal chains is a key part of economic reasoning. For example, one text that DMAP parses and reasons about is

> With high growth choked off by high interest rates, the deficit will be bigger, not smaller. [7]

To understand this text, a causal chain has to be found connecting interest rates to growth, and growth to the deficit.

When the parser reads "high growth choked off by high interest rates," it tries to specialize the economic causal "high interest rates cause reduced growth." If there is already is such a causal in memory, there is no problem, but if there isn't, then the specialization failures for economic causal structures are activated.

[7] Lester Thurow, *Newsweek*, September 21, 1983.

If there is no causal argument in memory mentioning interest rates or growth at all, the strategy ROTE-MEMORY will just add the input causal to memory. If there is instead a causal argument that says interest rates cause inflation, then the strategy attached to the specialization failure structure ECONOMIC-CAUSAL:CONSEQUENT-EXCEPTION will try to reconcile the input with memory by finding a second causal connecting increased inflation to reduced growth, hence building a causal chain.

Conclusions

We've discussed two techniques for integrating parsing and inferencing with a hierarchical memory. First, the direct memory access parsing algorithm attaches concept sequences to structures in memory and uses these sequences to guide memory search to the most specialized applicable packaging structures available. Second, specialization failures during parsing activate specialization failure structures, to which are attached reconciliation strategies that create new memory structures that will allow the input to be instantiated and installed into memory. The examples of reconciliation strategies given deal with one economist agreeing with another, and with causal chains. Our belief is that the reconciliation strategies are a significant part of the total inference processing that goes on during text understanding and evaluation.

Acknowledgments

The work described here is based on the joint efforts of the Direct Memory Access Parsing project, which consists of Charles E. Martin, Monique Barbançon, and Michael Factor.

References

Charniak, E. (Unpublished). *A Single-Semantic-Process Theory of Parsing.*

Feldman, J.A. and Ballard, D. (1982). Connectionist models and their properties. *Cognitive Science, 6* (3), 205–254.

Granger, R.H., Eiselt, K.P., and Holbrook, J.K. (1984). The parallel organization of lexical, syntactic, and pragmatic inference processes. In *Proceedings of the First Annual Workshop on Theoretical Issues in Conceptual Information Processing.* Atlanta, GA.

Hahn, U. and Reimer U. (1983, November). *Word expert parsing: An approach to text parsing with a distributed lexical grammar.* Bericht TOPIC 6/83. Universitat Konstanz, Konstanz, West Germany.

Quillian, M.R. (1969). The Teachable Language Comprehender: A Simulation Program and Theory of Language. *Communications of the ACM, 12* (8).

Riesbeck, C.K. (1981). Failure-driven Reminding for Incremental Learning. In *Proceedings of the Seventh International Joint Conference on Artificial Intelligence.* Vancouver, B.C..

Schank, R.C. (1975) *Conceptual Information Processing.* Amsterdam: North Holland/American Elsevier.

Schank, R.C. (1982) *Dynamic Memory: A Theory of Learning in Computers and People.* Cambridge University Press.

Schank, R.C. (1986) *Explanation Patterns: Understanding Mechanically and Creatively.* In preparation.

Small, S., Cottrell, G. and Shastri, L. (1982). Toward Connectionist Parsing. In *Proceedings of the AAAI-82.* Pittsburgh, PA..

Waltz, D.L. and Pollack, J.B. (1984). Phenomenologically plausible parsing. In *Proceedings of the AAAI-84.* Austin, Texas.

SPEAK OF THE DEVIL: REPRESENTING DEICTIC AND SPEECH ACT KNOWLEDGE IN AN INTEGRATED LEXICAL MEMORY[1]

Michael Gasser
Michael G. Dyer

Artificial Intelligence Laboratory, UCLA

INTRODUCTION

Consider the following dialogue:

```
1)  X:  I heard Sally got tenure.  [Sally approaches.]
    Y:  Well, speak of the devil.
```

The expression *speak of the devil* says, in effect: "What a coincidence! The person that we were just talking about has arrived on the scene." How would this notion and the lexical entry for this expression need to be represented in memory in order for Y to be able to generate it? The problem is that the concept involves a person who is definable only in terms of the ongoing discourse. A speaker should fail to select the expression *speak of the devil*, for example, when a coincidence referred to involves a person not under discussion.

We are developing a program called CHIE which simulates the English and Japanese speech production of a Japanese learner of English in conversational and narrative contexts (Gasser, 1985). For conversation issues like that associated with *speak of the devil* assume importance. The problem is one of lexical representation. We want lexical entries that can make declarative reference to elements of the utterance and the discourse context in which the lexical items occur. While our main concern is generation, we have been guided by the need to have lexical entries usable in parsing as well. In addition, since we are modelling a human language learner, a major consideration in the design of the lexicon has been learnability.

DEIXIS AND SPEECH ACTS IN CONVERSATION

Deixis

In all language use, but in conversational contexts in particular, certain expressions make reference to entities that are part of the discourse itself, i.e., the speaker, the listener, the time and place of the conversation, and the topic of discussion. Such expressions are called **deictics** (Anderson & Keenan, 1985). *Speak of the devil* is an example of a **discourse deictic**, an expression that points to something that is being discussed. Other categories include person, time, place, and social deictics.

Person deictics are expressions that refer to the speaker or listener of the utterance in which they occur. Sentence 2) has two person deictics, *I* and *your Honor*.

```
2)  Your Honor, may I question the witness?
```

Your Honor is an example of a **vocative**, an expression used to get the attention of a potential listener or to emphasize the contact between the speaker and listener. Vocatives are often highly context-dependent, and *your honor*, which occurs in the context of a trial, is a case in point. This context-dependence further complicates the lexical representation. Not only must the person referred to by *your Honor* be specified as the listener of the current utterance but also as the judge of the trial in which the utterance occurs. Context specificity is not confined to vocatives. In some cases particular expressions constitute basic events in

[1]The research reported on here was supported in part by a grant from the ITA Foundation.

scripts (Schank & Abelson, 1977) or other knowledge structures. An example is *I now pronounce you man and wife* from the wedding script.

Time deictics are expressions that are definable in terms of the time of the utterance. Examples include *ago* and *yesterday*. **Place deictics** point to the location of the speaker or listener. Examples are *here* and *come*. *Come*, for instance, refers to a movement in the direction of the speaker or listener.

Social deictics make reference to a social relationship between the speaker and someone else. Consider the following Japanese sentences.

```
3a)  Abe-san  wa     imooto     ni  yubiwa  o       kure-masita.
     Mr.-Abe  TOPIC  my-sister  to  ring    DIROBJ  give-PAST
     'Mr. Abe gave my sister a ring.'

 b)  Imooto     wa     Abe-san  ni  yubiwa  o       age-masita.
     my-sister  TOPIC  Mr.-Abe  to  ring    DIROBJ  give-PAST
     'My sister gave Mr. Abe a ring.'
```

Note that there are two separate verbs corresponding to English *give*. The notion of relative "social distance" is needed to define the distinction. *Kure-* is appropriate only when the speaker is socially "closer" to the recipient than to the donor. *Kure-* is selected in 3a) because the recipient is the speaker's sister, who is closer to the speaker than is Abe, the donor. With the donor and recipient reversed, as in 3b), *kure-* is not possible, and the default *age-* is selected. *Kure-* is not an isolated case; social deixis in various forms runs rampant through the Japanese lexicon.

Speech Acts

Every utterance realizes a **speech act** of some kind; that is, it is an attempt to achieve some goal of the speaker or writer (Searle, 1969). In conversational contexts, however, we deal with a greater variety of speech acts that we find in, say, narrative, and many are achieved through the use of lexical patterns. Speech acts resemble deictics in that their definition requires reference to the current speaker and listener. A **directive**, for example, has the goal on the part of the speaker of getting the listener to perform some action. The speech acts that we are interested in here are **indirect** ones such as 4).

```
4)  I was wondering if you could take out the garbage.
```

What we note about such utterances is that they have both a "literal" and an "indirect" sense. For example, in its "literal" sense 4) simply constitutes an assertion about the speaker's wanting to know something, while in its "indirect" sense it functions as a directive. The lexicon should be organized in such a way that both meanings are available to listeners who are interpreting the sentences (Gibbs, 1980).

THE LEXICAL MEMORY APPROACH

The last ten years have seen a move in theoretical linguistics toward the inclusion of increasing amounts of linguistic knowledge in the lexicon (e.g., Bresnan & Kaplan, 1983; Hudson, 1984). At the same time, because lexical items can be associated directly with concepts and also provide much of the structure of a phrase directly, language processing models have come to be mostly lexically- rather than syntactically-driven (e.g., for parsing, Dyer, 1983; for generation, McKeown, 1985). Finally, there is a new emphasis on idiomaticity in language, with provision for a large number of lexical entries specifying morpho-syntactic **patterns** of varying degrees of specificity. This view stems from work in linguistics (Fillmore, 1979), in psycholinguistics (Pawley & Syder, 1983; Peters, 1983), and within the **phrasal lexicon** framework and related approaches in AI (Becker, 1975; Jacobs, 1985; Wilensky & Arens, 1980; Zernik & Dyer, 1985). The arguments center on the sheer number of lexical patterns which cannot be accounted for in terms of compositional semantics, the productive nature of many of these phrases, and the computational savings that result from the inclusion of frequent phrases in the lexicon.

At the same time, the need to integrate these notions into the rest of the linguistic system has led to the use of hierarchical network or frame-like representation systems for linguistic knowledge (Hudson, 1984; Jacobs, 1985; Langacker, 1986). This approach results in greater modularity and permits the efficient sharing of features through the creation of categories at various levels. In addition, the linguistic knowledge base of a system can be represented using the same general framework as for other knowledge in the system, promoting uniformity.

We assume that linguistic knowledge is acquired by associating relatively fixed patterns with relatively specific concepts or contextual elements (Peters, 1983). When generalizations are made on the basis of instances of similar patterns, higher-level nodes are added to the hierarchy. Each new generalization is either knowledge about how a type of pattern conveys a semantic notion or about how a type of utterance can achieve a type of communicative goal.

Many uses of language are inseparable from the contexts in which they occur. *Your Honor* in 2) above provides one striking example. The implication is not only that lexical items need to be acquired in context (Zernik & Dyer, 1985) but also that lexical items comprising parts of knowledge structures such as scripts (Schank & Abelson, 1977) are learned along with the knowledge structures themselves. This position agrees with research emphasizing the contextualized nature of language acquisition (Hatch, Flashner, & Hunt, 1986).

In CHIE there is a language-independent hierarchy of concepts, including generalized states, events, actions, physical objects, and goals. Knowledge about morphology is organized in a separate hierarchy of words and word classes (Hudson, 1984). An **utterance** associates a concept with a sequence of one or more words, and a **speech act** associates a goal with an utterance. The basic units of linguistic knowledge (above the phonological and morphological level) are utterance and speech act generalizations. We call our scheme the **Lexical Memory** approach because of the importance of lexical information and because of the ways in which linguistic knowledge is integrated into the rest of memory.

INTEGRATING LEXICAL, CONCEPTUAL, AND PRAGMATIC INFORMATION

Generalized utterances (GUs) correspond closely to the pattern-concept pairs of phrasal lexicons. They take the form of frame-like structures and are organized in a hierarchy. In addition to roles (slots) for a **pattern** and a **content** (the concept referred to), each GU has "deictic" roles for **speaker, listener, time,** and **focus set,** the set of elements currently under discussion, plus a **language** role to distinguish patterns in CHIE's two languages.[2] GUs correspond both to lexical entries and to higher-level syntactic patterns. Figure 1 shows how the GU for the Japanese verb *age-* (sentence 3b) joins a node in the conceptual hierarchy to the hierarchy of words. The "G" links connect concepts to their generalizations (types); the "=" links join equivalent concepts. Role names appear in lower-case. The portion of the conceptual hierarchy shown in the figure includes ATRANS, a transfer of something from one person to another (Schank & Abelson, 1977) and a specialization of this concept, ATRANS-FROM-ACTOR, an ATRANS in which the person from whom the object is transferred is viewed as the actor (Jacobs, 1985). U.AGE- represents a class of utterances which refer to an instance of ATRANS-FROM-ACTOR and have a word of type "AGE_" as the head of their patterns. "AGE_" represents verbs with stem *age-*. Two subtypes of "AGE_" are shown in the figure, including "AGEMASITA", the form used in 3b).

Generalized speech acts (GSAs) are prototypical associations of goals with GUs which constitute plans to achieve these goals. They function like the goal-plan associations in Wilensky (1983). Each GSA has roles for an **objective** of the speaker (the goal) and a GU **plan.** The goal in a GSA **intends** the plan; that is, the execution of the plan is a consequence of the planner's desire to achieve the goal (Dyer, 1983). Like GUs, GSAs are arranged in a hierarchy.

---

[2]For simplicity the language role is ignored in the remainder of this paper. See Gasser (1985).

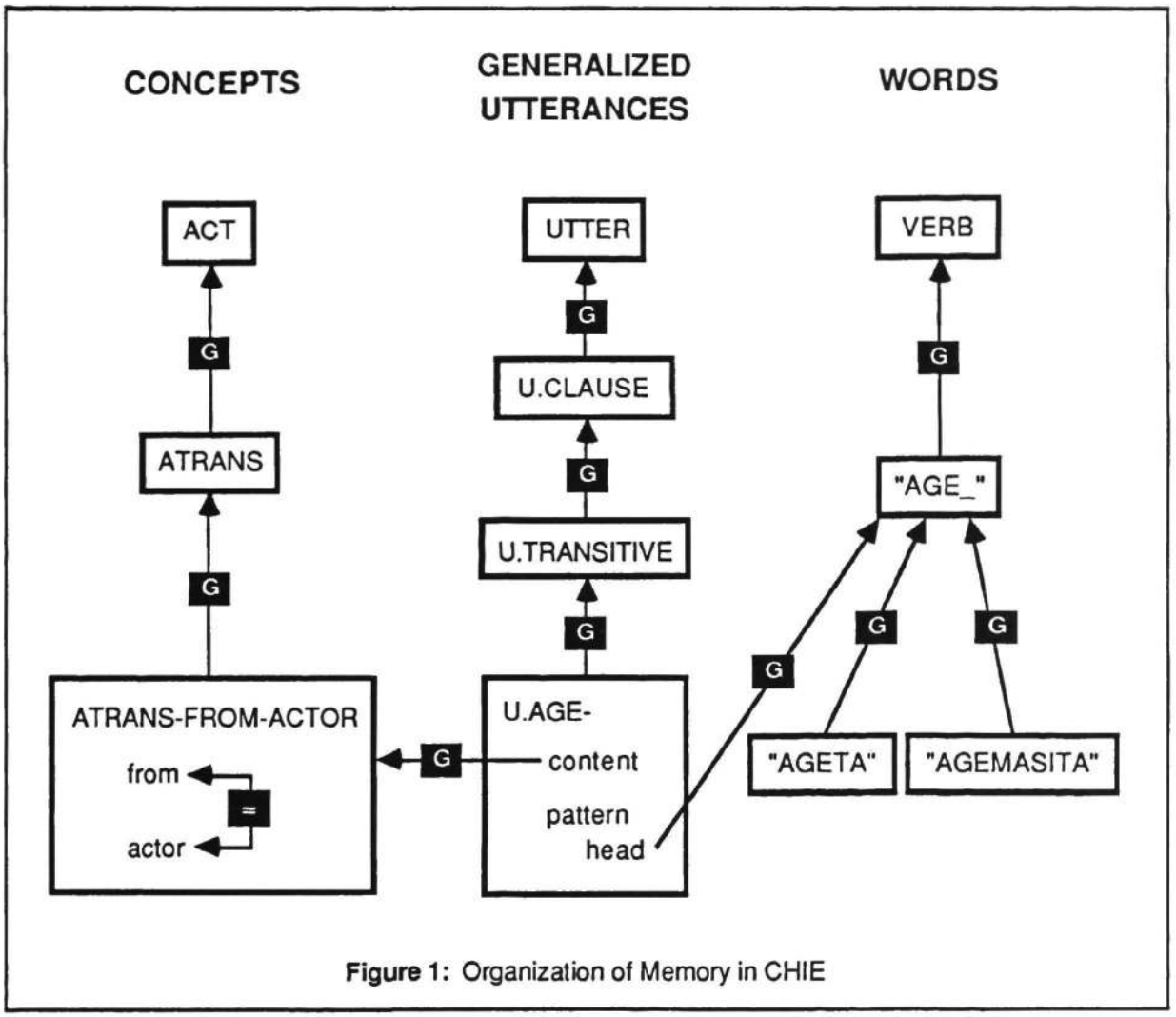


Figure 1: Organization of Memory in CHIE

In addition to the familiar illocutionary acts (Searle, 1969), GSAs include acts such as those realized by vocatives. One use of vocatives is embodied in the GSA SA.ADDRESS, shown in Figure 2. In this and subsequent figures restrictions on roles appear in parentheses following the role names. Thus the plan for SA.ADDRESS must be a kind of U.REFER. What SA.ADDRESS represents is roughly the following: for A to get B to ATTEND (Schank & Abelson, 1977) to an utterance of A's, A produces an utterance which refers to B (a U.REFER).

GUs and GSAs, like other nodes in the network, implicitly inherit all of the role information contained in their generalizations (Fahlman, 1979). For this reason a frame only makes reference to those roles which are restricted in that frame. For example, a general relationship which is true for all GSAs is that the person having the objective is also the speaker of the plan-utterance. This fact is not shown in Figure 2 because it is inheritable from the general GSA, SPEECH-ACT.

SA.ADDRESS is too general to provide specific information about the pattern for its plan-utterance, and it has specializations for particular categories of addressees and particular contexts. One of these is the GSA for addressing a judge during a trial, shown in Figure 3. SA.ADDRESS-TO-JUDGE is one of a set of speech acts which form elements of the trial script. *Your Honor* is appropriate only in the context of a trial and only for addressing the particular judge who is presiding at the trial. Embedding this GSA in the trial script allows us to represent these facts in a straightforward way.

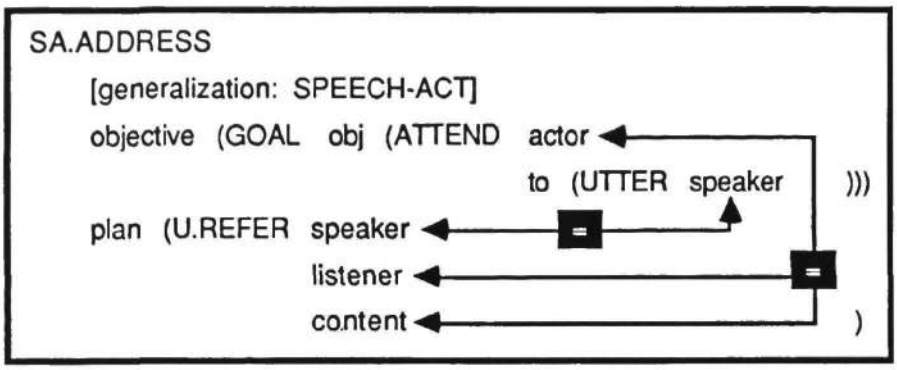


Figure 2: GSA for Address

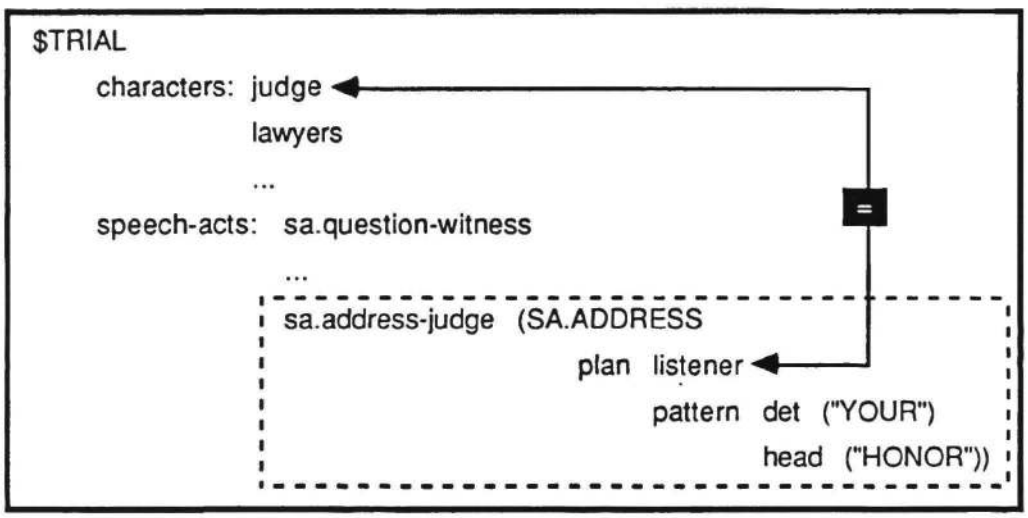


Figure 3: GSA for Addressing Judge

PROCESS OF GENERATION

Generation involves selecting GSAs and GUs in memory which match input goals of the speaker and conceptual content to be referred to. The GSAs and GUs yield patterns which are usually only partially specified. These are combined in a process analogous to **unification** (Kay, 1979) to produce complete patterns.

For the purposes of generation, GSAs are indexed by their objective roles and GUs by their content roles. Both may be indexed further in terms of the contexts in which they occur. Selection of an appropriate GSA or GU often involves classifying the goal or content instance on the basis of successive discrimination. In this sense the nodes in the GSA and GU hierarchies correspond to the nodes in a discrimination network (Goldman, 1975).

Consider the generation of *your Honor* in sentence 2). This occurs in the context of a trial, and we assume that the trial script has already been instantiated along with its various roles. The speaker's goal is that the judge attend to his/her question. This matches the goal in SA.ADDRESS. SA.ADDRESS has

specializations, distinguished on the basis of their contexts, so further classification is possible. One specialization is SA.ADDRESS-TO-JUDGE, which matches the speech act being formulated because it exists within the trial script and because the intended addressee is the current presiding judge. SA.ADDRESS-TO-JUDGE provides the full pattern for the utterance, *your Honor*. In most cases, one or more GUs would also need to be selected to complete the pattern.

DISCOURSE AND SOCIAL DEIXIS

Like other discourse deictics, *speak of the devil* makes reference to an element that is under discussion, i.e., a member of the current focus set (Grosz, 1984; McKeown, 1985). *Speak of the devil* is used to make the listener aware of a coincidence, namely, that a person who is a member of the focus set has just arrived on the scene. Since coincidences are usually worthy of mention, it is not surprising that languages have conventional ways of referring to particular kinds of coincidences. Thus corresponding to English *speak of the devil* is Japanese *uwasa o sureba kage*, literally, 'if (you) make a rumor (about a person), (his) shadow (appears)'. The GU for *speak of the devil* is shown in Figure 4. We represent a coincidence as a fact with three arguments, one for the common element and the others for two facts which are true of the common element. In the case of *speak of the devil* the common element is the person in question, one fact is the membership of this person in the current focus-set, and the other fact is the physical transfer of the person to the location of the speaker.

Consider now sentences 3a) and 3b). The verbs *kure-* and *age-*, like English *give*, both refer to the concept ATRANS-FROM-ACTOR, which is shown in Figure 1. What distinguishes the GUs for these two verbs is the provision regarding relative social distance for *kure-*. For such GUs we need to add the possibility of a **constraint**. A constraint is a role filled by a fact which associates other roles in the frame. Figure 5 shows the GU for this word. The constraint for U.KURE- is that the recipient of the ATRANS is socially nearer to the speaker than is the donor of the ATRANS. In generating 3a) the input concept matches the content roles of both U.KURE- and U.AGE-. The rule we apply in such cases is the **principle of specificity** (Anderson, 1983): when more than one item matches an input, select the one which better

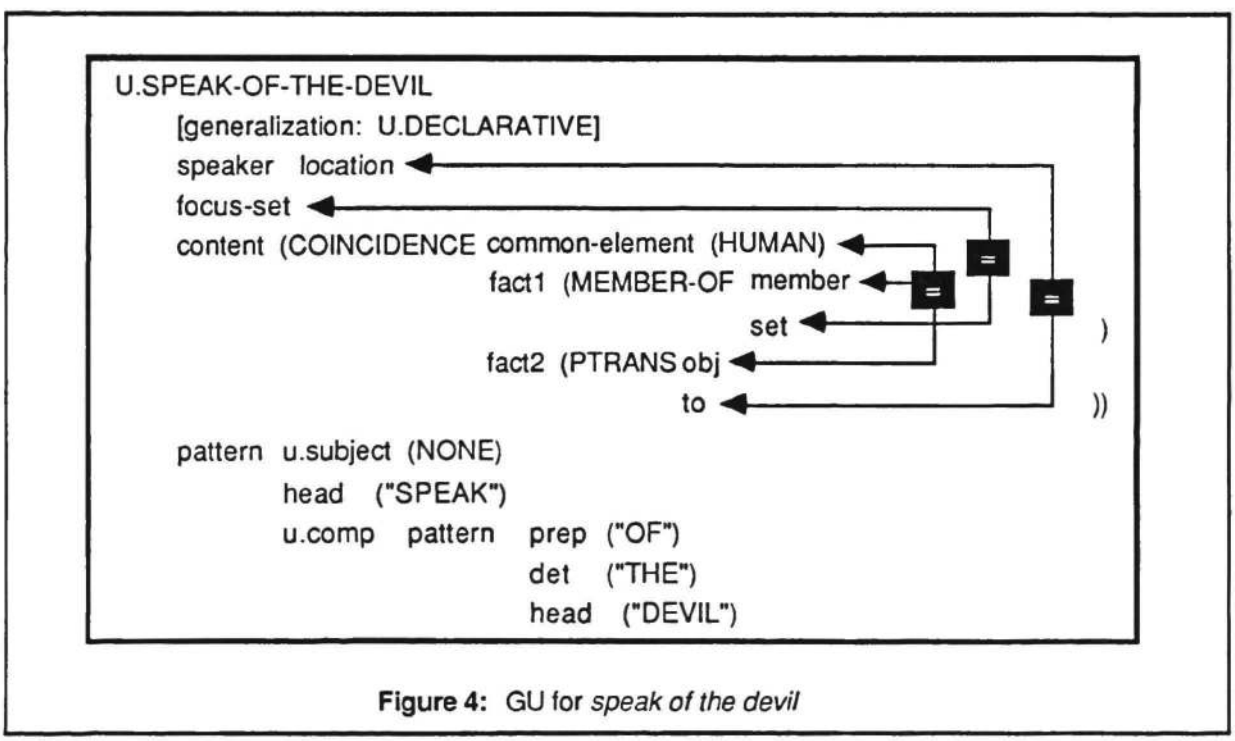


Figure 4: GU for *speak of the devil*

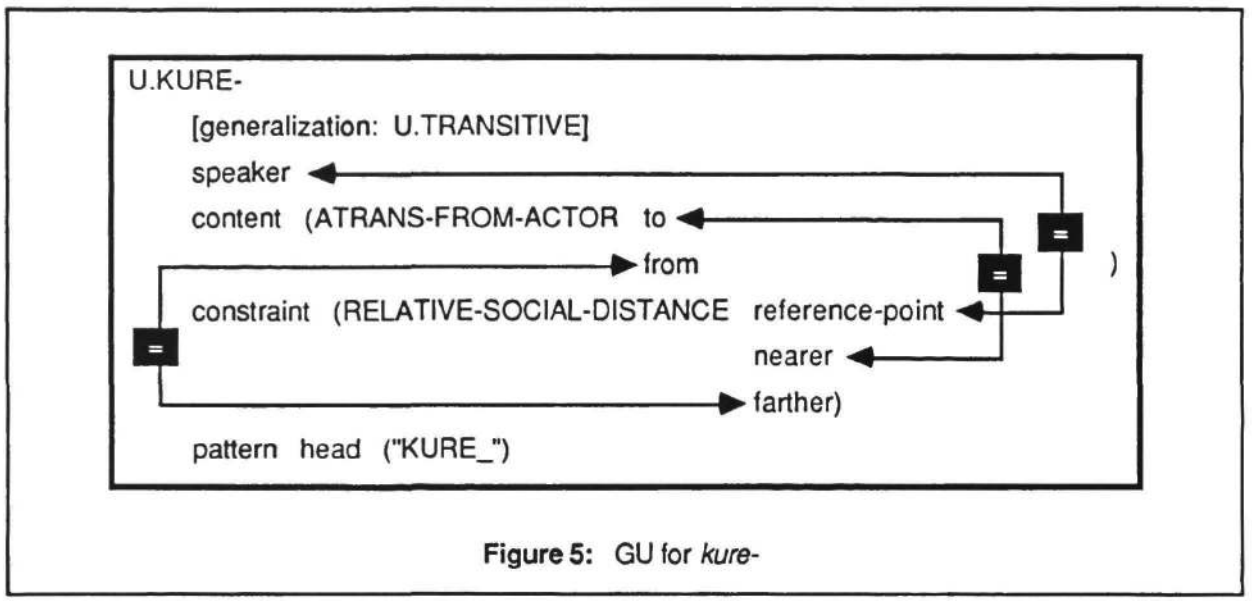


Figure 5: GU for *kure-*

characterizes the input. In this case this is U.KURE-, which matches on one more role than does U.AGE-.

INDIRECT SPEECH ACTS AND LITERAL MEANING

Now consider the directive in 4). For every directive there is a goal which the speaker hopes to achieve by having the listener perform some act. In other words, the plan to satisfy the top-level goal involves an **agency** subgoal (Schank & Abelson, 1977). The immediate goal behind a directive, then, is to have the listener *want* to perform the desired act. The GSA for directives is shown in Figure 6.

The specializations of SA.DIRECTIVE provide alternative ways of achieving the agency goal. These are distinguished on the basis of the speaker's expectation that the listener will perform the requested act. For **commands** the speaker is relatively certain of success. This expectation can derive either from the authority that the speaker has over the listener or the ease of the task. When the expectation of success is

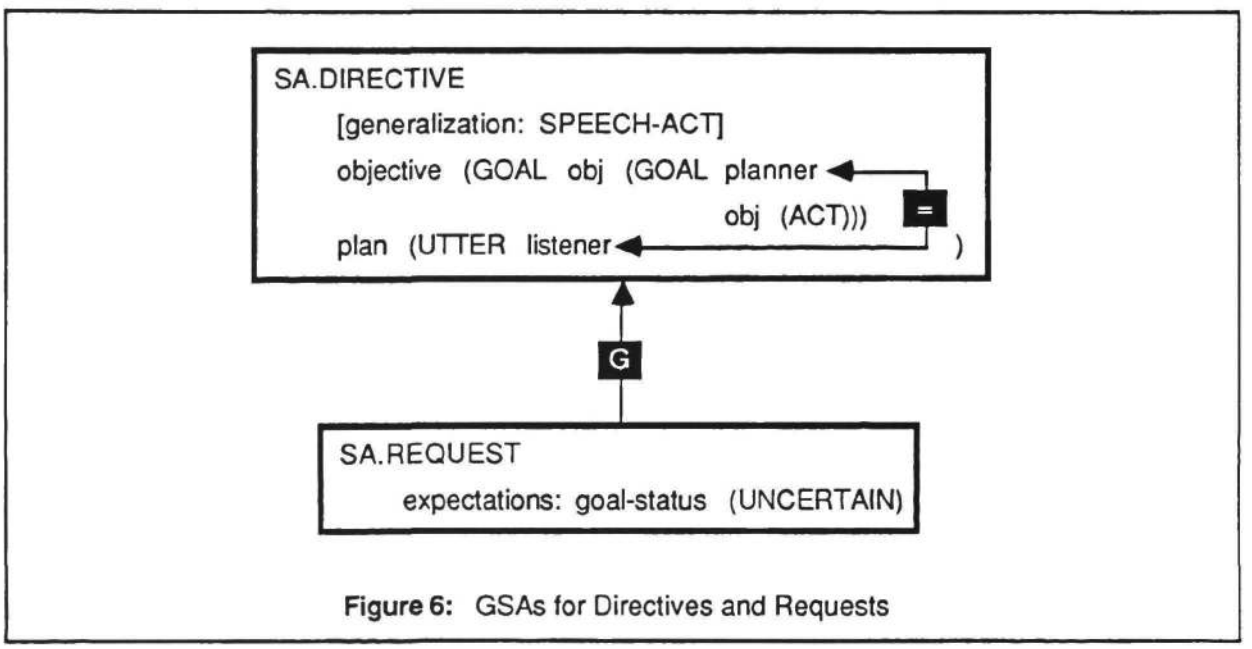


Figure 6: GSAs for Directives and Requests

not high, on the other hand, more polite directives, requests, are called for (Figure 6). Specializations of the request GSA specify particular request patterns. Figure 7 gives one of these, the GSA which is used in generating sentence 4). SA.WONDER associates its pattern with both its "indirect" and "literal" meanings. The indirect meaning is available because the objective of this GSA is that for requests. The literal meaning is accessible because the plan-utterance inherits from U.WONDER, the general GU for sentences with the verb *wonder*. The content of SA.WONDER's plan-utterance is inherited from U.WONDER. This concept, shown in the figure for convenience, is the literal meaning of *wonder*, i.e., a goal of the speaker to know some fact.

Note that the pattern provided by SA.WONDER is more general than *I was wondering if*. Rather it specifies that the subject refer to the speaker, that the head be a form of the verb *wonder*, and that the complement be a nominalized yes/no question. Thus the pattern also accommodates variants such as *we wonder whether*.

COMPARISON WITH OTHER APPROACHES

Existing natural language processing systems have been able to avoid some of the problems described in section 2 because they do not generally deal with unedited dialogue. Two areas that *have* received a good deal of attention, particularly in the language generation literature, are speech act planning (Appelt, 1985; Cohen & Perrault, 1979) and focus (Grosz, 1981; McKeown, 1985). While these researchers have made important progress in handling the pragmatic aspects of generation, all are limited by their failure to integrate the lexicon into the rest of memory. They maintain a functional separation of the grammar from the lexicon, and they make no attempt to incorporate lexical knowledge into other memory structures. This creates problems for representing lexical items that realize deictic or speech act functions or that are closely tied to particular social contexts. Work in the phrasal lexicon framework, on the other

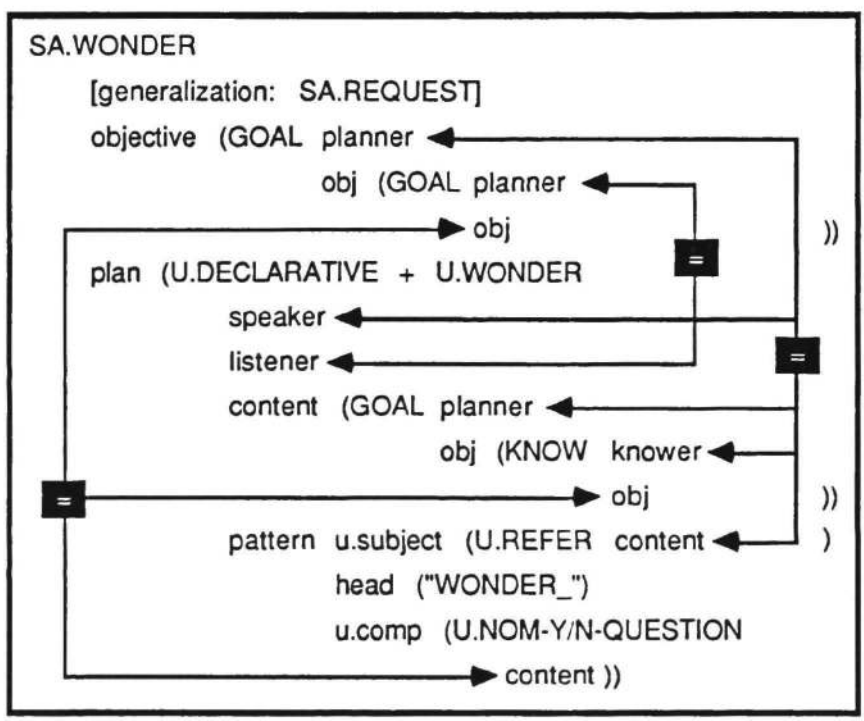


Figure 7: GSA for Requests Using *wonder*

hand, has to a certain extent eliminated the lexicon/phrase-structure distinction but has not generally addressed pragmatic issues.

In the **Lexical Memory** approach most patterns are amalgams of syntactic and lexical information. Those patterns which are purely syntactic, such as those specifying the structure for questions, are part of the same hierarchy as the lower-level "lexical" patterns. Because GUs have their deictic roles built in, any pattern may be assigned a deictic function. At the same time, because GUs fill the plan roles in GSAs, any pattern may also directly satisfy the goal behind a speech act type. The representation of context-oriented expressions like *your Honor* is also facilitated by the possibility of including GUs and GSAs as components of other memory structures, such as scripts.

The integrated nature of our approach distinguishes it in other ways from most mainstream linguistic theories. First, patterns in CHIE do not have a life of their own; they are defined only insofar as they fill a slot in a GU, which may in turn fill the plan slot in a GSA. In other words, they are always intimately associated with either semantic content or pragmatic intent. This contrasts with the autonomy of syntax position but is in agreement with "functional" views of syntax (e.g., Bates & MacWhinney, 1982; Langacker, 1986). Second, the fact that linguistic knowledge is represented in the same general framework as other knowledge and that lexical entries are integrated into conceptual memory is more than a computational convenience. This reflects an attempt to discover the extent to which linguistic knowledge shares properties with other kinds of knowledge. This approach stands in opposition to that of Chomskyan linguistics but is favored by an increasing number of linguists and psychologists (e.g., Anderson, 1983; Hudson, 1984; Langacker, 1986).

CURRENT STATUS AND FUTURE WORK

CHIE is implemented in AMI, a semantic network formalism based loosely on Fahlman's (1979) NETL. The program currently generates sentences for a small number of expressions of the type discussed in this paper. The work is still limited in several ways:

1. There is no account of where discourse goals come from; i.e., the program is not really ready to participate in conversations. A future goal of the project is to integrate the program into a model of the generation of conversation, such as that of Reichman (1985).
2. There is no facility for representing aspects of the listener's current knowledge state. Listener knowledge enters into the definition a number of lexical items, in particular, those of a "presuppositional" nature such as *again*. So that the program can make use of this sort of information in lexical entries, we are developing a procedure for "proving" whether the listener knows a certain fact or knows of a certain entity.
3. GUs and GSAs are *prototypical* associations, and CHIE cannot currently use them in atypical ways. A speaker may make *creative* use of an association in setting up an analogy between two contexts. The expression *your honor*, for example, might be used by a speaker addressing someone other than a presiding judge when a goal is to bring out the addressee's attempt to exercise judge-like authority in some context. We are currently working on a scheme for building analogies to satisfy generation goals.

CONCLUSIONS

In this paper we have presented an approach in which linguistic knowledge is integrated in various ways into the rest of memory:

1. GUs associate morpho-syntactic patterns directly with concept types.
2. GSAs associate goals with discourse plans that are realized as utterances. They are used in the same way as goal-plan associations that do not involve language.
3. Both GUs and GSAs are often embedded in and indexed to particular social situations.

CHIE extends the phrasal lexicon approach to allow lexical entries to make declarative reference to the elements of an utterance itself and to roles in the social context in which an utterance occurs. In addition, our approach has important implications for a model of language acquisition. GUs and GSAs consist of declaratively expressed generalizations about language behavior and thus present good candidates for the basic incremental units of language acquisition. Moreover, since GUs and GSAs may be components of non-linguistic memory structures, the acquisition of the linguistic units can be seen as part of the acquisition of the larger conceptual and situational structures. In this sense language acquisition becomes one aspect of learning about social interaction.

References

Anderson, J. R. (1983). *The architecture of cognition.* Cambridge, MA: Harvard University Press.

Anderson, S. R., & Keenan, E. L. (1985). Deixis. In T. Shopen (Ed.), *Language typology and syntactic description: Vol. 3. Grammatical categories and the lexicon.* Cambridge: Cambridge University Press.

Appelt, D. E. (1985). *Planning English sentences.* Cambridge: Cambridge University Press.

Bates, E., & MacWhinney, B. (1982). Functionalist approaches to grammar. In E. Wanner & L. R. Gleitman (Eds.), *Language acquisition: The state of the art.* Cambridge: Cambridge University Press.

Becker, J. D. (1975, June). The phrasal lexicon. *Proceedings of the Interdisciplinary Workshop on Theoretical Issues in Natural Language Processing,* Cambridge, MA.

Bresnan, J. W., & Kaplan, R. M. (1983). Lexical-functional grammar: A formal system for grammatical representation. In J. W. Bresnan (Ed.), *The mental representation of grammatical relations.* Cambridge, MA: MIT Press.

Cohen, P. R., & Perrault, C. R. (1979). Elements of a plan-based theory of speech acts. *Cognitive Science, 3,* 177-212.

Dyer, M. G. (1983). *In-depth understanding: A computer model of integrated processing for narrative comprehension.* Cambridge, MA: MIT Press.

Fahlman, S. E. (1979). *NETL: A system for representing and using real-world knowledge.* Cambridge, MA: MIT Press.

Fillmore, C. G. (1979). Innocence: A second idealization for linguistics. *Proceedings of the 5th Annual Meeting of the Berkeley Linguistic Society,* 63-76.

Gasser, M. (1985). Second language production: Coping with gaps in linguistic knowledge (Technical Report UCLA-AI-85-18). Los Angeles: University of California, Los Angeles, Computer Science Department.

Gibbs, R. W., Jr. (1980). Spilling the beans on understanding and memory for idioms in conversation. *Memory and Cognition, 8,* 149-156.

Goldman, N. M. (1975). Conceptual generation. In R. C. Schank (Ed.), *Conceptual information processing*. Amsterdam: North Holland.

Grosz, B. J. (1981). Focusing and description in natural language dialogues. In A. K. Joshi, B. L. Webber, & I. A. Sag (Eds.), *Elements of discourse understanding*. Cambridge: Cambridge University Press.

Hatch, E. M., Flashner, V., & Hunt, L. (1986). The experience model and language teaching. In R. Day (Ed.), *Conversation in second language acquisition*. Rowley, MA: Newbury House.

Hudson, R. (1984). *Word grammar*. Oxford: Blackwell.

Jacobs, P. (1985). A knowledge-based approach to language generation. Unpublished doctoral dissertation, University of California, Berkeley.

Kay, M. (1979). Functional grammar. *Proceedings of the Fifth Annual Meeting of the Berkeley Linguistic Society*, 142-158.

Langacker, R. W. (1986). Cognitive grammar. *Cognitive Science, 10*, 1-40.

McDonald, D., & Pustejovsky, J. (1985). Description-directed natural language generation. *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, 799-805.

McKeown, K. R. (1985). *Text generation*. Cambridge: Cambridge University Press.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication*. London: Longman.

Peters, A. M. (1983). *The units of language acquisition*. Cambridge: Cambridge University Press.

Reichman, R. (1985). *Getting computers to talk like you and me: Discourse context, focus, and semantics (an ATN model)*. Cambridge, MA: MIT Press.

Schank, R. C., & Abelson, R. (1977). *Scripts, plans, goals and understanding*. Hillsdale, NJ: Lawrence Erlbaum.

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge: Cambridge University Press.

Wilensky, R. (1983). *Planning and understanding: A computational approach to human reasoning*. Reading, MA: Addison-Wesley.

Wilensky, R., & Arens, Y. (1980). PHRAN--A Knowledge-based Approach to Natural Language Analysis (Memorandum UCB/ERL M80/34). Berkeley, CA: University of California at Berkeley, Electronics Research Laboratory.

Zernik, U., & Dyer, M. G. (1985). Towards a self-extending lexicon. *Proceedings of the 23rd Annual Conference of the Association for Computational Linguistics*, 284-292.

Repertory Actors' Memory for their Parts

William L. Oliver

University of Colorado

K. Anders Ericsson

University of Colorado

Most theories about how people perform everyday tasks and skills assume that knowledge is organized in memory so that related information can be activated or selectively retrieved as large structures. Entire constellations of knowledge (e.g., scripts or frames) are thought to be primed or activated in much the same way that words are primed by semantic associates. The idea that our memories are organized into large structures seems in agreement with our everyday experience. We have all experienced how memories suddenly rush into consciousness after brief encounters with once familiar material, like old lecture notes. Although there is no complete agreement on how memory is organized, it is generally agreed that many skills (e.g., chess, Chase & Simon, 1973) and language comprehension (e.g., Schank, 1982) require rapid access to very large amounts of information that has been previously acquired over periods of months or even years. Most laboratory studies of memory, however, look at memory for small amounts of material following brief learning episodes.

After speaking to several actors, it occurred to us that actors' memory for their parts provides an opportunity to study retrieval from large memory structures. Much of the information learned for parts is exactly specified by the scripts the actors studied. And it was evident from the performances that the actors gave on stage that they had learned their scripts very well. By knowing the content of a part, we could test the accessibility of information we knew the actors' had in memory.

Actors in repertory companies, must often remember parts from different plays and perform them during the same period of time. It is tempting to think that an actor's memory for different parts form distinct and very large memory structures. When an actor performs different roles, it is vital that the parts are remembered so that there are no confusions among them. Actors report to us that when they perform a particular role other roles in their repertory are completely blocked out, or, at least, not consciously accessed. In this sense, several parts in memory may be thought of as disassociated from one another.

The size of a memory structure corresponding to a part would have to be very large. To perform a major role, an actor must remember hundreds, even thousands of words spoken by his or her character. After studying and rehearsing a part, the actor must know what movements, gestures, and intonations of voice are to accompany almost every word spoken on stage. These details about the meaning and subtle nuances of performance are enacted in a very stereotyped way from one performance to the next, implying that these details are very well learned. Figure 1 represents schematically some of the information from an actor's part. Notice that the actor may represent information about his or her character or a scene somewhat apart from the actual wording of the part. Other information about specific gestures or emotions may be more directly encoded with the actual words spoken on stage.

It is not clear what information in memory actors must rely on to recall their parts. An actor may first have to recall what a speech is about, and then the meaning of the individual sentences in order to recall the exact words spoken on stage. In an extreme case, memory for a part might be context dependent (Godden & Baddeley, 1975) so that to recall the part efficiently the actor would have to be on stage interacting with other characters. Alternatively, actors might depend on a process of successive scanning (Rubin, 1977) where words or phrases would be cued by the words or phrases just uttered. In this case, the words would simply follow one another in much the same way that letters seem to automatically follow one another when we recite the alphabet. Notice that these different retrieval processes require that the actor be aware of differing amounts of information represented in

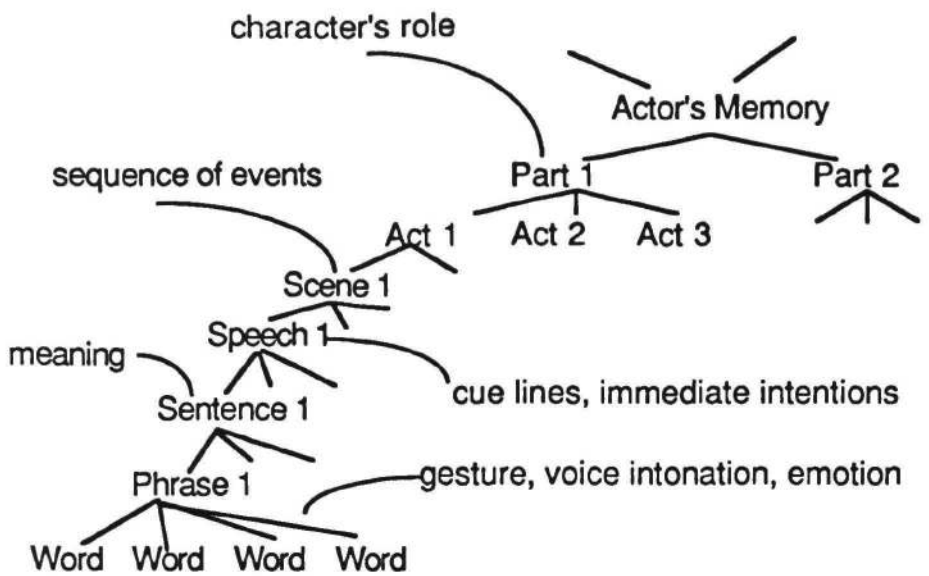


FIGURE 1. HYPOTHETICAL REPRESENTATION OF A PART

Figure 1 when he or she is recalling the verbatim wording of a part.

The subjects who participated in the present experiments were recruited from a local Shakespeare Festival held each summer in Boulder, CO. All of the subjects have received academic training and have considerable acting experience. We specifically recruited actors who either had very large parts (3000 or more words) or several reasonably large roles (1000 or more words). Before the experiments, we checked how accurately the actors could recite their parts from memory. We have found that the actors' memory of their parts is very exact, with only a very few words emended or substituted.

EXPERIMENT 1--DIRECT ACCESSIBILITY

The experiments we report in this paper used cued recall tasks to find out what information actors rely on to access the wording of their parts. We were surprised to find in our early studies that actors could respond correctly to probes consisting of only a few words taken at random from their parts (Oliver & Ericsson, 1984). These probes were so brief that they could provide little context or meaningful information for the subjects as retrieval cues. We will briefly describe a representative experiment in which the number of words in the memory probes was systematically manipulated.

Three actors participated in the experiment, each currently performing a large role. Equal numbers of one-, two- and four-word probes that occurred only once in the actors' parts were selected. The same line of text was never sampled more than once. An additional constraint required that no intervening punctuation appeared between any of the probe words or between the final word of the probe and the next word in the text, which served as the target word. Examples of some of these probes that were responded to correctly are shown in Table 1.

Probes were presented to subjects on a video display connected to a computer. The subjects were to respond out loud with the appropriate words from their parts as quickly as possible when the probes appeared on the screen. The elapsed times between the appearance of the probes and the voice responses were timed with a voice-actuated switch. On half of the trials the subjects were asked to report what they remembered thinking prior to responding to the probe. Unfortunately, we will not be able to discuss the analyses of these verbal reports in this brief paper. The subjects were given a maximum of 15 seconds to respond at which time the display went away and, depending on the condition, the subjects either gave verbal reports or quietly waited for the following probes.

TABLE 1. EXAMPLES OF PROBES FROM EXPERIMENT 1

born
for scaled
part of this anatomy

We found that longer probes were retrieved more successfully within 15 seconds than shorter probes, $F(2,360)=16.85$, $p<.001$. The one-, two-, and four-word probes were accurately retrieved .77, .91, and .99 per cent of the time respectively. There was also an effect of probe length on log-transformed reaction time, $F(2,276)=9.60$, $p<.001$, with the one-, two- and four-word probes being retrieved in 2.18, 1.93 and 1.67 seconds respectively.

These results indicate that very minimal cues can be used by actors to directly access their lines for many of the probes. Because the responses are extremely rapid, it is unlikely that large amounts of the parts are scanned, unless one assumes scanning times much faster than the usual scanning rates of approximately 38 milliseconds per item observed in Short Term Memory experiments (Sternberg, 1966). This direct accessibility came as something of a surprise to the actors themselves, as many of them claimed before the experiments that they would be unable to do the task. The additional words in the probes facilated performance by either providing additional surface structure or additional meaning to the subjects. It is important to reiterate, however, that the amount of information provided by a four word probe is minimal and that the probe could convey to the subject only fragmentary information about the speech it came from. Thus, the actors clearly need not depend on a context-dependent retrieval process.

EXPERIMENT 2--HIERARCHICAL ORGANIZATION

An actor's part is organized in way that would be expected to influence how it is represented in memory. A part itself is a unit distinct from other parts and is subsumed by the role the actor plays. The part is divided into acts and scenes, the scenes into speeches, the speeches into sentences, the sentences into phrases, and the phrases into words. This organization is shown schematically in Figure 1. Using the probe latency task, we have obtained evidence for the hierarchical organization of a part in memory. We will describe an experiment that extends some of our previous findings (Oliver & Ericsson, 1984).

All probes used in the experiment were four words in length and were drawn from the largest roles performed by the subjects. For half of the probes the subjects were to retrieve the words from their parts that immediately preceded the probes and for the other half they were to retrieve words that immediately followed the probes. Underlined blank spaces preceding or following the probes indicated to the subjects which words from their parts were to be retrieved. The first grouping of probes were all adjacent to sentence boundaries (demarcated by periods, question marks, and exclamation points) in the actors' parts. Sentence boundaries were selected randomly without replacement so that no more than one probe was ever associated with the same sentence boundary. Equal numbers of probes were selected such that they either immediately preceded or followed sentence boundaries or they were shifted one word to the left or right of the sentence boundary. For the second grouping of probes, equal numbers of four-word strings were sampled from the beginning and ends of speeches. When the probes began speeches, the subjects were to retrieve the last word of their immediately preceding speech; and, likewise, when the probes ended speeches, the subjects were to retrieve the first word of the speech that immediately followed. Figure 2 illustrates the various probe types. The probes were presented in a random order and the subjects were asked on half of the trials to give retrospective verbal reports.

It is clear from the examples given in Figure 2 that well-formedness of the probes differs systematically among the conditions. We expected that responding to well-formed probes in the forward direction (Forward-Within 1) would be the fastest condition and that retrieving backwards across

Forward-Across:
.....then he was when you sought him. **I** am.....

Forward-Within 1:
Do thou but close **our** hands with holy...

Forward-Within 2:
My dreams presage some joyful news at **hand**.

Backward-Across:
.....think it were not **night**. See how she leans her.....

Backward-Within 1:
Show me a mistress that is passing fair.....

Backward-Within 2:
.....and in **that** sparing makes huge waste.

Forward-Across Speeches:
.....to rejoice in splendour of mine own. [speeches by others] **Give** me a.....

Backward-Across Speeches:
...ground I cannot **move**. [speeches by others] I am too sore enpierced.....

FIGURE 2: EXAMPLES OF PROBES IN EXPERIMENT 2. THE PROBES ARE UNDERLINED; THE TARGET WORDS ARE IN BOLD LETTERING .

sentence boundaries (Backward-Across) would take much longer. Retrieval of the less well-formed units within sentences would fall somewhere in between, depending to some extent on whether retrieval was in the forward or the backward direction. Finally, we expected that subjects would be relatively slow at retrieving across speech boundaries. The experimental procedure was the same as in the earlier experiment.

The actors retrieved the correct responses 91 percent of the time. Due to a ceiling effect, analysis of accuracy is not particularly revealing. Analyses of reaction times, on the other hand, show clear differences between the boundary conditions, $F(7,206)=39.93$, $p<.001$. The mean reaction times for the different conditions are shown in Figure 3. Retrieving across sentence boundaries (Forward-Across and Backward-Across) was slower than any of the Within conditions. The reaction times for the different Within conditions did not greatly differ from one another. Crossing speech boundaries, particularly backwards, required a fair amount of time. It is interesting to note, however, that jumping to the first word of the following speeches took approximately the same amount of time as to go backwards across sentence boundaries to immediately adjacent words.

The relatively long times required by the subjects to cross speech and sentence boundaries indicates that processes beyond direct retrieval processes are sometimes required to retrieve words adjacent to the probes. Words within the same sentences can be retrieved quickly suggesting that they are encoded as part of the same unit that can be directly retrieved. These findings can be accounted for by assuming that the verbatim memory representation of a part is segmented into sentences and speeches. Sentences may in turn be segmented into phrase units or chunks, though additional findings or ours indicates that boundaries within sentences are not easy to identify. When the subject must retrieve a word from a chunk that is adjacent to the one the probe falls in, he or she must first retrieve the adjacent chunk and then unpack it. When normally reciting a part, associations between chunks, presumably in part mediated by meaning, permit orderly retrieval of the lines. Retrieval of the

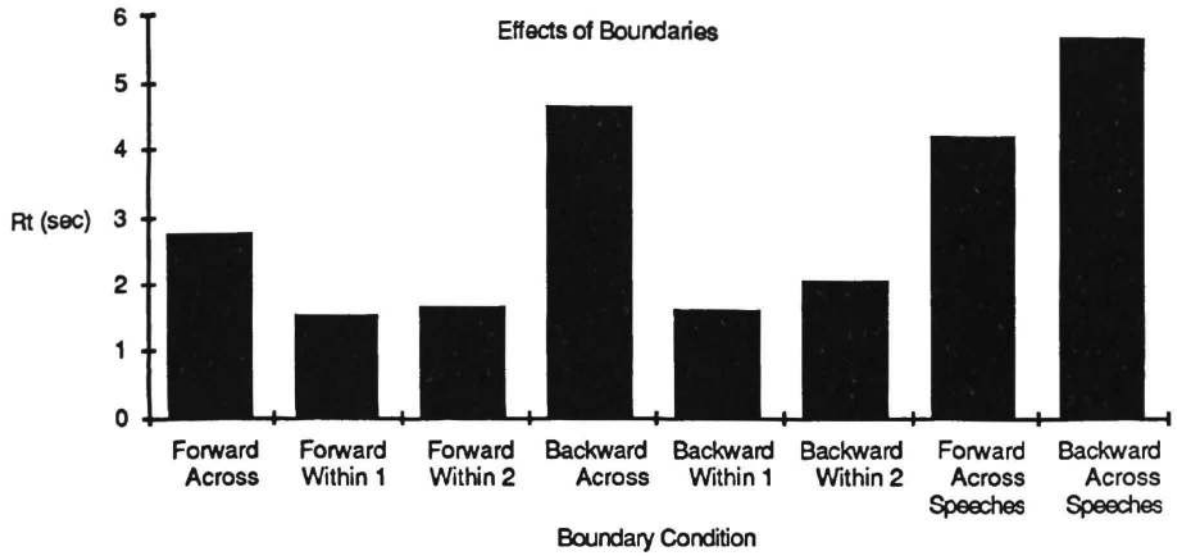


FIGURE 3. MEAN RETRIEVAL TIME AS A FUNCTION OF BOUNDARY CONDITION

beginnings of speeches, however, sometimes requires accessing organizing information in memory, such as cue lines spoken by other actors, action and positioning on stage, and so on. Thus, the hierarchical organization of the part shown in Figure 1 has clear consequences on retrieval, provided that the task requires that the subject make use of this organization, as, for instance, when jumping to the next sentence or speech. However, it appears that several words can cue recall of individual phrase units without mediation of other information in the hierarchy. The results from the present experiment also suggest that sequential scanning for information in parts must be a slow and effortful process because time-consuming retrieval processes are required to cross sentence and speech boundaries.

EXPERIMENT 3--INTERFERENCE BETWEEN PARTS

Thus far we have shown that actors can directly access very local, verbatim representations, i.e., phrases or sentences, of a single part on the basis of minimal cues. A natural extension to this line of research would be to look at retrieval of verbatim wording from more than one part in memory. In the first two experiments, subjects always knew what part the probes were drawn from, and thus could focus their search. By repeatedly retrieving from the same part, the memory for the part may have been activated or primed. The next three experiments that we will describe sought chronometric data to support our expectation that thematically organized information--in this case, parts and scenes-- might form memory structures that could be selectively searched or activated.

If a part forms a structure that can be strongly focused upon when it is searched in memory, there should be little interference from other parts in memory. Thus, when subjects know which part a probe comes from other instances of the same probe in a different memorized part should not interfere with retrieval. Lack of interference of this kind has been taken as evidence for separate memory structures consisting of thematically related concepts (e.g., Myers, O'Brien, Balota & Toyofuku, 1984) and entire lexicons of bilinguals (e.g., Scarborough, Gerard & Cortese, 1984).

Three actors were recruited for the present experiment who were performing two roles of comparable size in two different plays. Unique words were identified among the combined text of the two parts. In addition, words were identified that occurred exactly once in each of the two parts

(Overlapped probes). Probes were constructed by presenting the name of the part, e.g., DESDEMONA, followed either by the selected word (One-Word probe) or the selected word along with the previous word in the part (Two-Word probe) in equal proportions. The probes were presented in a random order, and the name of the part was presented to the subjects for three seconds before the memory probes appeared. The subjects were instructed to say out loud the words in their parts that immediately followed the probes. We chose to include two-word probes in the design because we were interested in whether two-word cues would form holistic cues that could be easily retrieved despite other occurrences of its constituent words elsewhere in the actors' parts.

The subjects were far less successful at retrieving the Overlapped probes (41%) than they were at retrieving Unique probes (74%), $F(1,82)=9.95$, $p < .005$. There were no main effects or interactions involving the number of words in the probes. Analyses of the response times yielded no reliable effects, perhaps because so few of the probes were retrieved resulting in reduced statistical power. The poor accessibility of the Overlapped probes strongly suggests that the part-information is not useful for retrieval in this task and that the principal cues for retrieval are the words.

EXPERIMENT 4--PART PRECUING

While it appears that actors may not always be able to restrict their search to a single part if instructed to, it is still unclear what role global information plays in the retrieval process. It was possible that telling the subject which part was to be retrieved from reduced, though did not eliminate, interference and that responding was generally facilitated. Global information was manipulated in the present experiment by precuing subjects with the name of the part they were to retrieve from on some trials, and telling them to retrieve from either of two parts on other trials.

The three subjects who participated in the previous experiment and an additional subject participated in the present experiment. From each actor's scripts equal numbers of unique words were selected from each part. Two-word probes were constructed by adding the subsequent words from their parts for half of these selected words. Subjects were shown the name of the part, e.g., VIOLA, or the word EITHER, followed by a one-word or two-word probe. On one half of the trials the part name preceded the probes, and on the other half of the trials the word EITHER preceded the probes. The subjects' task was to respond out loud with the words from their parts that followed the probes.

Probability of recall within fifteen seconds differed only as a function of number of words in the probe, $F(1,272)=18.41$, $p<.001$. The percentage of correctly recalled words were 62% and 86% for the one- and two- word probes respectively. The analysis of the log-transformed reaction times for correct retrievals showed a reliable difference for probe length, $F(1,202)=15.56$, $p<.001$, and the availability of part information, $F(1,202)=6.23$, $p<.05$, but no interaction between probe length and availability of part information. The average response time for the experiment was 2.37 seconds. Providing the part name facilitated response speed by .40 seconds; and providing an additional word to the one-word probes facilitated response speed by .35 seconds. In sum, adding an extra word to the one-word probes both reduced retrieval time and increased the probability of retrieval, whereas providing information about which of the two parts is to be probed only facilitated response speed.

EXPERIMENT 5--SCENE PRECUING

Because the precuing effect found in the previous experiment was relatively small, it is difficult to argue that part-information is important to the retrieval process. Part names may not be very effective cues because a part's representation in memory is too informationally dense (Hayes-Roth, 1977) or complex. Entire parts may simply encompass too many other units, such as speeches and scenes, to be activated as wholes. Scenes, however, may provide units that are better integrated. Reference to a scene might cue memory for constellations of events and emotions that could be held in mind and used as retrieval cues in a way that reference to an entire part could not.

In the present experiment three subjects were told on one quarter of the trials which of two scenes of a part the probe came from (Precue-Unique probes), and on a quarter of the trials they were

only told that it could come from either of two scenes (Either-Unique probes). Each probe appeared only once in the entire part. On the remaining one half of the trials, probes were presented that appeared once in a specified scene and once elsewhere in the part (Overlapped probes). Unlike previous experiments, only single words were used as probes. The subjects were always told which scene they were to focus on for these probes. Verbal protocols, as in the first two experiments, were collected on half of the trials. By having three types of probes we sought within the same experiment to see whether scene precuing could both facilitate retrieval and prevent interference.

The probability of successful recall within fifteen seconds did not greatly differ between the Precue-Unique probes (80%) and the Either-Unique probes (78%); however, the Precue-Overlapped probes (56%) were retrieved significantly less often, $F(2, 177)=587$, $p<.005$. An analysis of the log-transformed reaction times showed a similar pattern of results. The average response times for correct retrievals were 3.10, 2.97 and 3.85 seconds for the Precue-Unique, Either-Unique and Precue-Overlap conditions respectively, $F(2,121)=4.38$, $p<.05$. These results indicate that providing scene information did not significantly facilitate retrieval or prevent interference during the retrieval process.

CONCLUSION

The results from the last three experiments we have described taken together suggest that search cannot be focused at will upon a part or scene. The subjects were unable to search their memory for words from a part or scene without associative interference from another part or scene they had memorized. And, information specifying which part or scene was to be probed did not greatly facilitate retrieval. The small facilitation of part precuing on speed of retrieval stands as the single piece of evidence that parts can be separately activated. Further research is called for to better understand the conditions under which the effects of precuing are obtained.

Perhaps the most striking finding from the present study is that in relatively brief periods of time actors learn their parts so that they are directly accessible to minimal cues. A large body of information is simultaneously accessible and attention can be rather diffusely focused on several parts at once without seriously affecting retrieval. This finding is all the more surprising considering that the actors did not learn their parts with the aim of having direct access to individual phrase units.

The present study has focused on the accessibility of the verbatim wording of parts. As we pointed out, information about the plays' events, the characters' emotions, the meaning of all the lines, and so on, are also represented in the actors' memories. In other domains, the representation of knowledge is similarly complex, particularly in those domains where the expert must learn both rote knowledge for procedures and facts as well as knowledge that is used to understand problems. Thus, there may be interesting similarities between actors and experts, such as scientists, in how knowledge of a domain is represented and accessed in memory. We suspect that learning a huge body of directly accessible knowledge plays an important role in acquiring many skills. We should point out that learning this knowledge base may be much more difficult in many domains than it is for actors to learn their scripts for several reasons. First, experts must be concerned with what is worth learning and must invest much of their energies sifting through irrelevant facts. Second, the scripts actors learn do not differ from one study episode to the next, whereas the knowledge an expert must learn is less easily specified and must be learned in bits and pieces over a long period of time.

There are many obvious questions our study raises that we are not now in a position to answer. It is interesting to ask, for instance, whether models of memory can account for our findings. A number of current memory models are able to account for direct accessibility to large amounts of knowledge (e.g., Anderson, 1983; Eich, 1985; Raaijmaker & Shiffrin, 1980). It will be particularly interesting to look at issues of acquisition within the frameworks of these models. Other questions also concern acquisition: What amount of effort is required to learn a part so that it is directly accessible? How many parts can be simultaneously accessible? Yet other questions concern forgetting or maintenance of knowledge: Does memory for parts fade away like memory for isolated facts tend to fade away, or do parts fade from memory more or less as a wholes? At what point does an actor forget enough about a part so that it is no

longer directly accessible? What amount of practice is required to maintain a part in an accessible state? And finally, there are many questions about how meaning and affect are represented in actors' memory.

Note. This work was supported by the Office of Naval Research Grant N0014-84-K-0250, K. Anders Ericsson, Principal Investigator.

REFERENCES

Anderson, J. R. (1983). The architecture of cognition. Cambridge, MA: Harvard University Press.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. Cognitive Psychology, 4, 55-81.

Eich, J. M. (1985). Levels of processing, encoding specificity, elaboration and CHARM. Psychological Review, 92, 1-38.

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and under water. British Journal of Psychology, 66, 325-331.

Hayes-Roth, B. (1977). Evolution of cognitive structures and processes. Psychological Review, 78, 122-129.

Myers, J. L., O'Brien, E. J., Balota, D. A., & Toyofuku. (1984). Memory search without interference: The role of integration. Cognitive Psychology, 16, 217-242.

Oliver, W. L., Ericsson, K. A. (1984). Actors' memory for their parts. Talk presented at the 24th Annual Meeting of the Psychonomic Society in San Diego, November.

Raaijmaker, J. G. W., & Shiffrin, R. A. (1980). SAM: A theory of probabilistic search of associative memory. In G. H. Bower (Ed.), The psychology of learning and motivation: Advances in research and theory (Vol. 14, pp. 207-262). New York: Academic Press.

Rubin, D. C. (1977). Very long memory for prose and verse. Journal of Verbal Learning and Verbal Behavior, 16, 611-621.

Schank, R. C. (1982). Dynamic memory. New York: Cambridge University Press.

Scarborough, D. L., Gerard, L., & Cortese, C. (1984). Independence of lexical access in bilingual word recognition. Journal of Verbal Learning and Verbal Behavior, 23, 84-99.

Sternberg, S. (1966). High-speed scanning in human memory. Science, 153, 652-654.

A COMPUTATIONAL MODEL WHICH ADDRESSES ERRORS OF OVER-GENERALIZATION AND THEIR SUBSEQUENT DISAPPEARANCE IN EARLY CHILD LANGUAGE ACQUISITION

Jane C. Hill
Computer Science Department
Smith College

ABSTRACT

The model discussed here is offered as a prototype of the use of a computational model to explore alternate hypotheses and to suggest possible answers to some of the questions which have been addressed in the study of language acquisition. Why does not the child end up with an overly generalized grammar or lexicon? There is much evidence concerning the kinds of generalizations and over-generalizations that children make. However if we permit no overt and specific correction of the child's errors, then how is it that errors of over-generalization do not persist into adult speech? One answer to this question is proffered by attaching a system of weights to hypotheses. There are two related problems to be solved. Some mechanism in the model must allow erroneous hypotheses to be corrected; in addition there must be a way that more mature constructs can replace earlier ones. The model accomplishes these two tasks by means of a system of weights which represent confidence values and recency values. By this system more frequently matched constructs are preferred over less frequently matched constructs, and more recent hypotheses are favored for testing. This learning paradigm is illustrated by a set of procedures for learning the past tense of verbs in English. The scheme has the advantage that for a period of time when confidence factors are approximately in balance two or more constructs can co-exist. Thus we need not talk of rules or individual cases which have been learned or have not yet been learned but rather of a continuum in which rule schemas are either strong or weak.

WHY MODEL?

The development of computational models of language aquisition is a science which is in its infancy. The model discussed here is offered as a prototype of the use of a computational model to explore alternate hypotheses and to suggest possible answers to some of the questions which have been addressed in the study of language acquisition. Computational models such as ours may be used as a tool for linguists and psychologists in the task of exploring and contrasting alternative theories about the way in which children learn their native language. We would claim that such models enable the theorist to achieve a degree of explicitness which is virtually impossible to attain when working only with pencil and paper. Equally important are the

questions which such models raise which might in theory be overlooked but which must be addressed in the implementation of a computational model. Such models encourage a comparison of the results of a theory to actual data which in turn aids in refining and evaluating the theory.

There are of course many different approaches which a computational model of language acquisition may take. One approach is to proceed from a characterization of adult grammar and work backwards to see how the child might arrive at this characterization. An alternative approach, and the one embodied in our model, is to work forward from the evidence that the child provides toward some characterization of the adult language. Our model must ask not only under what conditions a language can be learned by a computational model, but also does the model learn the language in the same way that the child does? Particular attention must be paid to the errors that children make since child speech that differs from adult speech yields clues concerning the processes that the child is using to understand and produce speech. An equally important clue to the child's processes are those errors which the child typically does not make. One frequently noted phenomenon is the over-generalization of the plural of nouns and of the past tense of verbs in English. Such phenomena need explanation. Our computational model of language acquisition therefore must not only learn to understand and generate sentences of ever greater complexity as does the child, but to provide a satisfying explanation of the course of language acquisition the model must make the same kinds of errors that the child makes and eventually correct the errors after further learning has occurred. We have recently extended our model to attend to and learn past tense forms of verbs. It is gratifying to observe that the same processes which enabled the acquisition of a simple grammar extend very naturally to the learning of past tense verb forms.

In this paper we emphasize the somewhat different sorts of answers which computational models may suggest to traditional questions, principally because of an approach to language acquisition as acquiring a set of dynamic processes as opposed to acquiring a set of rules.

OVERVIEW OF MODEL

We will now describe our model of the acquisition of English in the two-year-old. Our model is an on-going research project, but the mechanisms discussed here have been fully implemented. We can give only a brief overview of the model and its assumptions in this paper. Readers interested in details of the model should consult Hill (1982, 1983) for full particulars of the model including examples of computer output together with corresponding linguistic data collected from a two-year-old child. The psychological validity of the model is defended in Hill and Arbib (1984) and in Hill (1984). The model is described as a member of the class of schema-theoretic models in Arbib, Conklin, and Hill (1986). A discussion of the use of the model in investigating the child's understanding and formation of coordinate structure is to be found in Hill (1985). It is characteristic of our model that the internal representation of the learning which takes place is more important than the output of the system, so the model must be described in terms of the

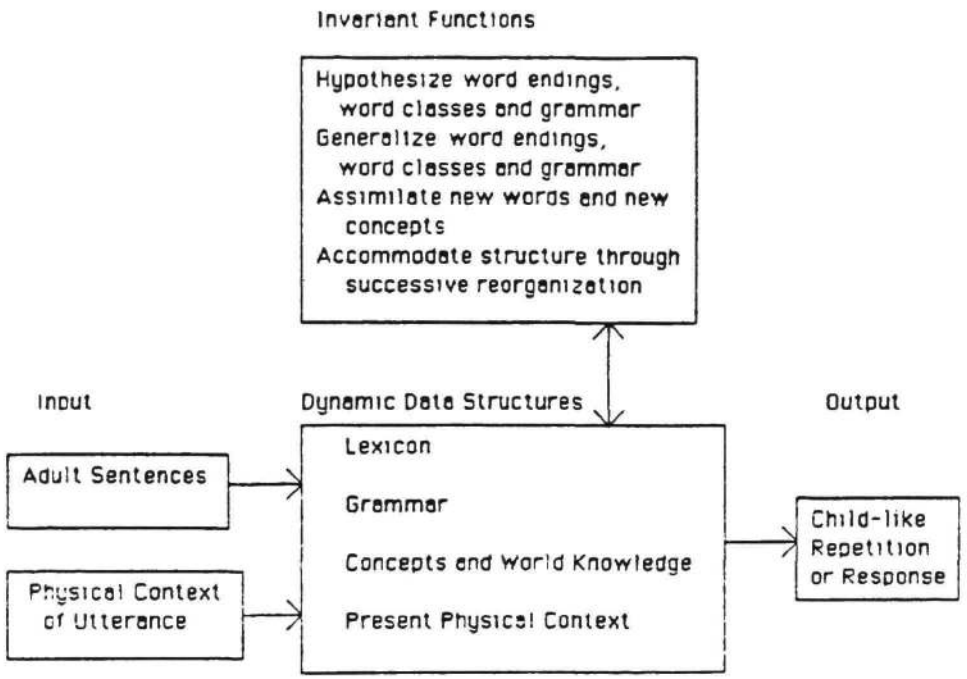


Figure 1
Components of Our Model of Language Acquisition in the Two-Year-Old

knowledge structures which are built as the model acquires language as well as in terms of the output.

Figure 1 provides a diagram of the components of the model. The model takes as its input adult sentences together with indications provided by the user, where relevant, of the physical context in which the sentences are uttered. Output from the model is a representation of child-like sentences repeating or responding to the adult input in accordance with the current state of the model's linguistic capacity. The child's knowledge is represented by dynamic data structures encoding the child's lexicon, the child's grammar, the conceptual knowledge of the child, and the physical context of the dialogue. The model is given a basic lexicon and a set of concepts with a mapping between the two. No assumptions have been made about the ultimate form of the adult grammar nor about what must be built-in to the model, but a precise account is kept of the knowledge and processes which are found to be necessary to be built-in to the model even for this elementary level of language understanding and production. Processes attend to the adult input and use rules of salience to focus on examples within the adult data which are used as the basis for language growth. The model is written in LISP using the semantic net language GRASPER (Lowrance 1978). The world knowledge is encoded in a semantic net as are the grammar templates and the lexicon.

The model uses its language experience to build a grammar which is at first a flat template grammar but which eventually evolves into a grammar which is actually a procedural grammar but which can be described by a set of recursive context-free phrase structure rules. The model notices and employs relations and word order, employs rules for concatenating relations and deleting words, and as the model grows word classes are formed, and new word forms are learned.

Let us explore the use of specific examples drawn from the input data. We will assume that gave has been identified as a relation-word and placed on the gleaning list. From an adult sentence such as "Daddy gave the toy to the boy" the model might initially respond with a single word such as toy. A subsequent presentation of the same sentence might cause the model to acquire a template for gave toy where gave would be classified as a relation-word and toy as a slot-filler. Yet another presentation of the sentence might cause the model to learn the template Daddy gave where Daddy was a slot-filler, and eventually the template (slot1 gave slot2) would be learned for Daddy gave toy. The learning is highly dynamic in that each time the same body of input is presented to the model a different set of grammar rules and additional lexical class information may be learned. What is learned in each presentation of the input depends upon the language experience of the model and what has been learned so far. No information is given the model about word classes, but hearing sentences such as "Mommy gave the toy," "John gave the book," "Sue gave the puzzle," would eventually cause the model to put toy, book, and puzzle all together in a word class meaning words which stand for possible objects of the relation-word gave. Note that it would not matter if the input sentences were far more complex than those used here for illustration. There is no requirement for correct and ordered exemplars. Typically we use as input adult sentences taken from a transcribed session of adult/child conversation. In the computer run from which illustrations have been chosen for this paper the adult sentences were taken from the Adam corpus which was collected by Roger Brown, Ursula Belugi, Colin Fraser, and Courtney Cazden prior to 1973 (Brown 1973) and which has been made available to us through the Child Language Data Exchange System. (For details see MacWhinney & Snow 1985). By the processes described above word classes are derived from the child's own ability to produce language. If the model is focussing on the word gave, then a sentence such as "Mommy gave the toy to Sue while she went into the store to buy groceries" would have just the same effect as the short sentences used above for an illustration. These processes result in a multiplicity of overlapping and intersecting word classes. The model requires schemas for word classification and template classification in order to grow, but the actual classes remain flexible. Processes of generalization eventually also permit the classifying of relation-words which might permit, for example, giving and bringing to be relation-words that could be classed together as words which have similar syntactic properties.

Successive reorganizations of the grammar and the lexicon occur as learning takes place. This process of gradual broadening of word classes and grammatical rules from applying to specific exemplars to sets of specific exemplars to more general categories has been defended in Kuczaj (1982), and Maratsos and Chalkley (1980). In this fashion the model suggests one way in

which language based initially on cognitive knowledge can grow into a syntactic system which will be increasingly independent of its semantic and cognitive foundations. It is important to note that although the rules embodied in the model are simple, their interaction is complex enough to necessitate the use of a computer model.

The model to date has attained only the level of a two-year-old producing sentences of up to six words in length. Initially the grammar acquired is entirely flat and is made up of rules for forming two-word utterances expressing relations and for combining those two-word relations into utterances which may be two to six words in length. It can understand and produce coordinate structure such as that found in the Adam corpus, ages two years three months up to two years eleven months, and with the enhancements to be described below it can now acquire a subset of English suffixes and prefixes.

DYNAMIC RULE SCHEMAS AND THE USE OF CONFIDENCE FACTORS IN THE MODEL

Why does not the child end up with an overly generalized grammar or lexicon? There is much discussion in the literature concerning the kinds of generalizations and over-generalizations that children make. (See for example Brown,1973, and deVilliers and deVilliers, 1978). We believe that it is important to focus on the errors that children make because of the insights which they yield concerning the processes that the child employs in language acquisition. Bowerman (1974) states this position very clearly. A study by Bybee and Slobin (1982) presents a careful examination of the acquisition of irregular past-tense forms of verbs in English. If, however, we permit no overt and specific correction of the child's errors, then how shall we explain the fact that errors of over-generalization do not persist into adult speech? Consider the verb, break. It is an empirical fact that children at the earliest stage of language acquisition typically learn the word broke and seem to use it correctly. One may assume that such forms have been learned by rote. Then at a subsequent stage of development the child will start to use the word breaked. Presumably this is because the child has formed a general schema for forming the past tense of verbs. Eventually of course children learn that break is an irregular verb and does not obey the general rule in the form of its past tense. But the puzzle is that for a period of time, sometimes for years, both forms exist in the child's vocabulary. How can this period of imbalance between the erroneous and the correct forms be explained? This behavior cannot be explained if the language mechanism is expressed in terms of explicit rules which the child either knows or does not know.

One answer to this question is proffered by our computational model which attaches a system of weights to hypotheses about word forms and grammar rules. These weights are meant to represent the relative strength of the various hypotheses. (An alternative model of this phenomenon is offered in the connectionist model of Rumelhart & McClelland, 1987, to appear.) There are two related problems to be solved. Some mechanism in the model must allow erroneous hypotheses to be corrected; in addition there must be a way that

more mature constructs can replace earlier ones. Our model accomplishes these two tasks by means of a system of confidence values and recency values. A confidence value is associated with each hypothesis and this confidence value is increased each time the hypothesis is instantiated in the adult speech input; the confidence value is increased to a lesser degree each time the hypothesis is instantiated in the child speech output. In this way more frequently matched constructs come to be preferred over (given a higher confidence factor than) less frequently matched constructs. Hypotheses must be reinforced to survive. If new hypotheses, however, are to start with very low confidence values they will have trouble "catching up" with earlier hypotheses. For this reason separate recency values are employed whose function it is to cause more recent hypotheses to be favored for testing. We will refer to The combination of confidence values and recency values as confidence factors. The use of weights to direct learning in computational models is by no means unique. Similar weighting schemes have been employed in many computational models. (See, for example, Kelley, 1967.) By means of the use of confidence factors our model attains the desired effects.

To illustrate, if the model were given a lexical pair such as tie and untie and the knowledge that untie is the reverse process of tie, then the model might hypothesize a set of similar process pairs such as cover/uncover, drop/undrop, hang/unhang. It might be that only after a great deal of language experience will the hypothesized lexical entries unhang and undrop be forgotten. This general scheme has the advantage that for a period of time when confidence factors are approximately in balance two or more constructs can co-exist, as, for example, in the case of the past-tense verb broke and the over-generalized form breaked for the verb break. Thus we need not talk of rules or individual cases which have been learned or have not yet been learned but rather of a continuum in which rule procedures are either strong or weak.

LEARNING PAST-TENSE VERB FORMS IN ENGLISH

It is especially interesting to explore the use of verbs in English in the developing language of the child since learning English is intimately tied to the learning of verbs. DeVilliers (1985) has found evidence that input language has a significant impact on the child's developing language with respect to verbs. The mother's use of verbs is a high predictor of the child's use of verbs. Note that it is not the frequency of the mother's use, but rather the variety of verb forms in the mother's use of a particular verb which is a significant predictor of the variety of forms the child will use for that same verb. This is interpreted to mean that the child is monitoring the input for clues about the grammatical prototypicality of forms of individual verbs. Wide differences in the use of verbs between subjects were found in the samples considered in her study. Verbs with a variety of heard uses were used with greater confidence by the child even in unheard contexts. Our model simulates this monitoring process. DeVilliers' analysis did not address the issue of over-generalization but it does lend credence to our processes which rely on the information gleaned from the input by focussing on different constructs at different times for the learning of the forms.

The learning paradigm is as follows:

1. Observe a correlation between past-tense verbs and **-ed** endings

2. Place **-ed** at the end of all action-verbs and give a modest confidence factor to all these past-tense forms

3. Proceed to modify the confidence factors depending on the experience of the model as follows.

   - add a small increment to the confidence factor of a given form each time the model produces that past-tense form

   - add a larger increment to the confidence factor of a given form each time the model notices that form in the adult input

Figure 2
Learning Past-Tense Forms of Verbs in English Through Use of Dynamic Rule Schemas and Confidence Factors

Figure 2 summarizes our model's paradigm for learning past-tense forms of English verbs. In order to observe the correlation between past-tense and -ed endings, the model must be given a representation of time-past in its cognitive knowledge, and the ability to identify action verbs in its lexicon. The model simulation begins by forming past-tense entries in its lexicon for all action verbs simply by adding -ed endings. Each of these forms is initially given a modest confidence factor. The model then proceeds to modify the confidence factors of the past-tense forms depending upon its language experience. A small increment is added to the confidence factor of a form each time that the model produces a past-tense form; a larger increment is added each time that the model finds a past-tense form in the adult input.

The choice of past-tense form is entirely dependent upon the input sentences used, so no conclusions can be drawn about the specific past-tense forms which are learned, but depending upon the input data

- The model may keep an erroneous -ed ending

- The model may proceed through a period of instability in which it will vacillate between -ed and irregular forms

- The model may discard the erroneous **-ed** form and replace it by the irregular form

Figure 3
Observed Behavior of the Model with respect to past-tense verb forms in English

Sample Model Output -- First Iteration

Don t you think you should pick it up and put it in the basket?
Basket

See the other one and put them together?
Together

We ll take his coat and hang it up all right?
Up

When you open the door and go outside what do you put on?
On

Oh Dale sit in the chair and watch the game?
In chair

Won t you please go over and take Ursula s pocketbook to her?
Pocketbook

Do you want Mommy to take it and put it up?
Up

Do you want Mommy to take the whistle and put it up?
Up

Why don t you pick the books up and take them over to Ursula?
Over

Yes, you fell down and broke your horn didn t you?
Horn

Yes, you put them in your bank, remember, and it broke your bank?
Bank

Do you want Mommy to take it and put it up?
Puted up

Remember we took the basket to her?
Basket

If you fall down you ll break the horn
Horn

Sample Model Output -- Second Iteration

Don t you think you should pick it up and put it in the basket?
Picked up and puted in basket

See the other one and put them together?
Together

We ll take his coat and hang it up all right?
Taked coat

When you open the door and go outside what do you put on?
Puted on

Oh Dale sit in the chair and watch the game?
In chair and watched game

Won t you please go over and take Ursula s pocketbook to her?
Taked pocketbook

Do you want Mommy to take it and put it up?
Puted up

Do you want Mommy to take the whistle and put it up?
Puted up

Why don t you pick the books up and take them over to Ursula?
Picked books

Yes, you fell down and broke your horn didn t you?
Broked horn

Yes, you put them in your bank, remember, and it broke your bank?
Broked bank

Do you want Mommy to take it and put it up?
Taked up and puted up

Remember we took the basket to her?
Tooked

If you fall down you ll break the horn
Failed down and broke horn

Sample Model Output -- Final Iteration

Don t you think you should pick it up and put it in the basket?
Picked up and put in basket

See the other one and put them together?
Together

We ll take his coat and hang it up all right?
Taked coat

When you open the door and go outside what do you put on?
What puted on

Oh Dale sit in the chair and watch the game?
In chair and watched game

Won t you please go over and take Ursula s pocketbook to her?
Taked pocketbook

Do you want Mommy to take it and put it up?
Taked up and puted up

Do you want Mommy to take the whistle and put it up?
Taked whistle and puted up

Why don t you pick the books up and take them over to Ursula?
Picked books

Yes, you fell down and broke your horn didn t you?
Broke horn

Yes, you put them in your bank, remember, and it broke your bank?
Broke bank

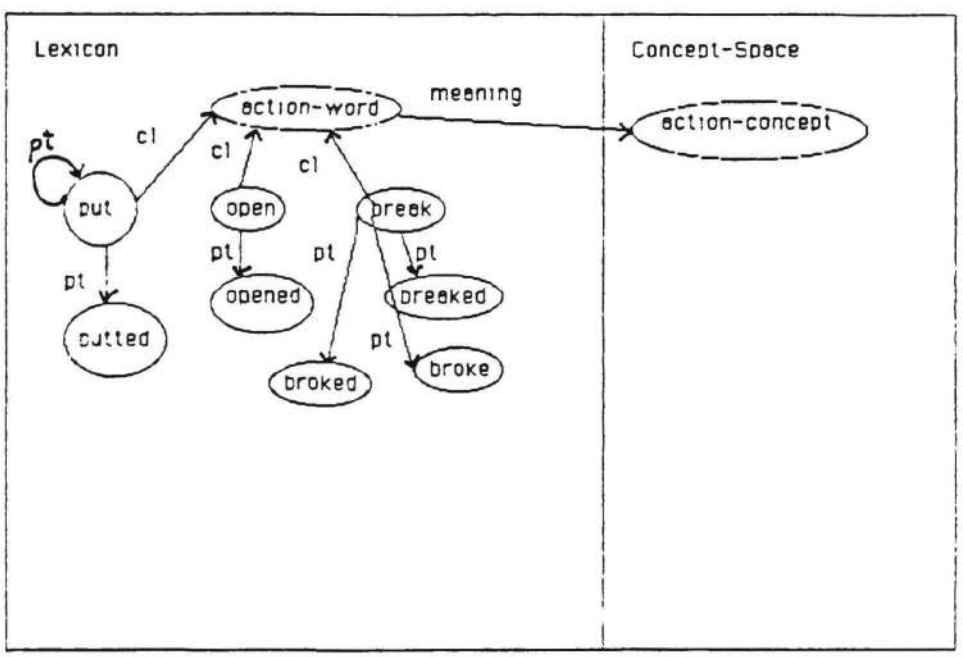
Do you want Mommy to take it and put it up?
Taked up and put up

Remember we took the basket to her?
Tooked basket

If you fall down you ll break the horn
Fell down and broke horn

Figure 4

The resulting behavior of the model is summarized in Figure 3. Since the choice of past-tense form is entirely dependent upon the input to the model, no conclusions can be drawn about the the specific past-tense forms which are learned, but depending upon the input data, the model may (1) keep an erroneous -ed ending, may (2) proceed through a period of instability in which the output vacillates between an erroneous -ed ending and the correct irregular form, or (3) the model may discard the erroneous form and replace it by the irregular form. All these forms of behavior are exhibited in the sample output from a series of computer runs which is presented for illustration in Figure 4. To make apparent the changing preferences of the model, only past-tense forms are shown in the output exhibited here. A portion of the internal representation of the model is graphically represented in Figure 5, the information given the model for this illustration is presented in Figure 6, and a summary of the confidence factors attached to the different past-tense forms is summarized in Figure 7.



cl = class
pt = past-tense form

Figure 5
A Schematic Representation of a Portion of the
Internal Representation of Our Model

Since developing the capacity to learn the past tense forms of verbs, we have successfully applied the same strategies to the learning of plural forms of nouns which name physical objects. We believe that this ability to generalize the processes employed by the model lends support to the validity of the processes.

Lexicon
away, bank, basket, bat, book, break, broke, chair, coat, come, Daddy, Dale, do, door, down, fall, fell, game, glove, go, hang, here, horn, in, Mommy, on, one, open, other, out, outside, over, paper, pick, pocketbook, put, racket, see, sit, take, that, time, together, took, try, up, use, Ursula, want, watch, what, when, whistle, you

Relation list
(possn-rein $owner), (breaking-rein break),(broking-rein broke), (coming-rein come), (doing-rein do), (down-rein down), (falling-rein fall), (felling-rein fell), (going-rein go), (hanging-rein hang), (here-rein here), (in-rein in), (on-rein on), (opening-rein open), (out-rein out), (outside-rein outside), (over-rein over), (picking-rein pick), (putting-rein put), (seeing-rein see), (sitting-rein sit), (that-rein that), (taking-rein take), (tooking-rein took), (trying-rein try), (up-rein up), (using-rein use), (wanting-rein want), (watching-rein watch), (what-rein what), (when-rein when)

Feature list
person:Dale, Daddy, Mommy, Ursula
place: here, outside

Slot-filling features
possn-rein $owner person
here-rein $slot1:pointing-obj

Closed class
and

Figure 6
Information Given the Model for this run

	First Run	Second Run	Final Run
breaked	.8	.8	.8
broke	.955	.99 *	.995 *
broked	.957 *	.957	.957
falled	.909 *?	.909	.909
fell	.75	.958 *	.981 *
felled	.909 *?	.957	.956
opened	.875	.941	.961
picked	.955	.983	.992
put	.944	.984	995 *
puted	.967 *	.99 *	.993
taked	.96 *	.986 *	.987
took	.875	.958	.979
tooked	.941	.976	.993 *

* wins

Figure 7

Confidence Factors Attached to Past-Tense Forms

CONCLUSION

We would emphasize that the purpose of this illustration is not to make any specific claims about the learning of past-tense forms in English, but rather to illustrate that a paradigm such as ours may be sensitive to the input data and my exhibit varied behavior. The use of dynamic rule schemas and confidence factors has been used to model the phenomena of generalization, over-generalization, and subsequent correction of over-generalized forms.

This discussion has been offered as an example of the kind of answers that computational models can proffer for consideration and experimentation. Other issues which the same model explores are (1) what variation occurs in the model as specific constraints are built-in or omitted, (2) how can the use of input filters focus on different aspects of the input data over time, (3) how does variation in meaning representation and sets of semantic features affect the learning process. We believe that the development of models such as ours will have a large impact on future work in language acquisition.

REFERENCES

Arbib, M. A., Conklin, J. E., and Hill, J. (1986) From Schema Theory to Language, Oxford University Press, New York, NY.

Bowerman, M. (1974) "Learning the Structure of Causative Verbs: A Study in the Relationship of Cognitive, Semantic, and Syntactic Development", in E. Clark, ed., Papers and Reports on Child Language Development, no. 8, Stanford University Committee on Linguistics, CA, pp. 142-178.

Brown, R. (1973) A First Language: The Early Stages, Harvard University Press, Cambridge, MA.

Bybee, J., & Slobin, D. (1982) "Rules and Schemas in the Development and Use of the English Past Tense", Language, vol. 58, 2, pp. 265-289.

deVilliers, J. (1985) "Learning How to Use Verbs: Lexical Coding and the Influence of the Input", Journal of Child Language, 12, pp. 587-595.

deVilliers, J., & deVilliers, P. (1979) Early Language, Harvard University Press, Cambridge, MA.

Hill, J. (1985) "Using a Computational Model of Language Acquisition to Address Questions in Linguistic Inquiry", Proceedings of the Seventh Annual Conference of the Cognitive Science Society, Irvine, CA.

_____ (1984) "Combining Two Term Relations" Evidence in Support of Flat Structure," Journal of Child Language, 11,pp. 673-678.

_____ (1983) "A Computational Model of Language Acquisition in the Two-Year-Old," Cognition and Brain Theory, 6(3),pp. 287-317.

_____ (1982) "A Computational Model of Language Acquisition in the Two-year-old," Ph.D. Dissertation, University of Massachusetts at Amherst, reproduced by the Indiana University Linguistics Club, Bloomington IN, February 1983.

Hill, J., & Arbib, M. A. (1984) "Schemas, Computation, and Language Acquisition," Human development, 27,pp. 282-296.

Kelley, K. (1967) "Early Syntactic Acquisition", Ph.D Dissertation, University of California at Los Angeles, also published as Report No. P-3719, The Rand Corporation, Santa Monica, CA.

Kuczaj, S.A. II, (1982) "On the Nature of Syntactic Development", in S. Kuczaj, ed, Language Development: Volume 1: Syntax and Semantics, Lawrence Erlbaum Associates, Hillsdale, N.J.

Lowrance, J. (1978) GRASPER 1.0 Reference Manual, COINS Technical Report 78-20, University of Massachusetts at Amherst (December 1978).

MacWhinney, B., & Snow, C. (1985) "The Child Data Exchange System", Journal of Child Language, 12, pp. 271-296.

Maratsos, M., & Chalkley, M. (1980) "The Internal Language of Children's Syntax: The Ontogenesis and Representation of Syntactic Categories", In K. Nelson, ed., Children's Language: Vol.2, Gardner Press, N.Y.

Rumelhart, D., & McClelland, J. (1987, to appear) "Learning the Past Tenses of English Verbs: Implicit Rules or Parallel Distributive Processes?", In B. MacWhinney, ed., Mechanisms of Language Acquisition, Lawrence Erlbaum Associates, Hillsdale, NJ.

REMARKS ON THE PSYCHOLOGICAL REALITY OF THE SUBSET PRINCIPLE: ITS RELATION TO UNIVERSAL GRAMMAR AS A MODEL OF THE INITIAL STATE[1]

Barbara Lust
Department of Human Development and Family Studies
Cornell University

Recent theory of language learnability has argued that the "Subset Principle", as defined in 1, significantly constrains the language learner's induction from primary language data (PLD) to the knowledge of language. This "Subset Principle" was argued formally to maximize learnability probabilities, in fact to be "necessary and sufficient for identifiability from positive evidence" given an inductive and deterministic mechanical model of language learnability (Angluin, 1980). Recently, there has been additional argumentation that the Subset Principle (SP) empirically characterizes acquisition of natural language (Berwick, 1982, 1985). In fact, it has been argued that "a single characterizing condition, the Subset Principle, subsumes all acquisition ordering constraints that have been proposed in the linguistic literature... This result is not surprising, since the Subset Principle is a necessary condition for acquisition from positive-only evidence" (Berwick, 1982, 240, cf. 1985, 275).

(1) Condition 1. An indexed family of nonempty languages satisfies Condition 1 if and only if there exists an effective procedure which on any input $i \geq 1$ enumerates a set of strings T_i, such that
i) T_i is finite
ii) $T_i \subseteq L_i$, and
iii) for all $j \geq 1$, if $T_i \subseteq L_j$ then L_j is not a proper subset of L_i
Angluin, 1980,121

Angluin (120) states that "informally, this condition requires that for every language L of the family there exists a 'telltale' finite subset T of L, such that no language of the family that also contains T is a proper subset of L. Moreover, it must be possible to enumerate effectively some such telltale finite set from any index for L. The point of the telltale subset is that once the strings of that subset have appeared among the sample strings, we need not fear "overgeneralization" in guessing L. This is because the true answer, even if it is not L, cannot be a proper subset of L. In this case, we will eventually see a conflict between the data and L, which will force us to change our guess." Intuitively, given a hierarchy of language classes, where $Li \subset Lj \subset Ln$, and a procedure for 'guessing' possible language classes on the basis of input data, such 'guesses' will be constrained by condition 1 to choose the smallest possible language class which includes the input data. The Subset Principle (SP) thus guarantees that an initial hypothesis (H) in 'guessing a language' on the basis of limited data will not be "too large," thus requiring correction by negative evidence.

If the Subset Principle (SP) can be validated psychologically as well as formally, this will provide a major advance in the study of acquisition of natural language for several reasons. (i) It is known that the set of hypotheses available to the language learner must be limited in order to

account for the acquisition of language in finite time. The SP would provide restrictions on both the number and the nature of possible hypotheses at any one time. (ii) Since the SP ensures inductive inference from positive-only evidence, it proposes a solution to the well known problem of learnability in finite time based on induction from input which does not include negative evidence (e.g., Gold, 1967). Gold shows that mechanical learnability in finite time of any class of languages over an alphabet which contains every finite L together with at least one infinite L is impossible in this case. The empirical facts of first language acquisition document little or no significant negative evidence in the acquisition process (e.g., Brown & Hanlon, 1970). (3) Since the SP principle formalizes a constraint on induction, it would provide a significant complement to UG which is primarily deductive. In fact, some constraint on induction is a necessary complement to UG, if UG is to be "psychologically real", i.e., if UG is to be consistent with the actual mapping from PLD to a specific language grammar in first language acquisition. This mapping occurs in real time, and thus necessarily involves successive hypothesis formation on the basis of successive input. (iv) Finally, the SP would provide a mechanistic, therefore theory-independent, definition of 'markedness.' No other theory-independent definition of 'markedness' exists. (According to the SP, the initial hypothesis (H) determined by the SP is 'unmarked'; that alternative H requiring experience for its confirmation is 'marked.') 'Markedness' would therefore mechanistically be defined in terms of amount of experience required for learnability.

In this paper we first (1) show that many of the predictions for English first language acquisition made by certain interpretations of the SP are disconfirmed by empirical data from studies of first language acquisition. We concentrate in particular on predictions made for acquisition of anaphora, given the centrality of anaphora to UG and to language knowledge (e.g., Chomsky, 1981, 1982).[2] (2) We then suggest that these ostensible failures to confirm the psychological reality of the SP arise at least in part from an ambiguity in the interpretation of this principle. We also suggest that where it appears that certain acquisition data confirm an extensional interpretation of the SP, an intensional interpretation in terms of UG is also possible. (3) We also suggest that recent attempts to apply the SP to cross-linguistic predictions for acquisition of anaphora, although not yet fully tested experimentally, can be problematic with regard to the linguistic analyses of anaphora in these languages and with regard to the consequent learnability problem they pose the child. Whether or not these problems exist also depends on how the SP is interpreted.

Specifically, in many cases the SP is interpreted in terms of <u>language</u> (L) (<u>viz</u>., in terms of the size or 'broadness' of language domains consistent with a certain H, e.g., number of sentences or sentence types which the H allows). Thus if the learner has a choice between Hi and Hj, and Hi leads to more types of sentences than Hj, and the set of sentence types allowed by Hj is included in the set of sentence types allowed by Hi, then Hj will be predicted to be the learner's initial 'unmarked' H, according to this interpretation of the SP. On the other hand, some predictions based on the SP are framed in terms of restrictiveness of the <u>grammar</u> (G) involved in each H, rather than in terms of the size of <u>language</u> domains which are the

extensional results of this grammar. In this case, if the learner has a choice between Hi and Hj, and Hi is a more 'restrictive' version of a grammar or a grammatical component (e.g., rule or principle), then Hi is said to be predicted to be the learner's initial H. This prediction is independent of size of the language domain (sets of sentence types) which follow from each H. It requires only that the G of H_i be included in the G of H_j.

We will argue that the principled Grammar/Language (G/L) distinction is critical to the formulation of the SP; and that these two formulations of the SP are not necessarily equivalent as they have been assumed to be. We will argue that the empirical predictions made by application of the SP to language acquisition are confirmed to a greater degree if the SP is interpreted in terms of G, not L. We will suggest that where data appear to support interpretation of the SP in terms of L, an alternative explanation in terms of a grammatical constraint on H is available. (4) We suggest that this restriction on interpretation of the SP in terms of grammar is in accord with its formal nature (Angluin, 1980); and with formal relations between 'language' and 'grammar'. (5) Finally, we suggest that this interpretation of the SP in terms of G, not L accords more closely with the theory of UG as a model of the initial state (e.g., Chomsky, 1981, Lasnik, 1981). In this model of UG, it is grammars which are proposed to be mentally represented, not languages. It is "constraints on the form of grammars which are of "central importance since they narrow the class of H that the child must consider" so that "...even where the cardinality of the class is not reduced (i.e., when the class remains infinite), the density of grammars compatible with fixed data will in general be lowered, thus facilitating acquisition" (Lasnik, 1981, 10).

In conclusion, although it has been claimed that the SP is "necessary and sufficient" to explaining learnability in a theory of UG, the facts reviewed here suggest that a specific linguistic theory of UG is necessary to the definition of the SP and to each instantiation of it. This conclusion raises the issue of how the SP relates to the PMF (Principle of Minimal Falsifiability) proposed by Williams 1981. This PMF defines 'most restrictive' and 'most unmarked' H as that which involves principles of Grammar "which require the least amount of evidence to acquire," but does not involve inclusion relations between alternative hypotheses and thus does not necessitate the SP.

THE GRAMMAR-LANGUAGE DISTINCTION

We assume here the formal distinction between 'language' and 'grammar': "Grammars (G) are formal systems". (Levelt, 1974, 1)."The LANGUAGE L(G) generated by G is the set of sentences generated by G" (1974, 5). Chomsky's theory of UG explicitly refers to this distinction:

> I have argued that the grammar represented in the mind is a "real object," indeed that a person's language should be defined in terms of this grammar, and that the vague everyday notion of language, if one wants to try to reconstruct it for some purpose, should be explained in terms of the real systems represented in the minds of individuals and similarities among these (Chomsky, 1980, 120)

We refer to statements that refer to languages, e.g. to sets of sentences or sentence types (e.g., by the size of these sets), as extensional in nature. We refer to statements that refer to grammars, or components of grammars, (e.g. to rules or principles or constraints), as intensional.

In general, the formal distinction between G and L is clear. It is well known from studies of the "Chomsky Hierarchy of Grammars" for example, (e.g., Levelt 1974), that a particular set of sentences (a language in the formal sense) may not map unambiguously onto a particular type of grammar. For example, a particular set of sentences which may be generated by a highly restricted type 3 'regular grammar' may also be generated by less restricted higher types of grammars (e.g., type 2 context free (CF) or type 0 unrestricted rewrite systems). Inclusion relations between such individual languages do not necessarily correspond to significant constrasts between grammars on this hierarchy. For example, language 1 may include the sentence types 'bbbbbb' and 'ababab', while language 2 includes only the sentence types 'ababab'. Both languages however are generable by highly restrictive regular grammars. The L-inclusion relation need not correspond to a significant G-inclusion relation, and therefore has no significance in itself. It could be argued in this case that some grammatical distinction other than that described by the Chomsky hierarchy does characterize the difference between these languages, and does establish a hierarchy between them, but such a grammatical theory would have to be developed to make the L-inclusion relation significant.

Language types on the Chomsky hierarchy (i.e., regular, context free, and context sensitive languages) do "show the same relations of strict inclusion as the grammar types" (Levelt, 1974,11); but only in so much as each language type refers to a 'class of languages'. Thus there are context free (CF) languages within the class of possible CF languages which are not regular, etc. Here 'class of languages' must be defined in terms of 'type of grammar'.

It follows thus that interpretations of the SP in terms of H's which bear an inclusion relation to each other need not be equivalent when formulated extensionally in terms of language domain, and when interpreted intensionally in terms of restrictiveness of Grammar.

Although the SP (as stated in 1 above) involves a "telltale set of sentences," the principle itself involves using these sentences to constrain inductive 'guesses' of possible 'classes of language;' where specification of 'class of L' requires determination by a Grammar. Thus the Hypotheses tested involve G, not L within the initial formulation of the SP.

Consequences of this G/L contrast for interpretation of the SP can be seen simply by the following example. The set of sentences with bare verbs (V) may be viewed as included in the set of sentences with verbs including verb phrases (VP) as well as bare verbs. (We will assume for the purpose of this argument that verbs can occur either transitively or intransitively, as in the English 'he ate' or 'he ate a meal'). On an extensional interpretation of the SP we could reason that the the SP would predict that the child's initial H was that sentences have bare verbs. Sentences with

bare V's therefore were more 'unmarked' than those with VP's. On an intensional interpretation of the SP, however, a more restrictive H grammatically (given certain theoretical assumptions) might reason that sentences with VP's represent a more restrictive H, because sentences with bare V's may involve deletions or empty categories. They would thus involve an expanded (less restricted) set of grammatical rules relative to full VP's. (Consider for example, the bare V's in "Who do you like e?" or in "Tom bought e and Sam ate the nougats".) Thus interpretations of the SP intensionally or extensionally may lead to distinct predictions relevant to the same data set.

Below we consider several proposals made in terms of either extensional or intensional interpretations of the SP, and evaluate available data from studies of first language acquisition which are relevant to these predictions. These data lead us to conclude that while extensional interpretations of the SP do not fare well, intensional interpretations are more often consistent with the initial acquisition data available. This conclusion has consequences for the nature of the SP and its integration with UG.

PREDICTIONS MADE BY THE SP FOR ACQUISITION OF ENGLISH WHEN FRAMED EXTENSIONALLY IN TERMS OF LANGUAGE.

Acquisition of Anaphora Types. In the current theory of Universal Grammar (UG) (e.g., Chomsky 1981, 1982), one module of UG, termed "Binding Theory" (BT) is specifically concerned with relations between nominal antecedents and their proforms, e.g. pronouns like 'he'/'him' or reflexives like 'himself' in English. A typology of such nominal categories distinguishes 'anaphors' as those elements like 'reflexives' which require an antecedent in the sentence, and have no independent reference, from those non-reflexive 'pronouns' which may be used deictically and do not require a grammatical antecedent. It is now known that 'anaphors' (e.g. English reflexives such as 'himself' in the examples 2, 4 and 6) and 'pronouns' (e.g. the lexical pronoun 'him' in examples 3, 5 and 7) must be differentiated in a theory of UG. The fact that 'anaphors' and 'pronouns' occur in different domains, as examples 2-7 show, suggests that they obey different principles. Principles of the 'Binding Theory" module of UG have been formulated accordingly to apply either to anaphors or to pronouns. By Principle A of BT, for example, "anaphors must be bound in their GC". By Principle B of BT, "pronouns must be free in their GC". ("GC" here refers to "Governing Category," one specification of a local domain).[3]

(2) $\underline{\text{John}}_i$ admires $\underline{\text{himself}}_i$.
(3) $\underline{\text{John}}_i$ admires $\underline{\text{him}}^*_{i,j}$.
(4) $\underline{\text{John}}_i$ wants $\underline{\text{Tom}}_j$ to admire $\underline{\text{himself}}_{*i,j}$.
(5) $\underline{\text{John}}_i$ wants $\underline{\text{Tom}}_j$ to admire $\underline{\text{him}}_{i,*j}$.
(6) $\underline{\text{John}}_i$ wants $\underline{\text{Tom}}_j$ to make $\underline{\text{Sam}}_k$ adore $\underline{\text{himself}}_{*i,*j,k}$.
(7) $\underline{\text{John}}_i$ wants $\underline{\text{Tom}}_j$ to make $\underline{\text{Sam}}_k$ adore $\underline{\text{him}}_{i,j,*k}$.

With regard to acquisition of anaphora types, , Berwick (1982, 300) considers that the initial H regarding anaphors and pronouns in acquisition of English may be constrained by the SP according to the following reasoning: "One can see . . . that the set of surface structures (now interpreted in an extended sense to include co-indexing), where pronominals can appear is larger than the set of surface structures in which some anaphors appear. One could therefore invoke the SP and establish an order in which H about the properties of an unknown NP element will be made." The initial H will be

that they are anaphors, not pronouns. Berwick notes that this interpretation of the SP "is not forced...since in fact pronominals and pure anaphors are in nearly complementary distribution"(1982,301) and therefore the domains at issue are not in an inclusion relation as application of the SP requires, according to 1. (cf. Manzini and Wexler, in preparation, on this issue). This formulation, however, is based on an extensional interpretation of the SP, since it refers to size of language domain allowed by each H being compared, (where 'set of surface structures' defines sentence types). It does not refer to Grammar, since the grammatical principles involved for 'anaphors' and 'pronouns' (Principles A and B) are complementary. These do not involve an inclusion relation between them, and neither of them is 'more restrictive' than the other.

The prediction, based on an extensional interpretation of the SP, that the initial H of children acquiring language would be that nominal pro-forms are anaphors, not pronouns, has been disconfirmed empirically, and theoretically on the basis of empirical first language acquisition studies (Wexler and Chien, 1985, Manzini and Wexler, in preparation, Padilla-Rivera, 1985; see Lust 1986 for review). Whether or not a child initially interprets an anaphor more like a pronoun or a pronoun more like an anaphor appears to be dependent on the context in which it appears.

Optional and Obligatory Anaphora. Another prediction of the SP interpreted extensionally was that the initial H of children would be that anaphora would be 'obligatory,' not 'optional.' This prediction can be interpreted in terms of the observation that "optional" language domains are extensionally larger than "obligatory" ones, and the sentence types resulting from the obligatory rule are included in the set of sentence types resulting from the optional rule. The grammatical principles of 'optionality' or 'obigatoriness' are not in themselves in an inclusion relation, but complementary. Therefore, the SP may not apply intensionally with regard to this domain of language knowledge.

This extensional prediction too has been disconfirmed empirically in certain situations. For example, Lust, Solan, Flynn, Cross and Scheutz, 1986 have found that young children assign an optional anaphora interpretation to sentences like 8a. They often convert the null sites in 8a (which are obligatorily subject-controlled in English adult grammar) to pronouns with verb tensing (in imitation) as in 8b; and they determine the interpretation of both 8a and b with sensitivity to pragmatic context, i.e., as though they were optional like the pronoun, in comprehension.

8a) $John_i$ saw Tom_j when $O_{i,*j}$ running down the street
b) $John_i$ saw Tom_j when $he_{i,j,k}$ was running down the street

Roeper 1986 reports another set of data which also appears to disconfirm this prediction. Roeper reports (from a preliminary experimental study) that children's initial H when interpreting the null subject in a sentence like 9a is similar to that in 9b. That is, children interpret the null subject in 9a as though it were a free lexical pronoun (with optional coreference thus), not as though it were a bound variable as it is in adult grammar. (For example, children attribute coreference between the person who "thinks" and the person who "wears a hat" in 9a.)

9a) Who does he think e wears a hat?
b) Who thinks he wears a hat?

PREDICTIONS MADE BY THE SP FOR ACQUISITION WHEN FRAMED INTENSIONALLY IN TERMS OF GRAMMAR.

Berwick 1982 also proposes that the SP may be interpreted to predict that children's initial H regarding empty categories is that they must be 'governed'. This hypothesis is defined independently of size of language domain. In fact, size of language domain resulting from this hypothesis is irrelevant, since it is either false or indeterminable that there are more sentences or sentence types with ungoverned than with governed empty categories. This H is 'intensional' in that it explicitly reflects a property of grammar, viz., the ECP (Empty Category Principle, which states that 'an empty category must be governed' in Universal Grammar (cf. Chomsky, 1981.)

This prediction, which involves an intensional interpretation of the SP, is consistent with certain empirical acquisition data. In Lust, Solan, Flynn, Cross and Schuetz, (1986), cited above in 8a for example, the child's main error was to tense the verb, thus governing the empty category subject, and allowing it to receive case. Thus it can and does appear as a lexical pronoun with case as in 8b. It is not clear how the SP actually applies to this H intensionally, however, since the grammar which includes the ECP is not included in a grammar which does not.

In cases, where data appear to support an extensional interpretation of the SP, an alternative (intensional) interpretation is possible. For example, in putative confirmation of the prediction of the SP as interpreted by Jakubowicz, and described above, it was reported that children make more errors on sentences like 10b and 11b than on 10a and 11a (cf. Deutsch and Koster, 1982, Wexler and Chien, 1985, Padilla-Rivera, 1985, Jakubowicz, 1984, and Lust, 1986 for relevant data here). Children reportedly often take the closest NP subject as antecedent for the pronoun in 11b, for example.

10a) $\underline{\text{John}}_i$ washed $\underline{\text{himself}}_i$
b) $\underline{\text{John}}_i$ washed $\underline{\text{him}}^*_{i,j}$
11a) John_i said that Peter_j washed $\underline{\text{himself}}^*_{i,j}$
b) John_i said that Peter_j washed $\underline{\text{him}}_{i,*j}$

An alternative explanation of these acquisition data (which are themselves now being tested for replication) might be the following. Children's initial H determined by UG regarding a nominal pro-form is that such proforms are bound if they are c-commanded. The child's initial H is that the minimal (possibly cyclic) constituent is assumed to be the relevant binding domain.

Cross-linguistic work on acquisition of anaphora in Chinese and English by Lust, Mangione and Chien (1984, and in preparation) is consistent with this alternative intensional hypothesis. Here children acquiring Chinese appear to use plus/or minus c-command to determine whether an empty category is bound or free. Padilla-Rivera's (1985) experimental test of acquisition of Spanish anaphora (which varies constituent domains of anaphors and pronouns) is also consistent with such an intensional representation of the knowledge underlying the acquisition facts. Here children acquiring Spanish

were found to assume that the minimal domain (e.g. prepositional phrase) in which a proform appears, is a binding domain for both anaphors and pronouns.

CROSS-LINGUISTIC APPLICATION OF THE SP

The above remarks concern interpretations of the SP with regard to choice between H which involve language possibilities within a single L. It is possible to interpret the SP in terms of cross-language variation. In this case, the H's involved would involve grammars for one language vs another. A-priori, this interpretation of the SP is more closely isomorphic to the structure of condition 1. However, in cross-linguistic application, interpretation of the SP can also be either extensional or intensional. Again, extensional predictions are problematic.

The concept 'parameter'(P) in UG specifies significant dimensions of possible cross-language variation, and the values these variations may take. Only a few of these P's involve a possible inclusion relation across their values, however, and thus allow application of the SP.

The Pro Drop Parameter (e.g., Chomsky, 1981, Hyams, 1983). Languages may take one of the binary values on this parameter or another. If they are [+ pro-drop] they will allow null subjects in tensed sentences optionally; if they are [- pro drop] like English, they will not. There is no grammatical basis for distinguishing one value of this parameter as included in another, or as intrinsically more marked than another. (See Lasnik 1983 for discussion of this point.) Each value of the parameter has a distinct set of grammatical consequences. Interpreted extensionally, however, in terms of the sizes of domains of language, or sentence types that each value of the parameter allows, then the [+ pro drop] is clearly more inclusive than the [- pro drop] hypothesis. Available acquisition data, however, do not support the prediction that children's initial H cross-linguistically is that the language they are acquiring is [- pro drop]. See Hyams 1983 and Mazuka et al to appear for alternative treatments of the acquisition facts.

Locality of binding domains for anaphors. Another aspect of language variation which has been predicted to be accessible to the SP involves 'locality' of anaphor binding. Japanese, unlike English, but like many Dravidian and other languages (e.g., Malayalam, Tamil, Telagu, Kannada, for example) allows 'long distance binding' of reflexives (cf. N. Akatsuka McCawley 1972, and K.S. Yadurajan 1986 for example). Compare English 4 and Japanese 12 below for example. It has been proposed that Japanese and English anaphora may be described in terms of a subset relation, wherein English is a subset of Japanese. Jakubowicz, 1984, 162, following Berwick's extensional proposal, interprets the SP as predicting that languages with long distance control of anaphors are "marked" by the SP relative to those languages without long distance binding, because those without it 'contain only one type of output string (sentences where anaphors are locally bound)', whereas those with long distance control 'contain two types of output sentences". This prediction is based on an extensional interpretation of the SP.

(12) [[Zibun$_i$ ga nusunda] to Sarah$_j$ ni iwareta] to John$_i$ ga omotta]
self nom stole that dat was said that nom thought
(John thought that Sarah said that he/self stole Ø)

On the grounds of this observation, it has been suggested that the SP predicts (if interpreted extensionally) that children's initial H (in Japanese as well as English) is that anaphors are locally bound. Although this H has not yet been extensively tested experimentally, there are several problems with this extensional instantiation of the H. Intensional and extensional interpretations again make different predictions for application of the SP.

(i) On the basis of recent cross-linguistic analyses, it appears that 'local' anaphors and 'long distance' anaphors may be grammatically distinct. In some languages, e.g. Malayalam, different forms distinguish these, but in other languages (e.g. Chinese or Japanese) different grammatical facts distinguish these syntactically and semantically. (See Giorgi 1984, Harbert to appear, Bowers in preparation, Tang in preparation, and Yadurajan 1986 for independent proposals of this type across several different languages which allow long distance 'anaphors'. If these grammatical analyses are correct then grammatically the local anaphor is not in an inclusion relation with the 'long distance' anaphor, since these are qualitatively distinct. Thus only an extensional, but not an intensional, intepretation of the SP would predict that 'anaphors' are locally bound in children's initial H.

(ii) Wexler and Manzini(to appear) and Manzini and Wexler(in prep) observe that the SP makes different predictions for pronouns than it does for anaphors. On an extensional interpretation, Icelandic is less restrictive than English with regard to anaphors because it allows long distance binding. However, it is more restrictive with regard to pronouns. Harbert argues for example, that English and Icelandic share sentence types with pronouns like 13, but Icelandic does not allow sentence types with pronouns with a coreferential reading like the English 14 does.

13) $They_i$ help $them_j$

14) $They_i$ expect me to help $them_{i,j}$

On an extensional interpretation of the SP, then, children's H would be predicted to be more like Icelandic than English, because there are fewer such sentence types with pronouns in Icelandic than in English. Harbert (in prep) argues, however, that grammatically the facts in 13 and 14 result from the fact that English has a more restrictive grammar than Icelandic (i.e, there are more restrictions on what consistutes a binding domain). Harbert supports the suggestion that it is the Icelandic grammar which is more marked on the basis of its low frequency across languages and on the basis of language change data. If this analysis is correct, however, then again, an intensional and an extensional interpretation of the SP will make distinct, in fact, opposite predictions for acquisition.,

iii) Wexler and Manzini (to appear) and M&W(in prep) have argued that the parameter of locality must be multivalued, ranging from English and other languages with 'very' local binding to Icelandic and other languages with less restrictive binding, to Japanese and other languages with least restrictive binding. The scale here is in terms of extensional size of domain of anaphora. W&M make the interesting and highly significant argument, however, that intensionally, i.e., grammatically, the substantive universals which characterize Binding Domains across languages are themselves in an

inclusion relation which correlates with the extensional size of domain of anaphors across languages.

However, any interpretation of a non-binary dimension of language variation in this way e.g., '$L_i \subset L_j \subset L_n$' will always formally require a violation of the SP. That is, whenever a child's H is in accord with L_i when L_n must be acquired, there will be an 'intervening language' L_j, thus offending condition 1 (SP). The only way that a subset offense can be avoided is if the child is constrained to hypothesize that Lj (and every intermediary language successsively on the inclusion hierarchy) is the correct language, before it hypothesizes that Ln is.

Certain languages (e.g., Chinese, Huang, pc, Tang, 1985 and in preparation) allow binding of the reflexive dzji (a) in minimal local domains as in 15; (b) in long distance domains as in 16; and (c) critically, in either maximal or minimal, but not intermediary domains in cases like 17. (Examples are from Tang, 1985). A phenomenon similar to 17 reportedly occurs in several other languages as well(e.g., Icelandic, Harbert, pc.; cf. also Giorgi, 1984). If the SP is interpreted extensionally, and the child is forced to formulate successive H up the hierarchy of locality, then a child acquiring Chinese will never acquire both 15 and 16 without making an error on 17. (If the inclusion hierarchy is formulated grammatically, then the grammar for a language which allows the intermediate clause antecedent in 17 must not be included in the grammar which allows the farthest antecedent in 16, in order for there not to be a required SP offense in this interpretation.)

15) Lisz_i dwei dzji_i mei syinsyin
no confidence
(Lisz has no confidence in himself) (=Tang #1)

16) Jangsan_i jrdau [Lisz_j tauyan $\text{dzji}_{i,\ j}$]
know dislike
(Jangsan_i knew that Lisz_j disliked $\text{himself}_{i,\ j}$)

17) Jangsan_i renwei [Lisz_j jrdau [Wangwu_k dwei $\text{dzji}_{i,*j,k}$
think know
mei syinsyin]]
no confidence (=Tang #83)
(Jangsan_i thought that Lisz_j knew that Wangwu_k had no confidence in $\text{himself}_{i,*\ j,\ k}$)

Note that if the grammatical parameter for distinction of binding domains across languages would be a binary one (as in Harbert, to appear, for example or as in a conflation of the W & M hierarchy), then the above issue need not arise.

CONCLUSIONS

These arguments do not motivate the conclusion that the SP is invalid. They do motivate the conclusion that in order for the Subset Principle to apply it is necessary not only that two languages (or two domains of language data) are in an inclusion relation, but also that the grammars (or some aspects of those grammars) which generate the language are in an inclusion relation. These arguments also motivate the conclusion that the interpretation of the SP must be restricted to close integration with theory of UG, if it is to be efficacious in a psychologically real model of the Initial

State. The current work of Wexler et al. (to appear) and Manzini and Wexler (in preparation) and Berwick, 1985 do represent advances in this regard. Wexler et al. and Manzini and Wexler, for example, pursue the linguistic/grammatical nature of what constitutes variation in Binding Domains across languages, i.e., of what the proper linguistic formulation of the Parameter or Parameters of language variation in this area are. The results above suggest that this contribution to UG is critical to specifying the dimensions of language variation which children may consult in formulating hypotheses, and the values these H may take. Empirically valid and successful application of the SP to children's test of these H in first language acquisition depends on the success of this linguistic theory of UG.

FOOTNOTES

1 I thank Jim Gair, and Wayne Harbert, John Bowers, Reiko Mazuka, Shin Oshima, and Kazuyo Otani for discussion of issues raised in this paper. I thank Reiko Mazuka, Shin Oshima , Kazuyo Otani and Wendy Snyder for Japanese data, and Jim Huang for Chinese data. I am thankful to Reiko Mazuka for critical assistance in paper preparation, with the assistance of Nancy Goss.

2 We will use the term "anaphora" in a general sense to refer to the relation between nominal antecedents and proforms (whether phonetically realized or null) as in examples 8a and b. This relation of anaphora determines the interpretation of the proform (e.g., either the pronoun 'he' or the null subject in examples 8) by reference to the antecedent. The relation of anaphora is often represented by co-indexing (as in the subscripts in 8), signifying possible coreference. (See various papers collected in Lust 1986 for study of this issue.)

3 In general, the reflexive 'himself' in sentence 2 reflects the fact that it is 'bound' in that it is 'c-commanded' by its antecedent, and obligatorily coreferential with it. ('C-command' is a concept defined over the geometry of tree structure; it involves dominance.) The pronoun 'him' in sentence 3 reflects the fact that it is 'free' in that it is not 'c-commanded' by its antecedent, may not refer to the main clause subject which c-commands it, and may refer to an antecedent outside the clause.

REFERENCES

Angluin, D. (1980). Inductive inference of formal languages from positive data. Information and Control, 48, 117-135.

Berwick, R. (1982). Locality principles and the acquistion of syntactic knowledge. (Vols. 1-2). Unpublished doctoral dissertation, MIT.

Berwick, R. (1985). The acquisition of syntactic knowledge. Cambridge: MIT.

Bowers, S. (In preparation). Non-anaphoric reflexives. Cornell University.

Brown, R., & Hanlon, C. (1970). Derivational complexity and order of acquistion in child speech. In J. R. Hayes (Ed.), Cognition and the development of language (pp. 11-53). New York: John Wiley & Sons.

Chomsky, N. (1980). Rules and representations. NY: Columbia University Press.

Chomsky, N. (1981). Lectures on government and binding. Dordrecht: Foris.

Deutsch, W., & Koster, J. (1982). Children's interpretation of sentence internal anaphora. PRCLD, 21, 39-45.

Giorgi, A. (1984). Toward a theory of long distance anaphors: A GB approach. The Linguistic Review, 3.

Gold, E. M. (1967). Language identification from given data. Information and Control, 10, 447-474.

Harbert, W. (in press). Binding, Subject and accessibility. In R. Friedin (Ed.), Proceedings of the Princeton Workshop on Comparative Grammars.

Hyams, N. M. (1983). The acquisition of parameterized grammars. Unpublished doctoral dissertation, The City University of New York, New York.

Jakubowicz, C. (1984). On markedness and binding principles. Proceedings of the North Eastern Linguistic Society, 14, 154-182.

Lasnik, H. (1981). Learnability, restrictiveness, and the evaluation metric. In C. L. Baker & J. J. McCarthy (Eds.), The logical problem of language acquisition (pp. 1-21). Cambridge: MIT Press.

Lasnik, H. (1983). On certain substitutes for negative data. Unpublished manscript, University of Connecticut, Storrs.

Levelt, W. J. M. (1974) Formal grammars in linguistics and psycholinguistics. (Vol. 1). The Hague: Mouton.

Lust, B. (1986). Introduction. In B. Lust (Ed.), Studies in the acquisition of anaphora: Vol. 1. Defining the constraints. Dordrecht: Reidel.

Lust, B., Mangione, L., & Chien, Y-C. (1984). Determination of empty categories in first language acquisition of Mandarin Chinese. Cornell University Working Papers in Linguistics, 6, 151-165.

Lust, B., Solan, L., Flynn, S., Cross, C., & Schuetz, E. (1986). A comparison of null and pronoun anaphora in first language acquisition. In B. Lust (Ed.), Studies in the acquisition of anaphora: Vol. 1. Defining the constraints (pp. 245-277). Dordrecht: Reidel.

Mazuka, R., Lust, B., Wakayama, T., & Snyder, W. (in press). Distinguishing effects of parameters in early syntax acquisition: A cross-linguistic study of Japanese and English. PRCLD.

Manzini, R., & Wexler, K. (in preparation). Parameters and learnability in binding theory. Unpublished manuscript, U.of California, Irvine.

McCawley, N. A. (1972). A study of Japanese reflexivization. Unpublished doctoral dissertation, University of Illinois, Chicago.

Oshima, S. (1985). Anaphora: A GB approach. Unpublished manuscript.

Padilla Rivera, J. (1985). On the definition of binding domains in the first language acquisition of anaphora in Spanish: The modular role of binding principles and lexical factors. Unpublished doctoral dissertation, Cornell University, Ithaca, NY.

Roeper, T. (1986). How children acquire bound variables. In B. Lust (Ed.), Studies in the acquisition of anaphora: Vol. 1. Defining the constraints (pp. 191-200). Dordrecht: Reidel.

Tang, C. C. J. (1985). A study of reflexives in Chinese. Unpublished master's thesis, Graduate Institute of English National Taiwan Normal University.

Wexler, K., & Chien, Y-C. (1985). The development of lexical anaphors and pronouns. Papers and Reports on Child Language Development, 24, 138-149.

Wexler, K., & Manzini, R. (in press). Parameters and learnability in binding theory. In Roeper & Williams (Eds.), Parameters & linguistic theory.

Williams, E. S. (1981). Language acquisition, markedness, and phrase structure. In S. Tavakolian (Ed.), Language acquisition and linguistic theory. (8-34). Cambridge, MIT Press.

Yadurajan, K. S. (1986). Reflexives in Dravidian. Paper presented at SALA, Urbana, IL. Unpublished manuscript, CIEFL, Hyderabad, India.

A THEORY OF DISCOURSE STRUCTURE

Barbara J. Grosz
Artificial Intelligence Center, and
Center for the Study of Language and Information
SRI International

ABSTRACT

1. Overview

In this talk I will present the basic elements of a computational theory of discourse structure.[1] A proper account of discourse structure is needed both as the basis of an account of discourse meaning (a semantic task) and to underlie a model of discourse processing. It provides the former by specifying the basic units a discourse comprises and the ways in which they can relate. It plays a key role in discourse processing by stipulating constraints on those portions of a discourse to which any given utterance in the discourse must be related.

An account of discourse structure is closely related to two questions: What individuates a discourse? What makes a discourse coherent? That is, faced with a sequence of utterances, how does one know whether they constitute a single discourse, several (perhaps interleaved) discourses, or none? Likewise, how does a discourse participant know where and how an utterance belongs in a discourse.

In the theory, discourse structure is intimately connected with two nonlinguistic notions: intention and attention. Intentions play a primary role in explaining discourse structure, defining discourse coherence, and providing a coherent conceptualization of the term "discourse" itself. Attention is an essential factor in explicating the processing of utterances in discourse.

A major claim of the theory that Sidner and I are developing [Grosz and Sidner 86] is

---

[1]Substantial portions of this abstract are taken from a paper coauthored by C.L. Sidner [Grosz and Sidner 86]. The theory I describe is being developed jointly with her. More details may be found in the paper.

that the structure of any discourse is a composite of three distinct but interacting components: (1) the structure of the actual sequence of utterances in the discourse; (2) a structure of intentions; (3) an attentional state. In examining attentional state, I will show that it has at least two separate constituents: one corresponds to the *global* focus of attention of the participants in a discourse and the other to their more local *centering* of attention at any given utterance in the discourse.

The distinction among the three components is essential to an explanation of various linguistic phenomena, including interruptions, the differential use of certain types of referring expressions, and the use of certain phrases that at times function solely to affect discourse segmentation and structure. Examples of each of these phenomena and the role of the components in explaining them will be given in the talk.

By providing an overall framework within which to answer questions about the relevance of various segments of discourse to one another and to the overall purposes of the discourse participants, the theory has implications for natural-language processing work in general. In particular, various properties of the intentional component suggest problems with approaches to discourse coherence based on selecting discourse relationships from a fixed set of alternative rhetorical patterns (e.g., [Hobbs 79], [Mann 83], [Reichman 81]) as well as indicating several issues that must be confronted in expanding speech-act-related theories (e.g., [Allen 80], [Cohen 80], [Allen 83]) from coverage of individual utterances to coverage of extended sequences of utterances in discourse.

Most of this presentation will be concerned with specifying an abstract model of discourse structure; in particular, the definitions of the components will abstract away from the details of the discourse participants. Both the construction of a computer system that can participate in a discourse (i.e., one that **is** a language user) and the specification of a psychological model of language use require the appropriate projection of this abstract model onto properties of a language user and specification of additional details; for example, memory for linguistic structure must be specified and means for encoding attentional state, and appropriate representations of intentional structure must

be provided). Although I will not address such issues completely, I will examine certain key processing questions.

2. Linguistic Structure

Utterances--the actual saying or writing of particular sequences of phrases and clauses--are the basic elements of the linguistic structure. Just as the words in a single sentence form constituent phrases, the utterances in a discourse are naturally aggregated into *discourse segments*. The linguistic structure consists of these discourse segments and embedding relationships among them.

The utterances in a segment, like the words in a phrase, serve particular roles with respect to that segment. In addition, the discourse segments, like the phrases, fulfill certain functions with respect to the overall discourse. Although two consecutive utterances may be in the same discourse segment, it is also common for two consecutive utterances to be in different segments. It is also possible for two utterances that are nonconsecutive to be in the same segment. An individual segment may include a combination of subsegments and utterances only in that segment (and not members of any of its embedded subsegments).

One claim of the theory is that the embedding relationships are a surface reflection of relationships among elements of the intentional structure. However, various elements of the linguistic structure are crucial in conveying information about the intentional structure so that these two components are mutually constraining. Because the linguistic structure is not strictly decompositional, various properties of the discourse (most notably the intentional structure) are functions of properties of individual utterances and properties of segments. In addition to giving examples that illustrate the nonstrict decompositionality of discourse, I will discuss problems that arise in determining the intentional structure because of it.

There is a two-way interaction between the discourse segment structure and the utterances constituting the discourse: linguistic expressions can be used to convey information about the discourse structure; conversely, the discourse structure constrains

the interpretation of expressions (and hence affects what a speaker says and how a hearer will interpret what is said). Not surprisingly, linguistic expressions are among the primary indicators of discourse segment boundaries. The explicit use of certain words and phrases (e.g., "in the first place") and more subtle clues, such as changes in tense and aspect, are included in the repertoire of linguistic devices that function, wholly or in part, to indicate these boundaries [Reichman-Adar 84, Cohen 83, Polanyi 83]. I will give examples that show these linguistic boundary markers can be divided according to whether they explicitly indicate changes in the intentional structure or in the attentional state of the discourse. The differential use of these linguistic markers provides one piece of evidence for considering these two components to be distinct.

Because these linguistic devices function explicitly as indicators of discourse structure, it becomes clear that they are best seen as providing information at the discourse level, and not at that of the sentence; hence, certain kinds of questions (e.g., about their contribution to the truth conditions of an individual sentence) do not make sense. For example, in the utterance "Incidentally, Jane swims every day," the "incidentally" indicates an interruption of the main flow of discourse rather than affecting in anyway the meaning of "Jane swims every day." Jane's swimming every day could hardly be fortuitous.

Just as linguistic devices affect structure, so the discourse segmentation affects the interpretation of linguistic expressions in a discourse. Referring expressions provide the primary example of this effect. The segmentation of discourse constrains the use of referring expressions by delineating certain points at which there is a significant change in what entities (objects, properties, or relations) are being discussed. For example, there are different constraints on the use of pronouns and reduced definite-noun phrases within a segment than across segment boundaries. While discourse segmentation is obviously not the only factor governing the use of referring expressions, it is an important one.

3. Intentional Structure

A rather straightforward property of discourses, namely, that they (or, more accurately, those who participate in them) have an overall purpose, turns out to play a fundamental role in the theory of discourse structure. Thus, intentions of a particular sort and a small number of relationships between them provide the basic elements of the intentional structure.

Although typically the participants in a discourse may have more than one aim in participating in the discourse (e.g., a story may entertain its listeners as well as describe an event; an argument may establish a person's brilliance as well as convince someone that a claim or allegation is true), we distinguish one of these purposes as foundational to the discourse. We will refer to it as the *discourse purpose (DP)*. From an intuitive perspective, the discourse purpose is the intention that underlies engaging in the particular discourse. This intention provides both the reason a discourse (a linguistic act), rather than some other action, is being performed and the reason the particular content of this discourse is being conveyed rather than some other information. For each of the discourse segments, we can also single out one intention--the *discourse segment purpose (DSP)*. From an intuitive standpoint, the DSP specifies how this segment contributes to achieving the overall discourse purpose.

Typically the initiator of a discourse (e.g., the writer of a text or the first speaker in a dialogue) will have a number of different kinds of intentions that lead to initiating a discourse. One kind might include intentions to speak in a certain language or to utter certain words. Another might include intentions to amuse or to impress. The kinds of intentions that can serve as discourse purposes or discourse segment purposes are distinguished from other intentions by the fact that they are intended to be recognized (cf. [Allen 80], [Sidner 85]), whereas other intentions are private; that is, the recognition of the DP or DSP is *essential* to its achieving its intended effect. Discourse purposes and discourse segment purposes share this property with certain utterance-level intentions that Grice [Grice 69].

It is important to distinguish intentions that are intended to be recognized from other

kinds of intentions that are associated with discourse. Intentions that are intended to be recognized achieve their intended effect only if the intention is recognized. For example, a compliment achieves its intended effect only if the intention to compliment is recognized; in contrast, a scream of "boo" typically achieves its intended effect (scaring the hearer) without the hearer having to recognize the speaker's intention.

We have identified two structural relations that play an important role in discourse structure: *dominance* and *satisfaction-precedence.* An action that satisfies one intention, say DSP1, may be intended to provide part of the satisfaction of another, say DSP2. When this is the case, we will say that DSP1 *contributes to* DSP2; conversely, we will say that DSP2 *dominates* DSP1 (or DSP2 *DOM* DSP1). The dominance relation invokes a partial ordering on DSPs that we will refer to as the *dominance hierarchy.* For some discourses, including task-oriented ones, the order in which the DSPs are satisfied may be significant, as well as being intended to be recognized. We will say that DSP1 *satisfaction-precedes* DSP2 (or, DSP1 *SP* DSP2) whenever DSP1 must be satisfied before DSP2.[2]

The range of intentions that can serve as discourse, or discourse segment, purposes is open-ended (cf. [Wittgenstein 53], paragraph 23), much like the range of intentions that underlie more general purposeful action. There is no finite list of discourse purposes, as there is, say, of syntactic categories. It remains an unresolved research question whether there is a finite description of the open-ended set of such intentions. However, even if there were finite descriptions, there would still be no finite list of intentions from which to choose. Thus, a theory of discourse structure cannot depend on choosing the DP/DSPs from a fixed list (cf. [Reichman-Adar 84], [Schank 82], [Mann 83]), nor on the particulars of individual intentions. Although the particulars of individual intentions, like a wide range of common sense knowledge, are crucial to understanding any discourse, such particulars cannot serve as the basis for *determining* discourse structure.

---

[2]These two relations are similar to ones that play a role in parsing at the sentence level: immediate dominance and linear precedence. However, the dominance relation, like the one in Marcus and Hindle's D-theory [Marcus et al. 83], is partial (i.e., nonimmediate).

What is essential for discourse structure is that such intentions bear certain kinds of structural relationships to one another. Since the CPs can never know the whole set of intentions that might serve as DP/DSPs, what they must recognize is the relevant structural relationships among intentions. Although there is an infinite number of intentions, there are only a small number of relations relevant to discourse structure that can hold between them.

4. Attentional State

The third component of discourse structure, the attentional state, is an abstraction of the participants' focus of attention as their discourse unfolds. Attentional state contains information about the objects, properties, relations, and discourse intentions that are most salient at any given point. It essentially summarizes information from previous utterances crucial for processing subsequent ones thus obviating the need for keeping a complete history of the discourse. It is inherently dynamic, recording the objects, properties, and relations that are salient at each point in the discourse. Thus attentional state serves an additional purpose: namely, it furnishes the means for actually using the information in the other two structures in generating and interpreting individual utterances.

In this theory, the global component of attentional state is modelled by a set of *focus spaces*; changes in attentional state are modelled by a set of transition rules that specify the conditions for adding and deleting spaces. We call the collection of focus spaces available at any one time the *focusing structure* and the process of manipulating spaces *focusing.*

The global focusing process associates a focus space with each discourse segment; this space contains those entities that are salient--either because they have been mentioned explicitly in the segment or because they became salient in the process of producing or comprehending the utterances in the segment. The focus space also includes the DSP; the inclusion of the purpose reflects the fact that the CPs are focused not only on what they are talking about, but also on why they are talking about it. Note however, that although each focus space contains a DSP, the focus structure does *not* include the

intentional structure as a whole.

The local component of attentional state is modelled by assigning to each utterance a *backward-looking center* and a set of *forward-looking centers* [Grosz et al. 83, Grosz et al. 86]. Centering rules provide constraints on how various entities can be realized in an utterance (e.g., what surface forms can be used to refer to a given object). I will give examples that illustrate differences in acceptability according to whether these rules are followed or not.

It is important to note that attentional state component is not equivalent to cognitive state, but is only one of its components. Cognitive state is a richer structure, one that includes at least the knowledge, beliefs, desires, and intentions of an agent, as well as the cognitive correlates of attentional state as modelled in this paper.

Acknowledgements

The research described in this abstract was made possible by a gift from the System Development Foundation.

References

[Allen 80] Allen, J.F., and Perrault, C.R.
Analyzing intention in dialogues.
Artificial Intelligence 15(3):143-178, 1980.

[Allen 83] Allen, J.F.
Recognizing Intentions from Natural Language Utterances.
In M. Brady and R.C. Berwick (editors), *Computational Models of Discourse*, pages 107-166. Massachusetts Institute Technology Press, 1983.

[Cohen 80] Cohen, P.R. and Levesque, H.L.
Speech Acts and the Recognition of Shared Plans.
In *Proc. of the Third Biennial Conference*, pages 263-271. Canadian Society for Computational Studies of Intelligence, Canadian Society for Computational Studies of Intelligence, Victoria, B. C., May, 1980.

[Cohen 83] Cohen, R.
A Computational Model for the Analysis of Arguments.
Technical Report CSRG-151, Computer Systems Research Group, University of Toronto, October, 1983.

[Grice 69] Grice, H.P.
Utterer's Meaning and Intentions.
Philosophical Review 68(2):147-177, 1969.

[Grosz and Sidner 86]
Grosz, B.J. and Sidner, C.L.
Attention, Intentions, and the Structure of Discourse.
Computational Linguistics 12, 1986.

[Grosz et al. 83] Grosz, B.J., Joshi, A.K., Weinstein, S.
Providing a Unified Account of Definite Noun Phrases in Discourse.
In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, June, 1983.

[Grosz et al. 86] Grosz, B.J., Joshi, A.K., and Weinstein, S.
Towards a Computational Theory of Discourse Interpretation.
, 1986.
in preparation.

[Hobbs 79] Hobbs, J.
Coherence and Co-references.
Cognitive Science 3(1):67-82, 1979.

[Mann 83] W.C. Mann and S.A. Thompson.
Relational Propositions in Discourse.
Technical Report RR-83-115, Information Sciences Institute, November, 1983.

[Marcus et al. 83]
Marcus, M. P., Hindel, D., and Fleck, M. M.
D-Theory: Talking about Talking about Trees.
In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*, pages 129-136. The Association for Computatinal Linguistics, Cambridge, MA, June, 1983.

[Polanyi 83] Polanyi, L. and Scha, R.
On the Recursive Structure of Discourse.
In K. Ehlich and H. van Riemsdijk (editors), *Connectedness in Sentence, Discourse and Text*, pages 141-178. Tilburg University, 1983.

[Reichman 81] R. Reichman.
Plain-speaking: A theory and grammar of spontaneous discourse.
PhD thesis, Department of Computer Science, Harvard University, 1981.
Also, BBN Report No. 4681, Bolt Beranek and Newman Inc., Cambridge, MA.

[Reichman-Adar 84]
Reichman-Adar, R.
Extended Person-Machine Interface.
Artificial Intelligence 22(2):157-218, March, 1984.

[Schank 82] Schank, R.C., Collins, G.C., Davis, E., Johnson, P.N., Lytinen, S., Reiser, B.J.
What's the Point?
Cognitive Science 6(3):255-275, July-September, 1982.

[Sidner 85] Sidner, C.L.
Plan parsing for intended response recognition in discourse.
Computational Intelligence 1(1):1-10, February, 1985.

[Wittgenstein 53] Wittgenstein,L.
Philosophical Investigations.
Oxford Press, 1953.

THE USE OF REMINDINGS IN PLANNING*

Kristian Hammond
Department of Computer Science
Yale University

Abstract

In recent years, much research has been aimed at the study of episodic reminding and its relationship to the process of understanding in both man and machines. Because of the ill defined nature of the functionality of episodic remindings in the understanding process, however, little progress has been made in uncovering the nature of the memory organization that supports these remindings or how they are used to help in understanding.

This paper will present a functional view of episodic reminding from the perspective of planning rather than understanding. Because of the clearer functionality of planning, a more straightforward view of the use of these remindings and the memory organization that supports them is possible. This paper will discuss three situations in which episodic memories can be used in planning: problem anticipation, plan construction and plan repair. The memory organization that each of these uses implies will also be discussed. This view of memory will be presented in terms of a case-based planner, CHEF, which makes extensive use of episodic remindings in constructing new plans.

1 Episodic Remindings in Understanding and Planning

Interest in the occurrence and use of episodic remindings abounds in both Artificial Intelligence and psychology. Studies of analogical reasoning, learning from experience and episodic memory organization all center on the notion of using features in one situation to access memories of a past one. In AI, the focus of interest has been on the structures that are required to simulate these remindings in machines. In psychology, the approach has been to experimentally clarify the features that are used to access them. Unfortunately, both of these approaches seem to ignore the most important aspect of these remindings: their function.

Without a clear idea what function episodic remindings serve, there are strong limits on what can be said about the memory organization that supports them and about the information that is available to any process that accesses them. Because of this, most of the discussion about remindings has been reduced to examining lists of the common features between an episode and the reminding it elicits. There has been little discussion about how functional needs determine the vocabulary that may be used to index episodes in memory or about the processes that chose the features in an input episode to use in accessing past episodes in memory. There is a great deal of discussion of remindings that ignores the constraints that would rise out of a theory of what purpose they serve.

One reason that many of these theories lack a functional view of remindings is that most of them come out of the context of natural language understanding in which the use of episodic

*This report describes work done in the Department of Computer Science at Yale University. It was supported in part by ONR Grant #N00014-85-K-0108.

remindings is somewhat ill-defined. One way around this problem is to shift the study of episodic reminding into an area where the task is clearer and the possible functionality of the remindings is more straightforward: planning. The goals and requirements of a planner are somewhat better defined than those of an understander. Likewise the situations in which episodic remindings would be of use to a planner are also better defined. In planning there is a clearer notion of when a reminding would be of use, what it would be used for and what information the planner might have at hand to find it. This clarity provides strong constraints on the vocabulary used to store episodic memories in the first place and on the choice of the features in an situation that are going to be used to search for a past episodes for use in that situation.

The discussion in this paper will center around the memory structures used in the computer program CHEF, a case-based planner in the domain of Szechwan cooking that makes heavy use of its past planning experiences to deal with current goals. The organization used by this planner provides memories for use in initial plan construction, problem anticipation and plan repair. These memory organizations provide what we call *directed remindings*, because they are intentionally sought by processes actively seeking solutions to present problems using past memories.

2 Remindings in Understanding

In 1980, Schank [Schank 82] proposed a view of remindings as the natural by-product of goal tracking while understanding stories. He suggested a set of memory structures that would account for the association of certain episodes in memory and allow access of one from the other. In his initial work, however, he did not suggest a model of how these remindings could be used in the understanding process. Following this proposal, Dyer [Dyer 82] designed an implementation of many of the organizational ideas suggested by Schank. His understander, BORIS, was "reminded" of memory structures similar to Schank's, but did not make any use of them in actually understanding stories. More recently Dyer [Dyer 85] has continued to look at computer simulations of understander's that have remindings but still has done no more in the way of explaining their use.

In Schank's current work [Schank 86], he has suggested using remindings for verification of explanations. He also has students working on reusing past explanations to fill in causal gaps in stories being read by an understanding program. Even in this work, however, there is some difficulty in defining the function of the remindings, which undercuts the effort to demonstrate how and when they arise.

In psychology, a great deal of work on analogical reminding has been done by Gentner [Gentner and Landers 85] and Holyoak [Gick and Holyoak 83] in trying to pin down the features used to access memory. They have both taken the approach of providing subjects with understanding and planning tasks that elicit remindings and noting the commonalities between the reminding and the initial episode. The object of this work has been to derive the features that are used to organize memory.

Unfortunately, both Gentner and Holyoak seem to ignore the fact that a reminding of one episode from another is a reminding out of all of memory. They forget that their task is not to just explain why one episode was found but also explain why others weren't. They are satisfied to explain a reminding from an episode to one in the past by simply listing the features that they

have in common. But the major problem with this work is that it avoids the fact that memory is used for something. Both Gentner and Holyoak seem to be looking at memory and reminding as though it were a phenomenon in a vacuum, without any function. Because of this, they have failed to ask the questions about the purpose to which the reminding they have studied are being put, and thus have failed to ask why and how they were retrieved in the first place.

A striking exception to the non-functional approach to the study of remindings in understanding is the work of Seifert [Seifert 85]. While her experimental paradigm is similar to Gentner's, her materials were aimed at discovering the processing situations that evoked remindings as well as the features used to index them. Her initial work seems to indicate that remindings are sought to fill in explanatory gaps in narratives and that the process that retrieves the remindings is more directed than it has previously been thought to be.

The general problem of discussing episodic remindings from the perspective of understanding is that the tasks involved in understanding are somewhat ill-defined. Because of this, the function of remindings in those tasks is equally ill-defined. Without an idea of the function of remindings, the theories that have arisen to explain them have lacked any notion of when the remindings will occur, what features are actually used to access them and what form the memories themselves will have to take in order to be useful. The similarities between episodes that have been noted within remindings have been interesting, but they only tell part of the story.

3 Remindings in Planning

The task of planning is somewhat more straightforward than the task of understanding. As a result, the specific functions to which an episodic memory can be put are somewhat more well defined than those in understanding. In planning, remindings of past situations can be used to suggest planning strategies, repair faulty plans, provide complete plans and warn of potential failures. These functions are such that they actually define what the planner needs and what it has on hand to search for a reminding to satisfy those needs. In planning, a theory of reminding is more than a list of the similarities between two situations. It is a theory that includes the vocabulary required to get the reminding, the process which chooses the features in one situation that are used to search for one and a description of the kind of information a planner is trying to get from it.

Carbonnel [Carbonell 83] has suggested one of the the more successful approaches to the use of remindings in planning and problem solving in his model of problem solving through analogy. His model centers around the idea that past solutions to problems can provide strategic information which can be applied to new problem solving situations. The vocabulary used to store past instances is based on a partial processing of problems. If the processing steps in one problem match those taken in a past one, the past steps are accesses from memory and applied to the current situation. The problem solver is always trying to find past solutions that can be applied to the present case. The information that the problem solver transfers from the old situation to the new consists of the the steps taken in the old case. While Carbonell's view of remindings gains a great deal from the function served by the memories retrieved, he suggests only one use that these memories serve.

4 Remindings in CHEF

CHEF is a case-based planner that tries to use past memories in all aspects of its planning activity. There are three areas in which this has been most successful:

- Plan Construction, which makes use of a memory organization of past plans indexed by the goals that they satisfy and the problems they avoid.
- Problem Anticipation, which makes use of a memory of the planner's own failures, indexed by the goal features that predict them.
- Plan Repair, which makes use of a memory of past repairs, indexed by descriptions of the problems that they solve.

Before looking at how CHEF makes use of remindings in planning, it is important to get an idea of how it plans in general.

CHEF's input is a set of goals for different tastes, textures, ingredients and types of dishes and its output is a single recipe that satisfies all of its goals. Its basic algorithm is to find a past plan that satisfies as many of the most important goals as possible and then modify that plan to satisfy the other goals as well.

Before searching for a plan to modify, CHEF examines the goals in its input and predicts any failures that might rise out the interactions between the plans for satisfying them. If a failure is predicted, CHEF adds a goal to avoid the failure it to its list of goals to satisfy and this new goal is also used to search for a plan. For example, if it predicts that stir frying chicken with snow peas will lead to soggy snow peas because the chicken will sweat meat into the pan, it searches for a stir fry plan that avoids the problem of vegetables getting soggy when cooked with meats. In doing so, it finds a past plan for beef and broccoli that solves this problem by stir frying the vegetable and meat separately. The important similarity between the current situation and the one for which the past plan was built is that the same problem rises out of the interaction between the planner's goals, although the goals themselves are different.

4.1 Plan Construction - Memory of Past Plans

CHEF's approach to plan construction is somewhat different than most planners'. Rather than build plans up from primitive steps, it searches for a past plan in memory that satisfies as many of its goals as possible and then modifies it to satisfy all of the planner's other goals. Because CHEF also has the ability to anticipate planning problems (see Section 4.2) before they arise, it tries to find past plans that avoid the problems that it anticipates while also satisfying its active goals.

The plan memory that is defined by this function is straightforward: past plans are indexed by the goals that they satisfy and the problems that they avoid. A CHEF plan for stir fried beef and broccoli, for example, is stored by the fact that it is a stir fry plan, that it includes beef and broccoli and by the fact that it avoids a problem endemic to stir frying in which the liquid produced by the stir fried meat makes the vegetable fried with it soggy.

CHEF begins its search of plan memory with two sets of features in hand, the goals that it needs to achieve and the problems it wants to avoid while doing so. These goals are ranked so that CHEF is able to search for plans that satisfy its most important goals first. This ranking is based on the difficulty with which it can add the steps for a goal to an existing plan.

Because it is rare that a past plan will satisfy all of the goals that CHEF has in hand, it also needs a memory organization that reflects similarity between different goals and the different plans that satisfy them. A plan to satisfy the goal to stir fry beef, for example is similar to the plan to stir fry pork. As a result, a complex plan for beef can easily be modified to satisfy the goal to include pork. This notion of similarity is implemented in an abstraction network of goals, in which similar goals are close together in the network.

CHEF's plan memory is implemented as a discrimination net, in which plans are indexed by the goals that they satisfy and the problems that they avoid. Access to this memory is controlled by CHEF's plan retriever, a module that takes a set of goals, including the goals to avoid problems, and retrieves the past plan in memory that satisfies as many of the most important goals as possible. The planner first uses its knowledge of the relative difficulty of modifying plans to prioritize the goals it is handed. It then uses the highest priority goal to drive into memory and abandons the search on the basis of that goal only after it has failed to find a plan that even partially satisfies it.

When a goal fails to match any of those satisfied by an existing plan, the plan retriever backs off on the goal and tries to find a plan that at least satisfies a more general version of it. If there are multiple plans that do this, the planner further discriminates between them on the basis of individual features of the goals that the plans satisfy. The features it chooses to do this discrimination are dynamically determined by past failures that are associated with them and by a static ordering of features that is used if none of the features is problematic. After a goal has been used to discriminate through the plan memory, the next highest priority goal is called on to do any further discrimination that has to be done.

The remindings that CHEF gets from its plan retriever reflect the needs of the planner. They are plans that satisfy many of the planner's goals because its plan memory is organized around those goals. Because it also wants to avoid problems, the fact that a plan does so is also used to index plans. Because it can use plans that partially satisfy goals it has a notion of similarity that is also used to organizes this memory. And because it wants to minimize the work it has to do to modify a past plan to satisfy its current goals, it searches for past plans indexing on the most difficult to achieve goals first.

In dealing with a planning problem of making a stir fry dish with chicken and snow peas, CHEF anticipates the problem of the liquid from the chicken ruining the texture of the snow peas and so adds a goal to avoid this problem to the list of goals it has to achieve. Once this is done, it searches for a plan that satisfies as many of the goals as possible. Because the most difficult goal to achieve is the one to avoid the problem of the liquid ruining the snow peas, this one becomes the highest priority goal for this search. As a result, a plan that satisfies this goal and partially satisfies the others is used rather than another plan that is a stir fry dish for chicken but does not have the changes that would allow the planner to easily avoid the problem it has anticipated.

```
Searching for plan -

Placing goals in order of difficulty -
   Make a stir-fry dish.
   Avoid failure of type SIDE-EFFECT:DISABLED-CONDITION:CONCURRENT
 exemplified by the failure 'The broccoli is now soggy' in recipe
```

```
BEEF-AND-BROCCOLI.
  Include chicken in the dish.
  Include snow pea in the dish.


Driving down on: Make a stir-fry dish.
 Succeeded -
Driving down on:
  Avoid failure of type SIDE-EFFECT:DISABLED-CONDITION:CONCURRENT
  exemplified by the failure 'The broccoli is now soggy' in recipe
  BEEF-AND-BROCCOLI.
 Succeeded -
Driving down on: Include chicken in the dish.
 Failed - Trying more general goal.
  Driving down on: Include meat in the dish.
   Succeeded -
Driving down on: Include snow pea in the dish.
 Failed - Trying more general goal.
  Driving down on: Include vegetable in the dish.
   Succeeded -

Found recipe -> REC9 BEEF-AND-BROCCOLI

Recipe exactly satisfies goals ->
    Avoid failure of type SIDE-EFFECT:DISABLED-CONDITION:CONCURRENT
exemplified by the failure 'The broccoli is now soggy' in recipe
BEEF-AND-BROCCOLI.
    Make a stir-fry dish.

Recipe partially matches ->
    Include chicken in the dish.
          in that the recipe satisfies: Include meat in the dish.

Recipe partially matches ->
    Include snow pea in the dish.
          in that the snow pea can be substituted for the broccoli.
```

In CHEF, remindings of past plans serve a function and that function defines the memory organization that houses the plans and the access process that retrieves them. As a result, the planner can use its memory to find the plans that fit its needs. These plans best achieve the goals that CHEF has been given and avoid the problems that it has anticipated. They partially satisfy goals can be found along with those that completely satisfy them. CHEF's plan memory is designed to meet the needs of the planner at large and the features that are used to access it are those that best describe the problems that the plans searched for have to solve.

4.2 Failure avoidance - memory of past failures

To avoid failures, CHEF has to be able to anticipate them, to notice that they are going to occur in order to give the plan retriever the goal to avoid them. The process that predicts the occurrence of failures is CHEF's problem anticipator which uses a knowledge base of CHEF's memory of failures.

The function of a problem anticipator is to predict failures on the basis of the surface features of a situation that have caused similar failures in the past. It predicts, for example, that the plan for the goals of including a meat and a crisp vegetable in a stir fry dish will lead to the vegetable becoming soggy after being stir fried with the meat. As the last section demonstrated, once this prediction has been made, a plan that avoids this problem can be found. But before the problem can be solved, it must be anticipated. And to anticipate these failures, CHEF's memory of them has to be organized so as to link its memory of past failures to the features of the goals that predict them.

The organization of CHEF's memory of failures is less complex than its organization of plans. It is simpler because its function and the requirements of the process that uses it are simpler. The task of the anticipator is to infer the possibility of a failure from a set of surface features. The memory of failures that it uses, then, has to be organized in a way that connects those features to the failures they predict.

CHEF's memory of failures is a simple network of nodes, in which particular failures are connected to the goal features that predict them. In CHEF, the surface features of a situation all stem from the goals that it is asked to plan for. The goals that CHEF is planning for define its situation, so it makes no sense to link memories of failures to anything but these goals. The failure in the beef and broccoli situation is linked to the goal to include meat, the goal to include any crisp vegetable and the goal to have a stir fry dish.

When a particular feature of a goal can be identified as participating in a failure, a test is built for that feature and is associated with the most specific version of the goal that allows the most general use. If all members of a class of items is associated with the failure, a link is made directly from that class to the memory of the failure itself.

When a new set of goals is handed to CHEF, its first step is to activate the features associated with each of the goals which in turn spreads activations to any memoriies of past failures that are associated with those features. If all of the features that are required to predict any failure send markers to its node, it activates itself and the planner builds a goal to avoid that failure and adds it to the planner's goal list. In looking at the goals to have a stir fry dish that includes chicken and snow peas, then, the fact that the chicken is a meat and the snow peas are a crisp vegetable activates the memory of a past stir fry failure stemming from these features.

```
Searching for plan that satisfies -
   Include chicken in the dish.
   Include snow pea in the dish.
   Make a stir-fry dish.

 Activating features:
  Activating: The dish STYLE-STIR-FRY.

  Activating: The item a MEAT.

  Activating: The item a VEGETABLE.
              The TEXTURE of item CRISP.

     Chicken + Snow Pea + Stir frying = Failure
```

```
"Meat sweats when it is stir-fried."
"Stir-frying in too much liquid makes vegetables soggy."
 Reminded of BEEF-AND-BROCCOLI.
 Fired demon: DEMON0

Based on features found in items: snow pea, chicken and stir fry
Adding goal: Avoid failure of type
SIDE-EFFECT:DISABLED-CONDITION:CONCURRENT exemplified by
the failure 'The broccoli is now soggy' in recipe
BEEF-AND-BROCCOLI.
```

CHEF's memory of failures is organized around the features of the goals and states that predict them. This lets CHEF go directly from a set of goals to be planned for to a memory of those failures that it should be trying to avoid. By also associating the failures with the actual features that caused them, features that may not have been directly attached to a goal that the planner was working on when the failure occurred, it is also able to be reminded of failures in situations that have surface features different from those of the original situation. So CHEF's memory of failures can provide appropriate predictions of problems in situations that are similar to those it has seen before at the level of initial goals and at the level of the actual causes of particular problems.

4.3 Plan modification - memory of past repairs

After CHEF is reminded of a past plan, it has to modify it to satisfy any goals that it does not already achieve. To do this CHEF uses a table of standard modifications that tell it how to add new goals to existing plans. In some case, however, CHEF needs to make changes that go beyond the scope of the modifications suggested. These are cases where past experience has shown that the standard modifications will lead to a failure.

Because CHEF is able to anticipate problems that it has encountered before, it is able to predict when these cases will arise. It is also able to use this prediction to find the past repair it used to solve the problem in the earlier instance. This is because it stores the memory of the repair it made in memory, indexed by the problem that it solves. Like complete plans that are indexed by the problems that they avoid, specific plan alterations are indexed by the problems that they fix.

One example of this occurs when CHEF is building a plan for Duck Dumplings. After running a basic plan that just replaces duck for pork in an existing recipe, it finds that the fat from the duck makes the recipe too greasy. It repairs this problem in its new plan by removing the fat from the duck before grinding it. It also stores the memory of this repair under the prediction of the failure. Later, when it is planning for a pasta dish with duck, it predicts that the same problem will occur but cannot find a pasta dish that avoids it. The best it can find is the recipe for Ants Climb a Tree, a pasta dish with pork, that it puts the duck into. Because it has predicted a problem with grease, however, it has access to the past repair that deals with it and can apply it before the failure occurs again.

```
Modifying recipe: DUCK-PASTA
```

```
to satisfy: Include duck in the dish.

Role substitution of duck for pork in recipe DUCK-PASTA.

Placing some duck in recipe DUCK-PASTA


Considering reminding:
After doing step: Bone the duck
 do: Clean the fat from the duck
 because: The duck is now fatty
 avoiding: SIDE-FEATURE:GOAL-VIOLATION
- Reminding applied.
```

By storing the memory of the past repair in terms of the problem that it solves, the planner can be reminded of it when the problem arises again and use it to solve the problem before it leads to an actual planning failure.

5 Expanding the Scope of Remindings

While CHEF makes good use of *directed remindings* there are still many other ways in which remindings of past situations could be put to good use. Two of these are in the area of plan repair. Briefly, *directed remindings* could be used to build explanations as to why a plan has failed out of past explanations of similar failures. It also could use memories of past repairs to fix current failures.

Just as CHEF reuses past plans rather than chaining together new plans out of primitive actions, an explainer could reuse past explanations from similar situations rather than chaining together new ones to explain current failures. A plan repair mechanism could also use memories of past experiences to avoid the replanning operations normally associated with plan repair.

6 Conclusions

The CHEF planner makes use past memories of its own experiences, in the form of *directed remindings*, in order to do much of its planning work. The view of reminding that it presents is an active one in which the needs of the planner define the different memory organizations that it uses. There is a relationship between the function to which a memory is going to be put and the features that are used to find it. The features are not random features that are taken from a situation and used to search memory, they are actively chosen by a planning process and reflect the needs of the process. As a result the remindings that occur also reflect those needs.

References

[Carbonell 83] Carbonell, J., Derivational Analogy and its Role in Problem Solving, *Proceedings of the National Conference on Artificial Intelligence,* AAAI, Washington, DC, August 1983.

[Dyer 82] Dyer, M., *IN-DEPTH UNDERSTANDING: A Computer Model of Integrated Processing For Narrative Comprehension,* Technical Report 219, Yale University Department of Computer Science, May 1982.

[Dyer 85] August, S. and Dyer, M., Understanding Analogies in Editorials., *Proceedings of the Ninth National Conference on Artificial Intelligence,* AAAI, Los Angeles, CA, August 1985, pp. 845–847.

[Gentner and Landers 85] Gentner, D. and Landers, R., Analogical Reminding: A Good Match is Hard to Find., *Proceedings of the International Conference on Systems, Man and Cybernetics,* Tuscon, AZ, November 1985.

[Gick and Holyoak 83] Gick, M. and Holyoak, K., *Schema induction and analogical transfer,* Cognitive Psychology, 15 (1983), pp. 1–38.

[Schank 82] Schank, R., *Dynamic memory: A theory of learning in computers and people,* Cambridge University Press, 1982.

[Schank 86] Schank, R., *Explanation Patterns: Understanding Mechanically and Creatively,* In preparation, 1986.

[Seifert 85] Seifert, C., McKoon G., Abelson R., and Ratcliff R., *Memory Connections Between Thematically Similar Episodes,* Journal of Experimental Psychology: Learning, Memory and Cognition., (1985).

KNOWING WHAT TO ASK NEXT AND WHY:
Asking Pertinent Questions Using Cases and Hypotheticals [1]

Kevin D. Ashley [2]
Department of Computer and Information Science
University of Massachusetts
Amherst, Massachusetts 01002

1. Introduction

In this paper, we outline a program, **HYPO**, that models reasoning with cases and hypotheticals. HYPO works in the task domain of the law, in particular, the area of trade secret protection of software. As illustrated below with actual examples, experts in the law use cases and hypotheticals as primary tools for analyzing fact situations and making arguments about them.

HYPO uses a Case Knowledge Base ("CKB") of actual legal cases to perform its basic tasks of *Analysis, Fact-gathering, Argument, Explanation* and *Hypothetical Generation.* Given a fact situation, the CASE-ANALYSIS module locates cases in the CKB that are relevant, or potentially relevant, to a legal analysis of the fact situation. HYPO's other modules use this information as follows:

- FACT-GATHERING asks the user about significant additional facts;
- ARGUMENT makes and responds to points from a legal argument about the fact situation;
- EXPLANATION explains with case examples the argument exchanges and factual queries; and
- HYPO-GEN generates hypothetical fact situations tailored to the needs of argument making and explanation.

[1] This work was supported (in part) by: Grant IST-8212238 of the National Science Foundation, the Advanced Research Projects Agency of the Department of Defense, monitored by the Office of Naval Research under contract no. N00014-84-K-0017, and an IBM Graduate Student Fellowship.

[2] Copyright ©1986. Kevin D. Ashley. All rights reserved.

This paper focusses on how HYPO analyzes a fact situation, its heuristic method for deciding what factual questions are important to ask and which question to ask next, and how it explains questions with examples, real and hypothetical.

2. Examples of How Experts Reason with Cases and Hypotheticals

People expect human experts to be able to explain by giving examples, to pose hypotheticals to demonstrate the consequences of contemplated actions, to think up counterexamples of proposed rules, and to learn from past cases. People assess an expert's understanding of a domain in part by how effectively the expert uses examples and hypotheticals.

Experts reason with cases and hypotheticals in many fields such as mathematics, medicine, and business management, but particularly in law and especially in legal argument.

Oral arguments before the United States Supreme Court illustrate the uses that attorneys make of cases and hypotheticals as tools in argument. To the chagrin of counsel before the bar of the Supreme Court, the Justices frequently interrupt an attorney's presentation to pose hypotheticals. For example, in *Lynch v. Donnelly*, a case involving the constitutionality of the Christmas creche display of the City of Pawtucket, Justices posed the following hypotheticals:

> To the attorney for the City:
>
> Q: Do you think ... that a city could display a nativity scene alone without other displays such as Santa Claus and Christmas trees...?
>
> Q: [C]ould the city display a cross for the celebration of Easter, under your view?
>
> To the attorney opposing the display:
>
> Q: [S]upposing the creche were just one ornament on the Christmas tree and you could hardly see it unless you looked very closely, would that be illegal?
>
> Q: What if they had three wisemen and a star in one exhibit, say? Would that be enough? ... What if you had an exhibit that had not the creche itself, but just three camels out in the desert and a star up in the sky?
>
> Q: Well, the city could not display religious paintings or artifacts in its museum under your theory.
>
> Q: There is nothing self-explanatory about a creche to somebody ... who has never been exposed to the Christian religion.

> Q: Would the display up on the frieze in this courtroom of the Ten Commandments be unconstitutional then, in your view?
>
> Q: Several years ago ... there was a ceremony held on the Mall, which is federal property of course. ...[T]here were 200,000 or 300,000 people ... and the ceremony was presided over by Pope John Paul II. Would you say that was a step toward an establishment of religion violative of the religion clauses? ... Then you think it would be all right to put a creche over on the Mall? ... How do you distinguish a high mass from a creche? ... [T]here was a considerable involvement of government in that ceremony, hundreds of extra policemen on duty, streets closed for traffic control purposes, and all that sort of thing. That was a considerable governmental involvement, was it not? [*SUP, Lynch v. Donnelly*, Case No. 82-1256, Fiche No. 5, pages 9,11,32,37–45].

In the above questions, one can see the Justices modifying the fact situation along various dimensions: changing the location, focus, size, and symbolic religious content of the display, the nature of the viewer, and the degree of government involvement. Sometimes the purpose of the modifications is to compare the fact situation to actual cases previously decided by the court to test whether the current situation presents stronger or weaker facts. [3] Or the actual "case", like the Mall example, may be significant because it did *not* give rise to litigation.

Frequently, the Justices use the hypothetical to apply pressure to the rule proposed by an attorney for deciding the case. That can be seen in the Mall example above and in the following example from *New Jersey v. T.L.O*, a case involving the constitutionality of a high school vice principal's search of a female student's handbag for cigarettes after a teacher reported that she had been smoking in the girls room. A Justice asked:

> Q: Do you think then that a male teacher could conduct a pat-down search of a young woman at age sixteen to find the cigarettes?

In response, the attorney for the state took the position that the Fourth Amendment of the United States Constitution, which has been interpreted as prohibitting unreasonable searches by law enforcement authorities, does not apply to high school administrators. The Justice rejoined:

---

[3] See e.g., *Stone v. Graham*, 449 U.S. 39 (1980): Posting copies of Ten Commandments in schools held unconstitutional; *Gilfillan v. City of Philadelphia*, 637 F. 2d 924 (CA3, 1980): City-financed platform and cross used by Pope John Paul II to celebrate public mass held unconstitutional; *McCreary v. Stone*, 575 F.Supp. 1112 (SDNY 1983): Not unconstitutional for village not to refuse permit to private group to erect creche in public park.

> Q: And does that mean that their authority then to make searches, if the Fourth Amendment is completely inapplicable, extends to any kind of search, strip search or otherwise? [*SUP, New Jersey v. T.L.O.*, 1984 Term, Fiche No. 5, pages 13–22].

In the last example, although the altered fact situation posed by the Justice is still covered by the proposed rule, it is increasingly harder for the attorney to justify applying the rule to the hypothetical because the latter presents progressively weaker facts. The Justice "stacks" the hypothetical with more extreme facts that weigh against the party in the hypothetical who corresponds to the attorney's client. The attorney is forced to distinguish the hypothetical, to come up with some alternative explanation for why the hypothetical and the current fact situation need not be decided the same way.

To summarize, the above examples show how cases and hypotheticals are used as rhetorical tools in argument:

- To present, support and attack positions in an argument (e.g. by testing consequences of a tentative conclusion, pressing an assertion to its limits and exploring the meaning of a concept.)
- To relate a fact situation to significant cases from past experience.
- To factor a complex situation into component parts (e.g. by exaggerating strengths, weaknesses or by hypothetically eliminating features.)
- To control the course of an argument (e.g., by focusing attention of participants in a discussion on particular issues.)

3. The Problem – Asking the Right Questions

As suggested by the above examples, one way of making a legal argument is to cite prior cases as precedents. [4] In urging a court to decide for her client, a lawyer frequently cites a favorable prior case, one whose holding is in favor of a similar party on the same kind of claim. The lawyer will argue that the facts of the precedent are similar in a relevant

[4]For purposes of this research, legal *cases* are disputes between parties tried by a court whose decisions are reported in published opinions. The opinion sets forth the facts of the case, the claims made by one party against the other, and the court's holdings. The *facts* of the case are statements about events associated with the dispute that were proved at trial or which the court assumed to be true. A *claim* is a recognized kind of complaint for which courts will grant relief. Examples of claims are breach of contract, negligence, trade secrets misappropriation, or patent infringement. The decision of the court as to the legal effect on each claim of the facts of the case, either in favor of the plaintiff or the defendant, is called a *holding*.

sense to those involving her client. Her opponent naturally will try to distinguish the cited case by pointing out some significant difference between the two fact situations.

HYPO needs to perform roughly the same tasks. It analyzes a fact situation describing a legal dispute between a plaintiff and defendant, determines what claims the plaintiff can assert and makes and responds to legal points on behalf of one or other of the parties. To make a point for the plaintiff, the program selects the most relevant pro-plaintiff case from the CKB, and cites it with a description of the important facts that the case has in common with the current fact situation. To respond to a point, the program may, among other things, distinguish the cited case by pointing out facts in the current situation that favor the defendant.

At various stages, HYPO must decide when and what questions to ask the user about additional facts that may be true of the fact situation. If a particular fact would allow the program to cite a good precedent case for a party, the program should ask the user about the fact.

The research problems addressed in this paper are how the program, given a fact situation, can decide what are the pertinent questions and explain why it asks a question. The problems are important because the user's faith in the program depends, in part, on how well the program asks questions: how many questions it asks, whether the questions evidence an understanding of the legal implications of the fact situation, and whether it has a good explanation of why it asks a question.

4. How HYPO Analyzes a Fact Situation

HYPO employs hierarchical clusters of frames for representing the features of actual legal cases (plaintiff, defendant, claim, facts, etc.) [Rissland, Valcarce, & Ashley, 1984]. The CKB contains about twelve of the leading cases involving protection of trade secrets. Four of the cases are described briefly in Appendix Table 1.

In addition to cases, HYPO has domain-specific knowledge about what clusters of facts are relevant to the legal merits of a claim (i.e., for a particular kind of case, what collections of facts represent strengths or weaknesses in a party's position.) The short answer is that facts are relevant to a claim if a court has decided such a claim in a real case having expressly noted in its opinion the presence or absence of such facts. **Dimensions** represent that knowledge in the HYPO program. Examples of dimensions in HYPO's area of trade secret law are: *Secrets-voluntarily-disclosed*, *Disclosures-subject-to-restriction*, *Competitive-advantage-gained*, *Vertical-knowledge*. These dimensions are summarized in Appendix Table 2.

Each dimension is a frame with the following slots:

Prerequisites
Focal-slots
Ranges
Direction-to-strengthen-plaintiff
Significance
Cases-indexed

As an example, the prerequisites of the dimension called **Secrets-voluntarily-disclosed**, set forth in Appendix Table 2, are that two corporations, plaintiff and defendant, compete with respect to a product, plaintiff has confidential product information to which defendant has gained access and plaintiff has made some disclosures of the information to outsiders. The prerequisites are stated in terms of factual predicates, which indicate the presence or absence of a legal fact or attribute (e.g., existence of a product, existence of a non-disclosure agreement). The focal slot of the dimension is the number of disclosees and its range is a non-negative integer. To strengthen the plaintiff's position in a fact situation to which this dimension applies, decrease the number of disclosees. The best case would be 0 disclosees. The significance of the dimension is that courts have found the prerequisites and certain focal slot values are a reason for deciding a claim of trade secrets misappropriation in favor of the defendant. The dimension indexes at least two cases: *Midland-Ross* [5] in which the court held for defendant where plaintiff disclosed the secret to 100 persons and *Data-General* [6] in which the court held for plaintiff where plaintiff disclosed to 6000 persons. We have identified some thirty dimensions in the trade secrets domain (some of the others are described in [Rissland, Valcarce & Ashley, 1984]). The dimensions were gleaned from law journal articles describing the state of the (case) law in this area [Gilburne & Johnston, 1982].

In order to perform a legal analysis of a fact situation, HYPO determines which dimensions apply to the fact situation. The prerequisites, in effect, define antecedent conditions and a dimension (i.e., a possible reason for deciding a claim in a particular way) is the consequent. This process, depicted in Figure 1, is performed by the CASE-ANALYSIS module. The overall structure of HYPO and its internal workings are described in [Ashley and Rissland, 1985; Ashley, 1986].

The output of the CASE-ANALYSIS module is the *case-analysis-record* which contains applicable factual predicates, applicable dimensions, near-miss dimensions, applicable claims, and relevant CKB cases. As a first approximation, the cases indexed by the

[5] *Midland-Ross Corp. v. Sunbeam Equipment Corp.*, 316 F.Supp. 171 (W.D. Pa., 1970), Appendix Table 1.

[6] *Data General Corp. v. Digital Computer Controls, Inc.*, 357 A.2d 105 (Del. Ch. 1975), Appendix Table 1.

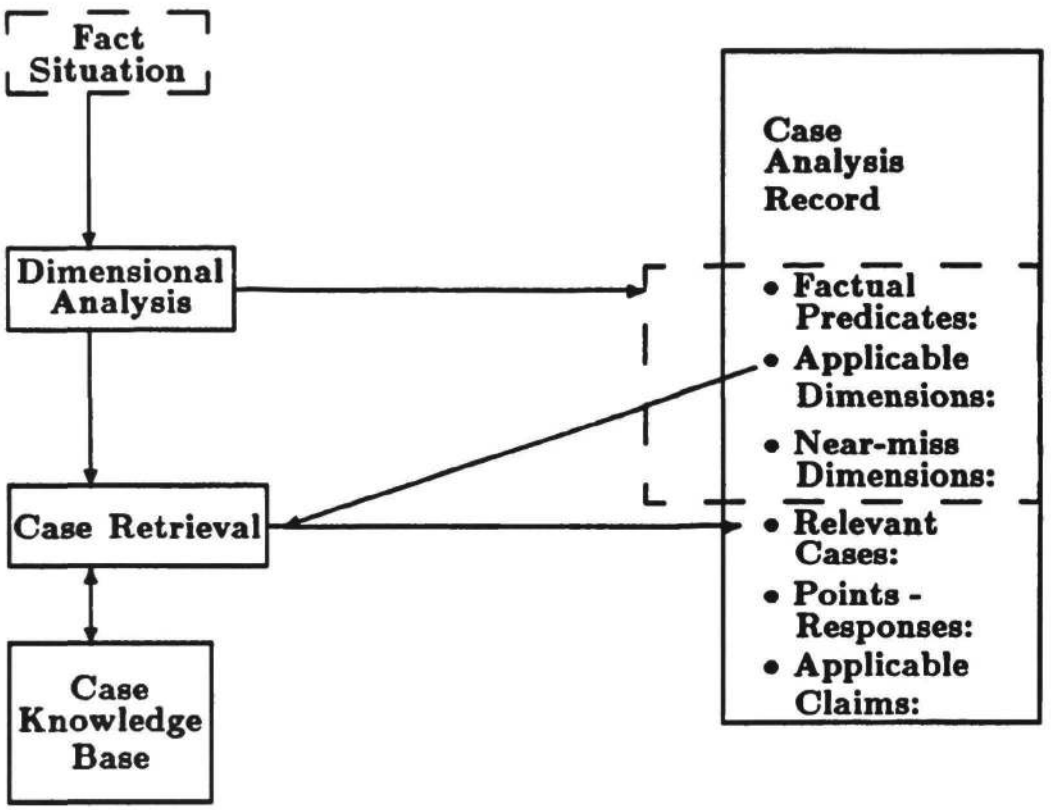


Figure 1: HYPO's CASE-ANALYSIS Module.

applicable dimensions are **relevant** to the fact situation. **Near-miss** dimensions are those for which the particular prerequisite associated with the focal-slot is the only prerequisite left to be satisfied by the fact situation. In *secrets-voluntarily-disclosed*, for example, the prerequisite associated with the focal-slot, number of disclosees, is whether the plaintiff made some disclosures.

5. How HYPO Asks a Question

Basically, what HYPO does is to start with a given fact situation and ask questions about possible facts that would make precedent cases apply in favor of or against various claims of the plaintiff. Since the space of possible facts known to the system is large, it needs to be searched heuristically so that the system asks only those questions which directly pertain to making an argument about the claim.

In effect, HYPO locates the fact situation in the CKB and incrementally moves out from it in various directions toward cases that would be useful for arguing in favor of a party. Dimensions provide the guidance for directing this search.

The process takes place in the following steps:

1. Analyze the fact situation to identify the near-miss dimensions.
2. Select the near-misses that are most *pertinent* to the argument goal.
3. Ask questions to see if the near-miss dimension applies and to determine where the fact situation lies on the dimension.

The first step is performed by the CASE-ANALYSIS module and results in the case-analysis-record described in the last section.

The second step is performed by the FACT-GATHERING module. In essence, the near-miss dimensions are ordered according to their utility in satisfying the current argument goal of the user/system interaction. There are two possible goals:

1. To "build a case" in favor of a plaintiff's claim, by turning up facts upon which a strong pro-plaintiff precedent case involving the claim can be cited;
2. To advise the user of weaknesses, real or potential, by finding facts that the defendant could use to cite a strong pro-defendant case or to distinguish a pro-plaintiff case. In other words, to play Devil's Advocate;

As illustrated below in an extended example, each goal guides the FACT-GATHERING module in selecting near-miss dimensions to ask about. If the program needs to build a case in favor of a claim of the plaintiff, then it screens dimensions that index pro-plaintiff cases involving that claim. If it is playing Devil's Advocate, it (a) screens dimensions that index pro-defendant cases or (b) screens dimensions that lead to facts that distinguish the plaintiff's cases.

The third step is performed in part by the FACT-GATHERING module. The module initiates a question about the missing prerequisite of the selected near-miss dimension. [7] If the dimension is found to apply, the module inquires about the dimension's focal slot value in the fact situation. The program compares the focal slot value in the fact situation with that of a real case indexed by the dimension. The relative values of the

[7] As part of the COUNSELOR project, the HYPO program has been designed to motivate the system to generate natural language questions and explanations by example and to understand the user's answers but it does not itself express or understand natural language outputs or inputs.

focal slots determine whether the fact situation is stronger or weaker than the real case along that dimension and whether the case can be cited on behalf of a party's claim in the fact situation.

In order to illustrate this process, we will use the following hypothetical case, *Widget-King v. Cupcake*, whose facts are as follows:

> Plaintiff Widget-King and defendant Cupcake are corporations that make competing products. Widget-King has confidential information concerning its own product. Cupcake gained access to Widget-King's confidential information. [8]

The parts of the initial case-analysis-record for *Widget-King v. Cupcake* that are relevant for the following discussion are:

applicable dimensions: nil
near-miss dimensions:
competitive-advantage-gained
secrets-voluntarily-disclosed;
vertical-knowledge
relevant CKB cases: nil

The FACT-GATHERING module uses the goal and the kinds of cases indexed by the near-miss dimensions to select among them. Of the three near-miss dimensions, *competitive-advantage-gained* indexes only a pro-plaintiff case (i.e., a case that the plaintiff won), *vertical-knowledge* indexes only a pro-defendant case and *secrets-voluntarily-disclosed* indexes both a pro-defendant and pro-plaintiff case. Although the number of cases in this example is very small, as the number of cases in the CKB increases, some dimensions will continue to index only cases that favor the plaintiff or the defendant, but not both. Others will index cases of both types.

[8] Another aspect of fact gathering is guiding the user through the process of providing an initial description of the fact situation. In the COUNSELOR project, we are experimenting with using scripts about stereotypical legal disputes involving trade secrets and about attorneys' representing clients to guide the initial questioning and understanding.

If the goal is to build a favorable argument for plaintiff Widget-King, the program will select ***competitive-advantage-gained*** as the first dimension to ask about. This is the first choice because the dimension indexes only pro-plaintiff cases. The CKB contains no case in which the advantage conferred on the plaintiff by virtue of the defendant's having saved development expense was so overshadowed by facts favoring the defendant that the defendant won. If the goal were to build an argument for defendant, ***secrets-volutarily-disclosed*** would be the near-miss dimension of choice.

If ***competitive-advantage-gained*** is selected, the program first asks about the missing prerequisite, whether defendant Cupcake saved any time or expense developing its competing product. If so, the program asks how much time or expense had been saved. If the relative amount of the savings is comparable to or greater than that in *Telex*, that case can be cited in favor of plaintiff Widget-King.

The ***secrets-voluntarily-disclosed*** dimension is the program's second choice in trying to build an argument for the plaintiff. The program needs to find out if Widget-King disclosed its secrets to anyone and whether there were 6000 or fewer disclosures. If so, the *Data-General* case, with its pro-plaintiff holding despite 6000 disclosees can be cited in favor of Widget-King. The dimension is the second choice because it also indexes a pro-defendant case, *Midland-Ross*. In other words, participation in this dimension does not necessarily help the plaintiff. The same dimension might have been selected by the program acting on a goal to build an argument for the defendant. In that event, the program needs to determine if there are 100 or more disclosures. Then *Midland-Ross* could be cited on behalf of defendant Cupcake.

Assume that in response to its questions, the program is told that Widget-King made 110 disclosures. The new fact has a number of effects. The ANALYSIS module updates the case-analysis-record to reflect that ***secrets-voluntarily-disclosed*** is now an ***applicable*** dimension. *Data-General* and *Midland-Ross* are added as relevant cases. The former can be cited in favor of the plaintiff; the latter can be cited in favor of the defendant. Significantly, with the addition of disclosures to the fact situation, the ***Disclosures-subject-to-restriction*** dimension becomes a ***near-miss***.

Switching to Devil's Advocate mode, the program tries to poke holes in the plaintiff's argument by distinguishing the *Data-General* case. That is, it tries to find some fact that makes the *Data-General* case stronger for the plaintiff than the fact situation. There is such a fact. In *Data-General* all of the disclosees were subject to the restriction that they would not tell the secret to anyone else. The dimension ***Disclosures-subject-to-restriction*** captures this information and indexes the *Data-General* case. If Widget-King's disclosees are not similarly restricted, *Data-General* does not really help Widget-King's argument; it is distinguishable.

The possibility of distinguishing *Data-General* makes it important for HYPO to ask the user whether any of Widget-King's disclosures were subject to restriction and how many. Remember that the *Disclosures-subject-to-restriction* dimension has just become a near-miss dimension and also applies to *Data-General*, the case the program is trying to distinguish. On that basis, the program asks about the dimension's missing prerequisite, whether any of Widget-King's disclosures are subject to restriction and if so how many. If the *Widget-King* case is weaker along *Disclosures-subject-to-restriction*, or if that dimension does not apply at all because none of Widget-King's disclosures are subject to restriction, then *Data-General* is distinguishable.

6. Explaining a Question

It is one thing to ask a question and another to explain why the question has been asked. In HYPO, the EXPLANATION module generates explanations of, among other things, requests for additional facts.

HYPO explains a factual query to the user by posing a case example, real or hypothetical, that illustrates the reason for the request. The examples are derived from the cases indexed by the near-miss dimension that motivates the question.

In the above example, if the user demands to know why the program asked whether Widget-King had made any disclosures, the program responds, for example, that it is trying to build an argument for the defendant, that in the *Midland-Ross* case, where plaintiff had disclosed the trade secrets to 100 people, defendant won. In giving the explanation, the program summarizes only the parts of the example case that are relevant to illustrating the significance of the solicited fact by referring to the prerequisites and focal slots of the near-miss dimension that motivated the question.

Since explanations of factual queries are made with examples, the query need not be expressed as a question. Instead of asking whether the trade secret information was disclosed to anyone and then waiting for the user to ask why, the COUNSELOR system could pose a hypothetical: "**What if** Widget-King disclosed the confidential information to 100 people? In the *Midland-Ross* case, the court held for the defendant where...." In so posing the query, the system takes the initiative, retains control over the question-asking process and asks a more meaningful, pointed question initially.

In the above example, HYPO creates a hypothetical fact situation to use as part of its question and explanation. The HYPO-GEN module constructs the hypothetical by modifying a seed case, the current fact situation, to include features of a real pro-defendant target case. The heuristics employed by the HYPO-GEN module to create hypotheticals, in this case making a hypo to which a near-miss dimension applies, are described in detail in [Rissland & Ashley, 1986].

With a specification provided by the EXPLANATION and FACT-GATHERING modules and its own heuristics for building hypos, the HYPO-GEN module constructs a hypothetical that is not only legally sensible but tailored to the needs of explaining a particular question. In the above example, the specification tells HYPO-GEN to build a hypo based on the current fact situation as seed to which the near-miss dimension ***secrets-voluntarily-disclosed*** applies and to make the hypo as weak for the plaintiff along that dimension as the target case, *Midland-Ross*. HYPO-GEN makes the near-miss dimension apply to the hypo by adding facts to the seed case corresponding to the dimension's missing prerequisite (i.e., 100 disclosures, as in *Midland-Ross*.) The hypo can be made stronger or weaker for the plaintiff by modifying the facts of the seed case corresponding to the focal slot of the dimension (e.g., increasing the number of disclosees.) The dimension's focal-slot, range and direction-to-strengthen-plaintiff slots contain the information necessary for making the hypo stronger or weaker along the dimension.

HYPO's facility for generating hypotheticals allows the program to guide the user through the space of possibly significant facts and precedent cases and to explain their significance.

7. Conclusion

In this paper, we have illustrated with examples of oral arguments before the Supreme Court how case examples and hypotheticals are primary tools for analyzing and making arguments about fact situations. We have described HYPO, a program that models reasoning with cases and hypotheticals, and shown how the program analyzes a fact situation, how it decides what are important questions to ask and which question to ask next, and how it explains questions with examples, real and hypothetical.

APPENDIX

Telex Corp. v. IBM Corp., 510 F.2d 894 (5th Cir., 1975).
Held for plaintiff IBM on trade secrets misappropriation claim where Telex gained access to IBM's confidential product development information by hiring an IBM employee, paying him a large bonus to develop a competing product. The employee used development notes he brought from IBM. Telex saved time and expense developing the competing product.

Midland-Ross Corp. v. Sunbeam Equipment Corp., 316 F.Supp. 171 (W.D. Pa., 1970).
Held for defendant Sunbeam on trade secrets misappropriation claim where Midland-Ross disclosed it's technical product development info to 100 persons.

Data General Corp. v. Digital Computer Controls, Inc., 357 A.2d 105 (Del. Ch. 1975).
Held for plaintiff Data General on trade secrets misappropriation claim where Data General disclosed its technical product development info to 6000 persons, all of whom were subject to nondisclosure agreements.

Automated Systems, Inc. v. Service Bureau Corp., 401 F.2d 619 (10th Cir., 1968).
Held for defendant SBC on trade secrets misappropriation claim where Automated-Systems' confidential info was about customer's business operations (i.e., vertical info).

Table 1: Sample Cases from Case Knowledge Base.

Secrets-voluntarily-disclosed:

Significance: Plaintiff's (P's) position stronger the fewer persons to whom secrets disclosed.

Prerequisites: P and Defendant (D) compete; D had access to P's product information; P made some disclosures.

Focal slot: Number of disclosees. **To Strengthen P:** Decrease number of disclosees. **Range:** 0 to N. **Cases indexed:** *Midland-Ross, Data-General*

Disclosures-subject-to-restriction:

Significance: P's position stronger the fewer disclosees not subject to nondisclosure agreements.

Prerequisites: Competition; access to info; some disclosures and nondisclosure agreements.

Focal slot: Number of disclosees subject to restriction. **To Strengthen P:** Increase percentage of disclosees subject to restriction. **Range:** 0 - 100 %. **Cases indexed:** *Data-General*

Competitive-advantage-gained:

Significance: P's position stronger the greater competitive advantage gained by D.

Prerequisites: Competition; access to info; D saved some expense.

Focal slot: Development expense saved. **To Strengthen P:** Increase expense saved by D. **Range:** 0 - 100 %. **Cases indexed:** *Telex v. IBM*

Vertical-knowledge:

Significance: P's position stronger if information technical, not vertical.

Prerequisites: P and D compete; D had access to P's product information; info about something.

Focal slot: What information is about. **To Strengthen P:** Make information about technical development of product. **Range:** {technical, vertical} **Cases indexed:** *Automated Systems*, et al.

Table 2: Sample Dimensions.

REFERENCES

[1] Kevin D. Ashley. *Modelling Legal Argument: Reasoning with Cases and Hypotheticals - A Thesis Proposal.* Project Memo 10, The COUNSELOR Project, Department of Computer and Information Science, University of Massachusetts, 1986.

[2] Kevin D. Ashley and Edwina L. Rissland. Toward Modelling Legal Argument. In Antonio A. Martino and Fiorenza Socci Natali, editors, *Atti preliminari del II Convegno internazionale di studi su Logica Informatica Diritto*, pages 97–108, Consiglio Nazionale delle Ricerche, Istituto per la documentazione giuridica, Florence, Italy, September 1985.

[3] M. R. Gilburne and R. L. Johnston. Trade Secret Protection for Software Generally and in the Mass Market. *Computer/Law Journal*, III(3), 1982.

[4] Edwina L. Rissland and Kevin D. Ashley. Hypotheticals as Heuristic Device. In *Proceedings of the Fifth National Conference on Artificial Intelligence*, American Association for Artificial Intelligence, August 1986. Philadelphia, PA.

[5] Edwina L. Rissland, E. M. Valcarce, and Kevin D. Ashley. Explaining and Arguing with Examples. In *Proceedings of the Fourth National Conference on Artificial Intelligence*, American Association for Artificial Intelligence, Austin, TX, August 1984.

[6] SUP The Complete Oral Arguments of the Supreme Court of the United States. University Publications of America, Frederick, MD.

Towards a Memory Architecture that Supports Reminding

Janet L. Kolodner
Richard E. Cullingford

School of Information and Computer Science
Georgia Institute of Technology
Atlanta, GA 30332

Abstract

The phenomena of *reminding* has been receiving quite a bit of attention in the past few years. Researchers have been looking at how previous experience can help in understanding and problem solving. As much of this work has shown, reminding is a complicated process. In this paper, we present requirements on a cognitive architecture that can promote and support these observations. The architectural requirements are those needed by a machine that can use experience to reason. To motivate these requirements, we present observations of reminding that are based on analyses of both actual remindings and hypothetical cases.

Why Worry About Reminding?

The phenomenon of *reminding* has been receiving quite a bit of attention in the past few years (e.g., Kolodner, 1983; Lebowitz, 1983; Reiser, 1983; Schank, 1982). Researchers have been looking at how previous experience can help in understanding (Riesbeck, 1981; Schank, 1982) and problem solving (Bain, 1985; Carbonell, 1983; Hammond, 1984; Holyoak, 1985; Kolodner et al., 1985; Ross, 1982). As much of this work has shown, reminding is a complicated process. Our own research on problem solving (Kolodner, 1983b, 1985; Kolodner & Kolodner, 1985; Kolodner et al., 1984) has led us to analyze many instances of reminding, and those instances along with the ones already reported in the literature (in, e.g., Reiser, 1983; Schank, 1982), lead us to believe that there is more regularity to reminding than is evident on the surface.

The questions we ask in this paper are the following: What are the fundamental memory processes and structures that support reminding? What kind of architecture do these processes and structures require? In this paper, we review and add to the set of things known about reminding. With each observation about reminding that we present, we also present the architectural features necessary to support it.

This research is motivated by our research into the role of experience in problem solving and understanding. We have built memory programs that can be reminded (e.g., CYRUS (Kolodner, 1983a, 1984)) and problem solvers that can use the results of reminding to solve problems analogically (e.g., MEDIATOR (Kolodner, et al., 1985), PERSUADER (Sycara, 1985), Consumer Advisor (Turner, 1986)). Each of these problem solvers is able to solve problems in fewer steps than would be required if they were solving the problems from scratch. Each gets better over time as a result of experience. Each, however, is very slow, primarily because of the overhead of maintaining and accessing the memory for experience. We have therefore been forced to ask ourselves the questions: Where does this slowness originate from and is there a way to get around it? At first blush, the answer to this seems obvious: The memory model we are using has slowness inherent in its search procedures because they are parallel procedures implemented on a serial machine (Kolodner, 1983a, 1983b, 1984). Certainly, if we had a parallel machine, we could speed up the memory search. But this is not the whole answer, and parallelism is not the only property we want out of a cognitive architecture that can support reminding. It should also support memory's other functions. We have therefore spent some time analyzing several cases of reminding to find out what kind of memory architecture would best support it and to find out how the process can be speeded up.

An Example

We begin by putting reminding into context with respect to other reasoning processes. The dialog below is from a conversation we recorded recently while attempting to explore the process of advice giving. It illustrates an interesting reminding that we use for illustrative purposes throughout this paper.

Advisee: I need to *buy* a bookshelf to use in my study at home. The study is carpeted. Since I'm in school now, I can't afford anything expensive. I'm planning to use the bookshelf for textbooks and paperbacks. Can you suggest what kind of bookshelf I need to get?

Advisor: I don't know what bookshelves run for these days. Have you ever considered *building* your own? My roommate did that last year. He was running out of space for his books and went out and bought some lumber one weekend and built some bookshelves ...

This reminding is striking. Although the advisee was asking for advice about "buying", the reminding was of a "building" episode. There are two possible explanations for this. The problem solver inside the person could have been working on solving the original problem -- buying bookshelves. In this case, it would have attempted to elaborate its buying plan taking the advisee's needs into account. When it came up against a dead end (in this case, the fact that the advisee is a student and therefore may not have enough money to buy a durable set of shelves), it considered the goal the advisee was trying to achieve (acquisition of bookshelves), began to generate a plan more amenable to the situation (in this case, building), and in considering the "building" option, was reminded of the building episode above. We call that episode the "Pete's bookshelves" episode.

Our other explanation is the following: Perhaps the problem solver was not doing much work at all. Rather, memory (a separate processor) came up with the previous case, handed it to the problem solver, and the problem solver used it to derive an answer. In other words, the reminding of the "building" episode came through memory's consideration of the "buying" request itself or some part of the buying request.

While the first explanation makes the problem solver bear the brunt of the advisor's reasoning, the second divides the work between several different processes: the memory, a case-based reasoner, and the problem solver. How can multiple reasoning processes interact to make remindings happen and to use them appropriately? Part of the answer to this question lies in *memory's representations*. The memory must know that "buying" is a plan that is usually instrumental to achieving an acquisition goal. It must also relate the acquisition goal to the kinds of plans that can achieve it, e.g., "buying" and "building". The problem solver and memory must share these representations. The general problems to be considered here are what kinds of knowledge structures the memory has and what kinds of links there are between the knowledge structures. Another part of the answer has to do with *understanding processes*. Here we refer to the hypothesis that understanders are always doing *goal tracking*. In other words, they are always trying to determine why an actor would do the actions reported in a text or conversation. Thus, inferring an acquisition goal is automatic when hearing about a "buy" plan. Another part of the answer revolves around *memory's organization*. There is probably agreement that the knowledge structure "build" is in some way referenced by the knowledge structure associated with "acquisition". But what are the connections between the *building episode* reported in the reminding and the knowledge structure "build"? How are the knowledge structures and the individual experiences reported in remindings connected? The fourth part of the answer lies in defining the *processing a memory is able and allowed to do* on this representation. Which links can be passively traversed? Which require strategic processing? What kinds of processing must be easy for the memory to do? Finally, we must consider the *interactions between the memory and the reasoner*. The reasoner might be a language understander or a problem solver. Either way, it needs memory's knowledge to do its work. By the same token, memory, which gets reminded of previous cases in the course of reasoning must always be aware of the reasoning done by the reasoner. There must thus be means of communication between the memory and reasoning processes.

Reminding

We continue by going over previous hypotheses about reminding. Schank (1982) postulates that the *interesting* kinds of remindings arise from processing considerations: As a new event or text is being understood or analyzed, the knowledge structures that can provide the most specific explanations of the new episode are accessed. Sometimes those knowledge structures are generalized structures, while sometimes they are particular previous episodes. The particular previous episodes that are remembered in the

course of understanding or analysis are what we call remindings, and they arise because of the need to make inferences.

Schank (1982) and others (Kolodner, 1983a, 1984; Reiser, 1983) describe a memory organization that supports such reminding. As new items are being processed by memory's structures (called MOPs), they are at the same time being indexed in those structures by a subset of their features not predicted by the generalized knowledge structures. That subset includes features that are predictive of other domain-related features (Kolodner, 1983a). This provides for several later functions. First, it allows a probe that describes an item to be used to retrieve that item from memory. Second, it stores the item such that if a later event that is distinct in ways similar to the first one is encountered, the first item will be found and available to inference mechanisms. Such a scheme allows an inferencer to recognize a situation that is prone to failure and also provides predictions of ways to get out of failure situations (Hammond, 1984; Kolodner et al., 1985; Riesbeck, 1981; Schank, 1982). It is thus useful for both understanding and problem solving.

CYRUS' "Diplomatic Meetings" MOP

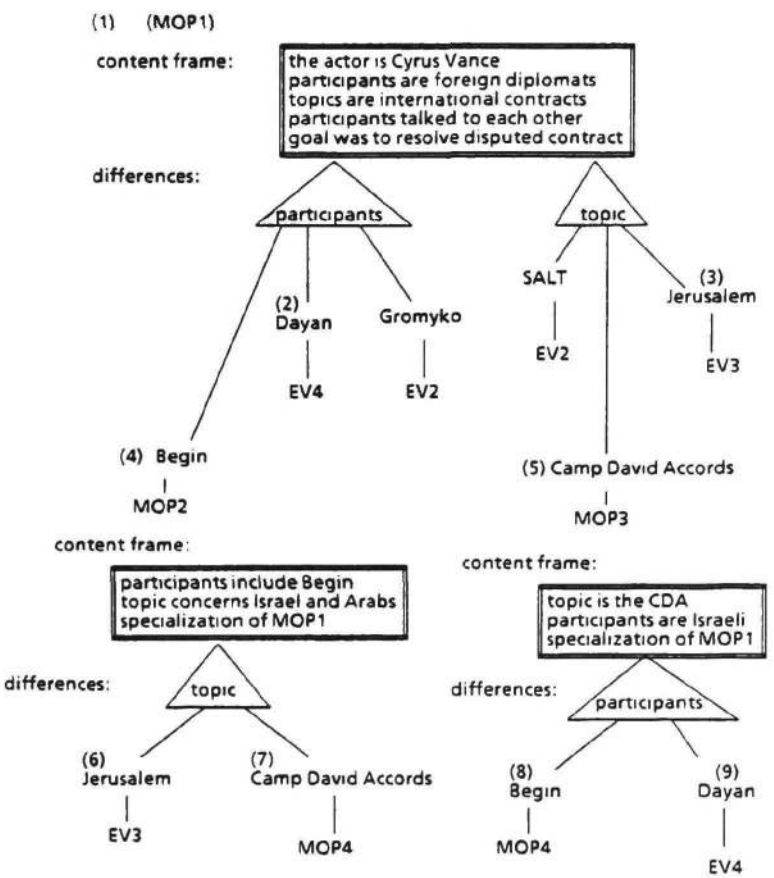


Figure 1

Figure 1 shows one of the MOPs used in CYRUS (Kolodner, 1983a, 1984). CYRUS kept track of the day to day events in the life of Cyrus Vance when he was U.S. Secretary of State. Thus, its memory structures correspond to the types of situations a secretary of state is involved in. This particular MOP organizes generalized knowledge about and individual instances of "diplomatic meetings." Since participants in such meetings can predict the topic under discussion, and topic in such situations can predict some other things, each is used for indexing in the "diplomatic meeting" context. This particular MOP holds meetings with Begin, Dayan, and Gromyko, and since there were several with Begin and all had

something in common (their topics were similar (see MOP2 in the figure)), a sub-MOP of "diplomatic meetings" representing those with Begin was created. For a similar reason, there is a sub-MOP (MOP3) representing meetings about the Camp David Accords. Based on these generalizations, if an understander or problem solver with this MOP in its memory were to encounter a diplomatic meeting about the Camp David Accords, it would be able to predict some of the participants based on the knowledge available in MOP3 that participants are Israeli. If it were to encounter a diplomatic meeting about SALT, it would be able to make predictions based on the specifics of that meeting.

Before going into the intracacies of reminding, we consider the traversal process that works on these structures and the basic architectural requirements this process puts on the memory. As described by Kolodner (1983a, 1984) and implemented in the CYRUS system, memory traversal is a parallel process. A probe is received by the memory. The first thing that is done is to determine which memory structures are to be traversed. The probe is then transformed to fit the representation of each of those memory structures. These two steps make up a process called *context instantiation*. If memory were given a probe such as "Vance talked to Gromyko," for example, it might choose "diplomatic meetings" as one of the memory structures to be traversed and would transform the original probe into "Vance attended a diplomatic meeting at which Gromyko was also a participant."

The next step is *traversal of the chosen memory structure*. The probe is compared to the generalized knowledge associated with the chosen memory structure. Those of its features that are not expected by the memory structure but that are salient to the knowledge structure's domain are extracted from the memory probe. Indexes associated with those features are traversed. If an event is reached, it is recalled. If another knowledge structure is reached, either it is recalled or the process is repeated to find more specific events. For the Vance talk with Gromyko, features that are not expected by "diplomatic meetings" but that are salient to diplomatic activities are the facts that Gromyko is a Russian, that he was an ambassador, and that he is Gromyko. Links associated with each of these features would be traversed and if an event were found, it would be returned. Addition of a new item to memory works in the same way. Instead of merely traversing links, however, a new link is created for each of the features that does not already have a link associated with it. For those features that do have links associated with them, any collisions with other items already in memory leaves open the potential for generalization, while collisions with other generalized memory structures allow consistency checks and updating of generalizations.

Because memory links can only be traversed if their labels are available as memory cues, interesting problems arise when a memory probe is too general (i.e., it describes many events) and when the memory probe specifies features that are not associated with memory links. In these cases, memory does *guided elaboration* of the memory probe in an attempt to generate additional cues. Elaborations of the diplomatic meeting referred to above might include its possible place (Washington, New York, or the UN Building), its possible topic (one that concerns the US and USSR, possibly SALT), other participants, or a particular situational setting (e.g., a summit conference), among a host of other possibilities. After elaboration, memory can be probed with the generated cues.

Elaboration can be either strategic or automatic. An automatic elaboration is one that is made on the basis of a close association. Gromyko, for example, lived and worked in New York, so an elaboration of New York as a possible meeting place can be fairly automatic. Russia has a set of important political topics associated with it, of which SALT was the most important at the time Vance was Secretary of State. That too, then, would be an automatic elaboration. Strategic elaboration is necessary when hypothetical features are more remote. Generating a hypothetical topic for a meeting based on what was going on in the world at the time of the targetted meeting would be a strategic elaboration. Strategic elaborations can be arbitrarily complex and may require a fair amount of problem solving.

What does this require in terms of an architecture? First and foremost, it requires *parallel traversal of memory links*. This seems necessary because of the large number of cues that are normally available for any memory probe. If memory links emanating from knowledge structures were to be traversed serially, there would be serious timing problems. Traversal is an easy process, and this seems like a process fit for parallelism. The memory architecture described in CYRUS works such that traversal is done concurrently for each available cue. Second, and equally important, it requires *associative mechanisms that acknowledge "locks" on each of memory's links*. Rather than allowing indescriminate traversal of memory

links as in a traditional spreading activation scheme, the MOPs scheme allows traversal only when a link's label is specified. This associative mechanism would fill the role of a quick pattern matcher. Third, it requires an architecture that can support both generalized structures and individual event descriptions at its nodes, and whose nodes can hold almost an arbitrary amount of generalized knowledge. This knowledge includes descriptive knowledge used for recognition and prediction as well as knowledge about salience and interestingness of features.

Observations about Reminding

The scheme presented above allows for reminding each time there are similarities between events. We continue by presenting several additional observations about reminding that add to the list of architectural requirements.

1. Remindings usually come from concrete scenes rather than abstract events.

Our analyses of remindings tell us that remindings tend to be of concrete situations. Remindings that seem to be of abstract events (e.g., a trip reminding someone of another trip) can usually be analyzed as coming from a concrete scene of the abstract episode (e.g., the reminding came after discussion of a particular taxicab ride that happened during one of the trips). To make this clearer, we must consider what an *abstract episode* is and what a *concrete scene* is. An abstract episode is made up of a variety of smaller events or scenes. It may be described in terms of an overall goal. Examples include trips and restaurant visits. A concrete scene, on the other hand, is usually relatively short in duration and strongly associated with a physical setting. It is a component of a larger episodic context and usually achieves a subgoal of that episode. The subgoal may be a contributor to the main goal (i.e., a precondition), the main goal itself, or a "clean-up" goal (i.e., a post-condition). Examples include checking in at a hotel and the paying scene in a restaurant.

To restate the premise in terms of knowledge structures, a reminding that seems to be based on an abstract knowledge structure (representing an abstract event) can usually be interpreted as a reminding based on a scene of that knowledge structure when analyzed in greater detail. While we observe people being reminded of whole episodes on the basis of similar episodes, further analysis usually leads us to conclude that the reminding derived from scene descriptions.

Consider, for example, one of Schank's (1982) restaurant examples. In this particular example, he goes for a visit to a company and eats in the company restaurant, where diners do their ordering by filling out a form. At a later time, he goes to another company, goes to eat in their restaurant, and must write down his order again. At this point, he was reminded of the previous restaurant experience. There are two different analyses of this event. In the first, one restaurant experience reminds him of another restaurant experience. In the second analysis, a particular scene deviation in one restaurant experience reminds him of the same deviation in another restaurant experience, and this scene reminding allows him to remember the rest of the previous episode. It is this second explanation that we claim is the more correct one, and Reiser's (1983) experiments bear this out.*

Schank's MOPs provide a structure that supports this kind of reminding. MOPs can have two purposes: they can package scenes and they can organize memories of events. Abstract knowledge structures, called MOPs (e.g., eating at a restaurant, going to the doctor, building bookshelves), are the ones that organize scenes. Scenes (e.g., ordering, sitting in the waiting room, selecting materials, installing the completed object), which are more specific and tend to be included in several different types of more abstract situations, organize memories of events.

This gives us one way of explaining the "Pete's bookshelves" reminding. Both "building" and "buying" of a large object require going to a store, selection of the object, installation of the object, etc. Each of these scenes is shared across both plans. "Installation" episodes, for example, whether they come from a "buying" or a "building" experience, are stored in the same place. If the advisor had been reminded of

*In fact, Reiser makes a stronger point in his experimentation. According to his experiments, scenes themselves are not enough for reminding. The situation in which the scene is embedded (i.e., the MOP) is also important. There is an implication in this for scene instantiation: scene instantiation is easy only when the abstract situation is available.

"Pete installing his bookshelves" or of some other "Pete's bookshelves" scene shared by "buying" and "building", then he would have been able to recall additional details of that episode, including details of doing the building. Figure 2 shows the memory structure that would allow this.

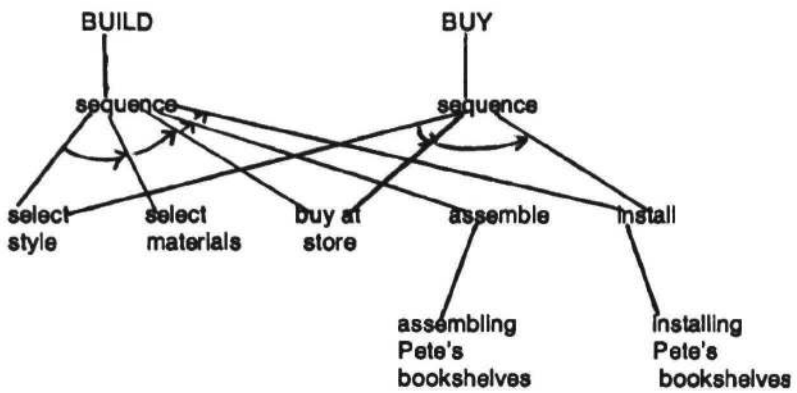


Figure 2

Now that we have the memory structure in place that can explain this reminding, we must consider the processing necessary to make the reminding happen at the right time. We have two competing explanations for this: One we will call the *elaborative hypothesis*, and the other we will call the *automatic traversal hypothesis*.

According to the *elaborative hypothesis*, some elaborative process is responsible for selecting a scene from the abstract situation (MOP).* That scene is then instantiated and traversed using the current scene description. In this case, the elaborative hypothesis might be applied several ways to result in the "Pete's bookshelves" reminding. The advisor might consider the sequence of buying a bookshelf, going through each of the major scenes, instantiating a hypothetical scene for this episode and reasoning about it. For each scene, he would take his scene description and use it to traverse the indexing network associated with the scene. If something similar to the hypothetical scene were in memory, it would be recalled at the scene level. In this case, there is a chance for reminding of "Pete's bookshelves" in any scene that "building" and "buying" have in common.

Alternatively, the advisor might be attempting to construct an image of a bookshelf with textbooks in it in a study. This in turn, might make him ask the question: How did it get there? This would lead to instantiation of the "installation" scene with the object being installed the one described by the advisee. At this point "Pete's bookshelves" would be remembered.

What does this require of a memory? It requires that *scene instantiation be fast*. This would allow an episode description at an abstract level to be described as specific scenes quickly. It also requires that the knowledge necessary for elaboration be easily available.

The alternative hypothesis, *automatic traversal*, introduces a new kind of parallel traversal of memory into the model. While previously, the only kind of traversal that was permitted was traversal of links whose labels were specified in retrieval cues, the new kind of traversal allows automatic traversal from abstract situations (MOPs) to well-known scenes. Under this hypothesis, when an abstract memory structure (MOP) is accessed, memory automatically instantiates each of the well-known scenes using the given specifications and traverses the structures of each in parallel.** This hypothesis places

*Particular strategies for elaboration are discussed and presented in (Kolodner, 1983a, 1984). Discussions of elaboration in general can be found in (Schank, 1982) and (Williams & Hollan, 1981).

**While Schank explains all the remindings in his book based on the elaborative hypothesis, we have been able to find explanations for each based on the automatic traversal hypothesis also. The advantage of the automatic traversal hypothesis in these cases is that an explanation of a problem solving failure does not have to be constructed before reminding happens. Rather, the explanation can be derived from the reminding or can be constructed on the basis of two episodes. While Schank claims that explanation is guiding reminding,

another requirement on the architecture: It requires that scene instantiation should be *automatic* in the architecture as well as fast.***

2. Remindings often seem very visual, but are only rarely of objects themselves. They are usually of objects in the context of some situation (i.e., a scene).

People often report remindings that are very visual (e.g., the legs on a particular table I saw while house hunting reminded me of the legs on my mother's dining room table). There are two ways to interpret such remindings: a view of an object may remind a person of a view of another object or a view of an object *in a situation* may remind a person of an object *in an analogous situation.* Our analyses of several such visual remindings lead us to believe that this second explanation is predominant. Usually the remindings people have of objects are not remindings of just the object itself, but of the object in a setting or scene. In the table example just given, standing in the position where the legs of the table could be viewed in a certain way reminded the person involved of a situation when she was standing in her mother's dining room and looking at the table legs int he same way.

Consider, again, the "Pete's bookshelves" example. One way we have tried to explain that reminding is by saying that the bookshelf description given by the advisee in the initial problem description reminded the advisor of a bookshelf he had *seen* that fit that description. But in that case, would the reminding come from a description of the bookshelves themselves, or from placing the given description of bookshelves in several situational contexts (e.g., "buying materials" "construction", "installation", "designing") and being reminded through one of those contexts.

We are more comfortable with the second description since visual reminding is then another example of reminding based on concrete scenes derived through elaboration. There are several hypotheses we can make based on this interpretation: (1) Organization of memories is rarely around objects themselves, but rather around situations the objects might be embedded in. (2) Visual memories might not be organized very much differently than other memories.

3. There may be many explanations of how a particular reminding happens. Remindings can be situation-specific (scene based) or thematic.

We again consider the reminding in the dialog above. Since it was not a reminding that one of us had, we cannot know exactly where it came from. We can, however, try to explain it reasonably, and we find that there are several reasonable explanations. The explanations we have given up to now were based on situation-specific knowledge structures (scenes) and in-context remindings. We can also explain the reminding in terms of thematic similarities between the advisee's problem and "Pete's bookshelves." Consider the following scenerio for reminding:

> The advisor knew that Pete had limited resources when he made his bookshelves, and additionally that he was a student or that his books were abnormally tall or heavy (like textbooks).

Our discussion so far has been about the usefulness of concrete situations represented by scenes for reminding. The scenerio above makes no references to concrete situations. Instead, it refers to a set of goals and conditions on those goals common to the two cases. We propose that the reminding comes through a common constellation of goals (acquisition of an object, preserve a limited resource) couples

---

that is not necessarily always the case. Rather, reminding can help with explanation. This is, of course, one of the roles experience should play in reasoning.

*** It may seem that we are inventing this piece of processing in an ad hoc way. Having invented it, however, we can explain its usefulness. Especially in planning, it is useful to be able to recall relevant planning failures as soon as possible. This is one way to enable that. Some may argue that doing it this way may result in a proliferation of remindings that we don't see in people. This is not necessarily the case, however, since instantiations of scenes will be fairly plain, and the only remindings we would therefore expect would be based on features of a scene if all ran according to the specifications so far.

with a set of common features novel to that goal set.

We explain this scenerio and others relying on plan and goal similarities by referring to thematic or cross-contextual knowledge structures in memory. The thematic structure that we call "Acquire Object; Limited Resources" (AO;LR) is important here. Such a knowledge structure represents knowledge about the interaction of goals, the interactions of plans and goals, and conditions on goal achievement. Schank (1982) calls these structures TOPs. This particular TOP represents knowledge about a particular kind of goal-goal interaction in which the normal plan for achievement of an acquisition goal violates a resource preservation goal.* In these cases, there is a goal of acquiring an object (usually done by "buy"), and there is also a goal of conserving a limited resource that would normally be used for acquisition (money or time). The kinds of situations represented by TOPs call for planning such that all conditions are taken care of, and there are often known ways of dealing with such situations.** In this case, buying second hand, buying at discount stores, and building are ways to conserve money, while using an agent or ordering from a catalog may be ways to conserve time.

TOPs organize plans that take multiple goals and conditions into account, but how? There are two ways. One relies on application of strategic planning knowledge. The other is more like the organization of events in MOPs: Events are organized by predictive differentiating features and found by automatic traversal of links whose values are specified in a memory probe.

Hammond (1984) gives a good explanation of the strategic organization. He claims that TOPs relate goal constellations to plans through complex questions asked of the input. The answers to the questions direct planning in appropriate directions. In AO;LR, the questions that might be asked are "what is the resource," "how important are the looks of the object," "is the object simple or complex in structure," "does the object have a lot of interacting components," "are cast-off objects of this type available," "can cast-off objects of this type be made to function appropriately," etc. These questions are the most important preconditions of the abstract plans that are available for acquiring objects with limited resources (e.g., buy second hand, buy from a catalog, build). Asking such questions of the advisee's problem description could result in either the "build" or the "buy second hand" plan being chosen. "Build" and "buy second hand" function here as abstract plans, but are examples of what we previously called *abstract situations*. They are represented by MOPs, and each has a set of more refined plan steps associated with it, represented by MOPs and eventually scenes at the most concrete levels. After choice of an abstract plan, the problem description would allow "Pete's bookshelves" to be remembered through one of those structures by traversing their scenes as described above.

According to this formulation of thematic structures, *at least some knowledge structures (thematic ones) must have strategic knowledge associated with them* that is made available to reasoning processes as soon as the thematic structure is accessed. This also gives an explanation of how situation-specific knowledge structures (MOPs) can be triggered, and provides an explanation of reminding of "Pete's bookshelves" in the course of considering several planning alternatives.

The second way TOPs organize plans is through indexing much like that done in MOPs. This organization provides an explanation for the reminding scenerio above. Here we assume that indexing in TOPs is according to features of a situation that make planning predictions. Thus, novel goals in tandem with those of the TOP would be used for indexing, as would other features (e.g., occupation, hobbies) that predict plans and plan steps appropriate to a particular individual. "Pete's bookshelves" in the scenerio above would thus be indexed in AO;LR by its novel conditions (e.g., advisee is a student). The reminding can be explained as coming from the TOP called "Acquire Object; Limited Resources" (AO;LR) by traversing an index associated with the other condition. "Pete's bookshelves," also indexed by that condition, would be remembered. If novel conditions or constraints of a new problem and an old one coincide, we expect reminding.* Figures 3 and 4 show some of the memory structures we have been discussing.

---

*It is a specialized case of a thematic situation in which plans for an achievement goal violate a preservation goal. Hammond (1984) calls this PVP.

**Hammond (1984) gives a good explanation of the use of these structures in planning.

*We do not consider here how TOPs are recognized or triggered.

Acquire Object; Limited Resources

Strategic knowledge:

What is the resource?

If time, "catalog buy"

How important are the looks of the object?

If not very important, and time is available, then "build" or "buy second hand"

Is the object simple or complex in structure?

If complex, rule out "build" unless actor has expertise

If simple, try "build"

Are cast-offs available?

If not, rule out "buy second hand"

Can cast-offs function appropriately?

If not, rule out "buy second hand"

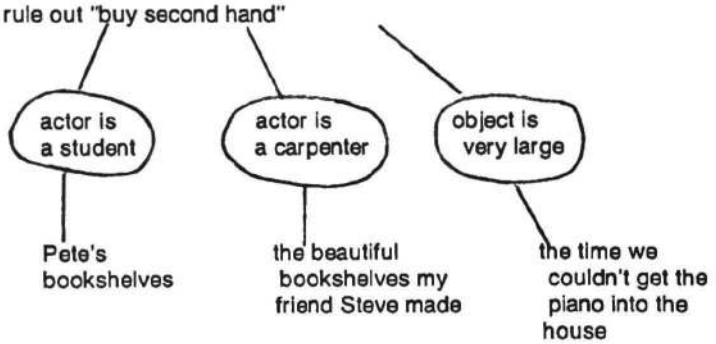


Figure 3

The BUILD plan

Preconditions:

Actor must be capable of building

Actor must have building instruments available

Sequence of events:

Select style (a plan)

Select types of materials (a plan)

Buy at store (a scene)

Assemble (a plan)

Install (a scene)

Figure 4

What kinds of processing are necessary to make all of this happen? The same ones required previously: parallel traversal of MOP/TOP indices, fast elaboration and scene instantiation, and the capability of associating almost arbitrary amounts of generalized knowledge with memory structures.

4. Remindings can be immediate or require a lot of effort.

Remindings can happen immediately when hearing of or reading about something, or might happen after a lot of understanding or problem solving effort has been expended. Remindings often require a fair amount of inference. Sometimes an explanation
for an anomolous event must be derived, and reminding is on the basis of the explanation. Sometimes reminding doesn't happen until after something goes wrong, when the reasoner is attempting to figure out what went wrong. Sometimes it comes in the process of coming up with a plan in the first place. In this case, reminding can be of past successes or failures (Kolodner, 1985). Remembering failures allows a problem solver to avoid past mistakes. Remembering successes may provide shortcuts in problem solving. Sometimes reminding happens in the course of determining whether a plan to be proposed is a good one. Suppose, for example, that Pete was known as a poor handiman (contrary to the preconditions for using the "build" plan), and not much is known about the skills of the advisee. In this case, reminding of Pete's bookshelves would confirm that the "building" plan can be done with only minor prerequisites.

The point to be made here is that if reminding can happen almost any time an agent is reasoning, then memory traversal and reasoning must be happening at the same time. This requires *an architecture in which memory is always cognizant of any reasoning that is being done.* If memory is always attempting to integrate what it sees the reasoner doing into its structures, then it can be reminded at any time based on whatever knowledge is currently available. At the same time, memory must have a way of interrupting reasoning processes and telling them that it has been reminded of a good case.

Summary: A Cognitive Architecture

We have presented a number of observations about and explanations of reminding, and have presented a set of requirements on a cognitive architecture. We summarize by listing the set of constraints on a cognitive architecture that promotes and supports reminding:

1. It must allow for highly parallel traversal of memory links.
2. It must have associative mechanisms that acknowledge "locks" on each of memory's links.
3. Memory's nodes must be able to hold either generalized structures or individual event descriptions.
4. At least some memory nodes must be able to hold the arbitrary amounts of generalized knowledge associated with generalized structures. This includes knowledge necessary for automatic elaboration and for strategic processing.
5. Instantiation of concrete scenes from memory structures representing abstract situations must be fast. Under one hypothesis about memory traversal, instantiation must also be automatic.
6. Memory processing must be concurrent with other reasoning processes and memory must be cognizant of other processing being done.

There is currently interest among both psychologists and AI researchers in reminding and the use of remindings in reasoning. At the same time, there is considerable interest among researchers in both areas in massively parallel architectures. It seems that the time is right for researchers studying high level memory processing and those studying cognitive architectures and building massively parallel machines to start talking to each other. A cognitive architecture that supports reminding requires at least some of the things being built into massively parallel hardware, in particular, the massive parallelism. It also requires several things that the architecture people have not considered: fast instantiation, high storage memory nodes, and support for strategic processing within the architecture. With hardware of this type, more massive episodic memories could be built and more sophisticated work on case-based reasoning could be done. At the same time, in attempting implementations of episodic memories on massively parallel architectures, we may be able to make our memory models more sophisticated. We hope this paper will be the beginning of an attempt to combine the needs and aims of the two groups, so that additional research will be done by both groups to make sure architecture and processing needs match.

References

Bain, W. (1985). JUDGE: A Case-Based Reasoning System. In *Proceedings of the International Machine Learning Workshop*, Skytop, PA, pp. 1-4.

Carbonell, J. G. (1983). Learning by Analogy: Formulating and Generalizing Plans from Past Experience. In Carbonell, J. G., Michalski, R. S. and Mitchell, T. M., (eds.) *Machine Learning*. Tioga Publishing Co., Palo Alto, CA.

Hammond, K. (1984). Indexing and Causality: The Organization of Plans and Strategies in Memory. Yale Department of Computer Science Technical Report #351.

Holyoak, K. (1985). The Pragmatics of Analogical Transfer. In Bower, G. (Ed.), *The Psychology of Learning and Motivation*, Vol. 19, pp. 59 - 87.

Kolodner, J. L. (1983a). Reconstructive Memory: A Computer Model. *Cognitive Science*, October, 1983.

Kolodner, J. L. (1983b). Towards an Understanding of the Role of Experience in the Evolution from Novice to Expert. *International Journal of Man-Machine Systems*, November, 1983.

Kolodner, J. L. (1984). *Retrieval and Organizational Strategies in Conceptual Memory: A Computer Model.* Lawrence Erlbaum Associates, Hillsdale, NJ.

Kolodner, J. L. (1985). Experiential Processes in Natural Problem Solving. Technical Report No. GIT-ICS-85/23. School of Information and Computer Science. Georgia Institute of Technology. Atlanta, GA 30332.

Kolodner, J. L., and Kolodner, R. M. (1985). The Role of Experience in Clinical Problem Solving: Introduction and Framework. Technical Report #GIT-ICS-85/21. School of Information and Computer Science, Georgia Institute of Technology, Atlanta, GA 30332.

Kolodner, J. L., Simpson, R. L., and Sycara-Cyranski, K. (1985). A Process Model of Case-Based Reasoning in Problem Solving. *Proceedings of the International Joint Conference on Artificial Intelligence*, Los Angeles, CA.

Lebowitz, M. (1983). Generalization from Natural Language Text. *Cognitive Science*, Vol. 7, pp. 1-40.

Reiser, B. J. (1983) Contexts and Indices in Autobiographical Memory. Cognitive Science Technical Report No. 24. Yale University. New Haven, CT.

Riesbeck, C. K. (1981). Failure-driven reminding for incremental learning. In *Proceedings of International Joint Conference on Artificial Intelligence*, Vancouver, BC.

Ross, B.H. (1982). Remindings and Their Effects in Learning a Cognitive Skill. Cognitive and Instructional Sciences Series CIS-19. Xerox Palo Alto Research Center, Palo Alto, CA.

Schank, R. C. (1982). *Dynamic Memory: A Theory of Learning in People and Computers.* Cambridge University Press, London.

Simpson, R.L. (1985). A Computer Model of Case-Based Reasoning in Problem Solving: An Investigation in the Domain of Dispute Mediation. Ph. D. Thesis. Technical Report #GIT-ICS-85/18. School of Information and Computer Science, Georgia Institute of Technology, Atlanta, GA.

Sycara-Cyranski, K. (1985). Precedent-Based Reasoning in Expert Labor Mediation. Technical Report #GIT-ICS-85/22. School of Information and Computer Science, Georgia Institute of Technology, Atlanta, GA 30332.

Turner, R. (1986). Using Preconditions in a Case-Based Advisory System. *Proceedings of the 1986 IEEE International Conference on Systems, Man, and Cybernetics*, Atlanta, GA, October, 1986.

Williams, M. W. & Hollan, J. D. (1981). The process of retrieval from very long-term memory. *Cognitive Science*, Vol. 5, pp. 87 - 119.

Application of Cognitive Science Methods to Psychotherapeutic Problem Solving: A Case Study and Some Theory

Rolf Pfeifer
Department of Computer Science
University of Zurich, Switzerland
and
Marianne Leuzinger-Bohleber
Department of Psychotherapy
University of Ulm, FRG

ABSTRACT

In this paper we present a case-study to demonstrate an application of concepts of knowledge representation from cognitive science and AI to problem solving in psychotherapeutic situations. In particular, a special type of frame, the so-called "Thematic Organization Point", or TOP, is used to characterize generic conflictive patterns of interaction, and to elucidate the meaning of a "psychotherapeutic interpretation". The concept of "failure-driven memory" is related to the process of evoking memories in patients. A belief systems analysis is used to explain why in some situations people are incapable of learning in spite of repetitive expectation failures. The underlying theory is summarized as a set of "Theorems". It is concluded that a cognitive science approach to therapeutic problem solving not only clarifies theoretical concepts but enables the derivation of powerful heuristics to be used by therapists in their practical work.

INTRODUCTION

Much of cognitive science research in one way or another has to deal with the notion of goals. One of the reasons is, of course, that behavior is controlled by the goals to be achieved. Another has to do with the fact that goals have a high explanatory value. Whenever we observe some kind of behavior, we typically interpret it as a plan to achieve some goal. Goals are ubiquitous in interpreting the environment, as well as in generating behavior, and therefore it seems natural to organize a knowledge base for problem solving (which is a kind of goal-directed behavior) around the notion of goals, goal-configurations and related ideas, such as plans, failures, etc. to be discussed below. These ideas will be applied to a special kind of expert behavior, namely the one of psychotherapists.

Psychotherapists are expert at interpreting unusual and unexpected behavior in their patients. For example, they search for "hidden" goals, i.e. goals the patient may not be aware of him or herself. One method frequently used to elicit a patient's knowledge structures is that of "free association" where memories related (by assumption) to the current behavior or experience are to be evoked. This method is used mainly in so-called *psychodynamic* therapies, i.e. therapeutic techniques dealing with the explanation and interpretation of behavior in terms of cognitive and emotional processes which may or may not be conscious. It is this kind of therapy on which our paper is focused. The memories thus activated can guide the way to an interpretation, i.e. they can help finding an explanation of some maladaptive or undesirable behavior. It is a basic tenet of many forms of

psychotherapy, that such an insight constitutes the basis of a learning process which can eventually lead to an elimination of certain behavioral patterns.

Psychotherapeutic theories are typically vague and at times even have a mythical touch. Cognitive science provides concepts and formalisms which should be capable of capturing at least a small portion of the phenomena involved. It is demonstrated in the sequel that not only concepts from psychodynamic theories can indeed be formalized in a transparent manner, but that this approach can directly support the practical work of the clinical expert.

After introducing the basic concepts, a case report from a psychotherapeutic treatment is presented. This case report will be analyzed in the subsequent sections with a particular focus on the therapist's hypotheses (her beliefs) about the mental representations of the patient (i.e. the patient's beliefs).

BASIC CONCEPTS

Much of the theory presented in the sequel is based on the work of the Yale Artificial Intelligence and Cognitive Science Programs (e.g. Schank, 1982). Basic familiarity with the concepts of frame theory in general, and Thematic Organization Points (TOPs) in particular, is assumed, and only those principles needed later on will be summarized as "Theorems"[1]. The theorems presented in this section deal with memory processes. They will be applied to elucidate concepts from psychodynamic therapy.

THEOREM 1: Reminding of episodes across contexts is enabled when the episodes are stored under the same TOP. The term "context" is used in its everyday sense.

Clearly, this is not the only way in which remindings can occur (they could e.g. also be visually based), but it is the most relevant one for goal-oriented and interactive behavior.

DEFINITION: A TOP (=Thematic Organization Point) is an abstract memory structure, a frame type, containing (at least) the following components: goal configuration, expectations about plans (or more generally: behavior) and outcomes, actual plans (behavior) and results, and explanation of discrepancies.

Theorem 2 deals with expectation violations.

THEOREM 2: (principle of "failure-driven memory") If an expectation generated from a frame is violated, a pointer is established from this frame to a representation of the situation in which the violation occurred. This enhances memory recall.

From Theorem 2 Theorem 3 can be derived (e.g. Schank, 1982):

[1]The term "Theorem" is not used in a rigorous mathematical sense but rather to designate theoretical concepts that have proved useful in many AI programs and for which there is considerable empirical support (e.g. Seifert et al., 1984), or heuristics on how to proceed in analyzing people´s knowledge structures.

THEOREM 3: The principle of "failure-driven memory" underlies a large class of learning processes.

The heuristic stated in Theorem 4 is a corollary of Theorem 3:

THEOREM 4: If learning processes are to be induced it is a good idea to violate existing expectations or to evoke expectations which can then be violated.

The theory presented so far can be used to elucidate the concept of a so-called "complete psychodynamic interpretation", and to explain why a psychotherapist should remain neutral towards a patient's goals and wishes if memories are to be evoked from the patient (see Boxes 1 and 2). Further "Theorems" will be introduced as needed in the case study.

BOX 1: A complete psychodynamic interpretation and TOPs

A major goal in psychodynamic therapies is to make hidden knowledge structures (e.g. goals or wishes) underlying maladaptive and undesirable behavior explicit in order to make them inspectable and potentially modifiable. Such hidden knowledge structures are frequently related to unresolved conflicts which have their origins in early childhood, and which can be observed in the patient's actual behavior. A central therapeutic mean to communicate hypotheses about maladaptive behavior to a patient is the one of a so-called interpretation which relates unresolved conflicts to behavior in different kinds of situations. According to Menninger (1958) an interpretation is only *complete* when a conflict has been understood with respect to three domains in the patient's life, namely: current therapeutic interaction, early childhood, and current life situation (work, lovers, etc.). Thus, similar patterns of behavior must be ferreted out and described in all three domains. It is exactly this similarity we are trying to capture with the (context independent) concept of a TOP. Such similarities are normally perceived by therapists in a rather intuitive and global fashion. They can be made more precise and transparent if TOPs are used for their description.

BOX 2: Therapeutic neutrality and "Failure-driven memory"

The practice of psychodynamic therapy has shown that in order to evoke memories which relate to hidden conflicts of a patient, it is appropriate for a therapist to take an attitude of neutrality: the therapist refuses to fulfill inadequate (i.e. "infantile") wishes of the patient, and thus earlier experiences come to mind and can be reflected in relation to current problem situations (see also Box 1). Wishes of a patient are considered inadequate if they imply any form of direct help or advice by the therapist. Therapeutic neutrality, as many other concepts in the psychodynamic literature, are based on intuitions and clinical experience and are not sufficiently grounded in theory. The principle of "failure-driven memory" (Theorem 2) explains why early childhood memories concerning infantile wishes can be evoked through a process of frustration in the therapeutic interaction: those memories relating to conflictive content typically have to do with expectation failures (e.g. unfulfilled wishes from a patient's distant past). Since, through expectation failures, those memories were originally tagged, their recall is enhanced if similar expectation failures occur within the therapeutic situation.

PFEIFER & LEUZINGER-BOHLEBER

CASE STUDY

Episodes from a psychotherapy of a 38 year old depressive woman, reported by Marianne Leuzinger-Bohleber: In July 1979 Gertrud, a German social worker, mother of seven children, married to a self-employed businessman, asked for an appointment at our counseling center in Zurich. She was in a deep crisis which was apparently triggered by her husband's starting his own business in the area of therapeutic instruction materials. It was her husband's wish that she give up her part-time job in a home in order to help him with his business. He was not a good businessman without her help and Gertrud was afraid that they might soon have to file bankruptcy.

In her psychotherapy she frequently had the feeling of giving herself up, that she would "never be able to free herself of me" and that she would come out of therapy as a "case". The sessions often had a tormenting quality and I couldn't reach her. Nevertheless, her marital situation had improved, among other things, through the fact that she had taken up working part time outside of her husband's firm again.

After some time she felt worse again. She complained about finding her life too taxing and that nothing was moving in her therapy. Casually she mentioned how nice it would be to have an additional cuddly silk cushion on the couch in the counseling room. I felt hard-pressed and had the impulse to go downtown and buy her such a cushion immediately.

In the following session I completely forgot about the cushion so that I couldn't even broach the subject. Gertrud did not say anything about it either but could hardly be reached in this and the following sessions. She felt distant, tacitly accusing and aggressive. But I had no idea that it could be related to the cushion.

The "cushion theme" re-emerged as Gertrud's fears of committing herself to therapy and the subject of terminating it were discussed. She explained how much she had been hurt by my forgetting the cushion and that her disappointment and her anger about the event were still present and would weigh heavily in her decision of whether to continue therapy or not. To her this had been proof that I was a cold, unempathetic person and that for me she was in fact only a "case," or a possibility to make money. This confirmed her experience once more that she had no bearing in the world. She felt exploited and didn't trust me any more.

This problem could be worked on therapeutically in the following weeks as she was beginning to feel that crucial infantile memories were being activated, though they were not conscious at the time. This lead to her decision to continue therapy in order to "go after" the core conflict.

After working on the cushion experience a host of memories emerge. We only render one example which will be used in our subsequent analysis: When she was about four or five (in 1946) Gertrud had to accompany her mother when at night she returned to the occupied zone to retrieve certain objects they had left behind, e.g. a typewriter. Her mother used her as a sort of "protection" against the Russian occupational forces. During these "excursions" she was scared to death but was not even allowed to cry.

Summary of report: In this paper the following key scenes are discussed:
the therapeutic situation: the patient communicates that it would be nice to have a silk cushion and doesn't get it ("cushion scene"),

- the patient's current real-life situation: the patient should help her husband with his business ("husband's business scene"),
  an early childhood experience: the patient had to accompany her mother into the occupied zone to "protect" her against the occupational forces ("occupational forces scene").

ANALYSIS OF KEY SCENES

The goal of this analysis is to make explicit the patient´s beliefs and memory representations, in particular the ones controlling her behavior. We begin with the "cushion scene". From the patient's utterance we infer that the patient has the goal to have such a cushion. This is not by logical necessity, and the attribution of goals to an actor is indeed a possible source of error. Moreover, we interpret her utterance as part of a plan to achieve her goal: the therapist is to be her "agent" (Schank & Abelson, 1977). The patient expects the therapist to fulfill her wish. However, this expectation is violated. As we find out later, the patient had an explanation of why her wish was not fulfilled, namely that the therapist is cold and non-empathetic. This preliminary analysis is summarized in Figure 1.

Patient´s structure

GOAL: SELF has silk cushion.

PLAN: SELF uses T as an "agent". SELF communicates goal to T.

EXPECTED RESULT: SELF has silk cushion.

ACTUAL RESULT: SELF does not have silk cushion.

EXPLANATION OF DISCREPANCY BETWEEN EXPECTED AND ACTUAL RESULT: T does not respond to the needs of SELF since T is cold and unempathetic.

FIGURE 1: Patient´s representation of the "cushion scene": preliminary hypothesis. "SELF" refers to the patient, "T" to the therapist.

For the purposes of the present discussion we will ignore the therapist's own goals and plans for the therapy and concentrate on the ideas she has about the patient. From the utterance of the patient she infers that the patient has the goal to have a cushion. But the patient does not explicitly talk about this wish and report her related thoughts, as could be expected in therapeutic settings. This implies that the therapist's expectations are violated: the patient wanted her to actually buy her a cushion rather than work out an appropriate interpretation. This discrepancy is also reflected in the therapist's emotional reaction. A major goal for the therapist now is to find appropriate explanations.

Frequently, when our expectations are violated and there is no adequate explanation of a person's behavior available this may be due to the fact that our hypotheses about that person's goals are inappropriate. In our example, the patient may never have had the goal to have an additional cushion. Theorem 5, adapted from Wilensky (1983), gives us a clue as to what some alternative goals might be.

THEOREM 5: The execution of a plan frequently activates "dormant" goals, i.e. goals that would otherwise not have been activated. Of particular interest are the so-called "preservation goals".

Patient's structure

T'S GOAL: To be a successful, financially independent therapist.

T'S PLAN: Use SELF as an "agent".

leads to

PRESERVATION GOAL: SELF does not want to be merely a "case" (i.e. an "agent") for the benefit of T.

PLAN FOR PRESERVATION GOAL: SELF wants from T some kind of behavior which does not conform to the rules of therapeutic technique (provide SELF with silk cushion). Mention goal in conversation.

EXPECTED RESULT FOR SELF: T gets her the cushion.

ACTUAL RESULT FOR SELF: T does not get SELF the cushion. Strong negative emotions. SELF is only a "case".

EXPLANATION OF DISCREPANCY BETWEEN EXPECTED AND ACTUAL RESULT: T deliberately ignores preservation goal of SELF. Thus SELF is merely an "agent" of T.

ACTUAL RESULT FOR T: T achieves goal: T makes money and is successful.

FIGURE 2: Patient's representation of the "cushion scene": more elaborate hypothesis. "SELF" refers to the patient, "T" to the therapist.

If we look at the intensity of the reactions of both the patient and the therapist to the "cushion scene" (the patient hardly talks to the therapist for a long time; the analyst "forgets" about it) we infer that there must be something major at issue. The major activity is the patient's involvement in the therapy. Her plan is to use the therapist as an agent to help her solving her problems. If we look through the case report for possible preservation goals which are activated by her involvement, we find a statement to the effect that she does not want to be simply a "case". Applying this idea to the "cushion scene" we find a different interpretation of the patient's utterance that it would be nice to have a cuddly silk cushion on the couch. Had the therapist assented to the implied wish of the patient, this would have been an action on the part of the therapist which lies outside of the range of therapeutic (neutral) behavior. In that case -- in the patient's beliefs -- the patient would not have been treated merely as a "case" but as a "real person"; the therapist would have fulfilled her wish even though this did not conform to the rules of psychotherapeutic technique. The utterance concerning the cushion can now be interpreted as a (clever) plan to achieve the preservation goal of remaining an independent "real" individual and not merely being a "case". Of course, such a plan could only be conceived by someone with profound knowledge of psychotherapeutic technique, a precondition the patient clearly fulfilled. It is interesting to note that the patient, in order to get her preservation goals met, tries to involve the same person as an agent who the patient holds responsible for her preservation goal in the first place (since the therapist is the one who uses her as an agent). This more elaborate analysis is summarized in Figure 2. Again, it is a hypothesis of the therapist about the

patient's mental representation. This structure is non-trivial and goes beyond a simple goal-subgoal analysis. As a matter of fact, we are dealing with a rather tricky and uncomfortable pattern of interaction.

If this interpretation of the patient's behavior is appropriate and if it is communicated to the patient, then corresponding memories should be activated from the patient's distant past (early childhood) and from her recent past (her real-life situation). And indeed, when this information could be made explicit the patient started reporting her early childhood memories. Since in early childhood, for example, there were no therapists, and no silk cushions, an abstraction has to be made from the current context, with the constraint that the relevant information concerning behavior and interaction is preserved. As mentioned earlier, these commonalities are captured by TOPs. The TOP corresponding to Figure 2 is shown in Figure 3. We will call it "GA-UP" (Goal Pursuit with Agency - Unsatisfiable Preservation Goal).

Patient's TOP

O'S GOAL: Important goal.

O'S PLAN: Use SELF as an "agent".

leads to

PRESERVATION GOAL: SELF does not want to be an "agent" of O. O is an "agent" for SELF.

EXPECTED RESULT FOR SELF: O understands and achieves goal of SELF.

ACTUAL RESULT FOR SELF: SELF is an "agent". Strong negative emotions.

EXPLANATION OF DISCREPANCY BETWEEN EXPECTED AND ACTUAL RESULT: O's goal has such high priority that O deliberately ignores preservation goal of SELF.

ACTUAL RESULT FOR O: O achieves goal.

FIGURE 3: Patient's TOP, called "GA-UP" (Goal Pursuit with Agency - Unsatisfiable Preservation Goal). "SELF" refers to the patient, "O" to the person the patient is interacting with (the "other").

For a complete psychodynamic interpretation (see Box 1) we need also a representation of an early childhood experience and of an episode concerning the patient's current life. With the TOP "GA-UP" it is straightforward to find the representations of the "husband's business scene" and the "occupational forces scene". The former is shown in Figure 4, the latter is left as an exercise to the reader (or see Pfeifer & Leuzinger-Bohleber, 1986).

Patient's structure

H'S GOAL: To be an independent, financially sound businessman.

H'S PLAN: SELF must help out in H's store.

leads to

PRESERVATION GOAL: Maintain independent life style (keep job in home). SELF does not want to be merely an "agent" to run H's business.

PLAN FOR PRESERVATION GOAL: SELF wants H to support preservation goal. SELF communicates goal to H.

EXPECTED RESULT FOR SELF: H helps SELF keep independent life style.

ACTUAL RESULT FOR SELF: H interprets needs of SELF as exaggerated. H does not support wish for independence of SELF. Strong negative emotions.

EXPLANATION OF DISCREPANCY BETWEEN EXPECTED AND ACTUAL RESULT: H does not understand the importance of independence for SELF. H exploits SELF to achieve his own goals.

ACTUAL RESULT FOR O: H achieves goal: Business of H improves.

FIGURE 4: Patient's representation of the "husband's business scene". "SELF" refers to the patient, "H" to her husband.

REPETITIVE BEHAVIOR

Frames can be used to interpret the environment as well as to generate behavior. From the patient's reports it is obvious that the TOP "GA-UP" is often activated and controlling her behavior, and each time it produces undesirable negative results. One question that immediately comes to mind is why the patient is apparently incapable of learning from her expectation failures, although they occur repetitively. This persistent repetition of some undesirable behavior in different contexts is called "repetition compulsion". Theorem 6 states one frequent reason.

THEOREM 6: If an individual is incapable of learning from her expectation failures, this may be due to the fact that a structure (a frame) different from the believed one is controlling her behavior.

Theorem 6 can be directly applied to our case study. The reason why the patient is not capable of changing her behavior, although she perceives it as undesirable, is that in her (conscious) belief, she is applying a simple planning frame (called, say, "use agent") for engaging someone as an agent to achieve her goals (and this planning frame normally works - in situations not described in the case report), whereas in "reality" she is activating the more intricate frame "GA-UP". "GA-UP" is activated whenever the following triggering conditions are fulfilled: the situation contains a person who is important to the patient; this person is pursuing an important goal (in the patient's beliefs, not in "reality"); and the patient is used as an "agent" in the plan of that person. These conditions are

fulfilled for all three key scenes described. But the patient cannot discriminate between situations in which she is applying the plan "use agent" and situations in which she is using "GA-UP", for the simple reason that she has no awareness of "GA-UP". Since her explanations of expectation failures are based on the wrong assumptions, she cannot change her behavior appropriately. One fact that adds to the severity of the problem is that "GA-UP", like "use agent" is context independent and can be instantiated in many situations.

There are two main reasons for believing that our analysis is appropriate. First, the patient's emotional reactions to her expectation violations are much too strong if the plan had simply been "use agent". The intensity can only be explained in relation to the negative and strong early childhood experiences. This relation is captured by "GA-UP" which relates the memory representations of early childhood to current experiences. Second, we had the opportunity to follow her therapeutic development. The patient experienced the recognition of "GA-UP" as a profound insight, and from that point in time, her behavior started to change: the preconditions for a learning process were met (for more detail, see Pfeifer & Leuzinger-Bohleber, 1986).

DISCUSSION

Knowledge representation: The combination of the idea of belief systems and the Yale concepts for knowledge representation enabled us to explain why certain types of behaviors can be transferred to a variety of different domains, and why some of them are so persistent even if they are perceived as undesirable. Moreover, it was possible to devise heuristics which may be used in therapeutic problem solving, and to make some of the rather opaque concepts from psychodynamic therapy more transparent. The powerful concept of a TOP was applied to the representation of patterns of interaction in different domains (e.g. Box 1). It is suggested that it may be a good heuristic for a psychotherapist to actively try to work out TOPs since they not only characterize conflictive interactions, but they may also help to activate pertinent memories (so-called "strategic reminding"; Schank, 1982).

Evocation of memories: Using the concept of "failure-driven memory" it was demonstrated why the therapist should remain neutral towards a patient's goals and wishes: memories can be evoked in this way. And memories are precious sources of information for therapeutic problem solving.

Related work: There have been many frame approaches to knowledge representation in AI and cognitive science research but for our own purposes we found the emphasis on expectations and explanations of expectation failures, as well as their relationship to memory processes (Schank, 1982) most convincing. Similar concepts were used, for example, by Dyer (1983), Kolodner et al., (1985), and Lehnert (1981) mainly for the representation of complex patterns of interaction. A closely related study is the one of Teller & Dahl (1981). They also use a frame concept to describe situations in psychotherapy transcripts. In their paper, similarity between situations is defined in terms of the rather technical notion of "frame overlap". In contrast, the concept of a TOP as a specific frame type provides similarity relations on the basis of a general theory of cognitive processes.

REFERENCES

Dyer, M.G.: *In-depth understanding*. Cambridge, Mass.: MIT Press, 1983 (originally published, 1982)

Kolodner, J.L., Simpson, R.L. Jr., Sycara-Cyranski, K.: A process model of case-based reasoning in problem solving. In: *Proceedings of the 9th International Joint Conference on Artificial Intelligence,* 1985, 284-290.

Lehnert, W.G.: Plot units and narrative summarization. *Cognitive Science*, 1981, *5*, 293-331.

Menninger, K.: *Theory of psychoanalytic technique*. New York: Basic Books, 1958.

Pfeifer, R., & Leuzinger-Bohleber, M.: Applications of cognitive science methods to psychoanalysis: A case study and some theory. *International Review of Psycho-Analysis,* 1986, *13,* 221-240.

Schank, R.C.: *Dynamic memory*. Cambridge, UK: Cambridge University Press, 1982.

Schank, R.C., & Abelson, R.P.: *Scripts, plans, goals, and understanding*. Hillsdale, N.J.: Erlbaum, 1977.

Seifert, C.M., McKoon, G., Abelson, R.P., & Ratcliff, R.: Memory connections between thematically similar episodes. Yale University Cognitive Science Technical Report #27, 1984.

Teller, V., & Dahl, H.: The framework for a model of psychoanalytic inference. In: *Proceedings of the 7th International Joint Conference on Artificial intelligence,* 1981, 394-400.

Wilensky, R.: *Planning and understanding*. Reading, Mass.: Addison-Wesley, 1983.

Encoding Planning Knowledge for Recognition, Construction, and Learning*

Charles Dolan
Hughes Aircraft AI Center
and
Artificial Intelligence Lab
Computer Science Department
UCLA

Michael Dyer
Artificial Intelligence Lab
Computer Science Department
UCLA

Abstract

This paper discusses a method for representing thematic level structures, i.e. abstract plan/goal combinations. We make the case that the processes for both recognition and construction of plans use the same memory structures. In particular, we are looking at the knowledge structures for recognizing and avoiding bad planning. The learning procedure we describe starts by observing the bad planning behavior of narrative characters and combines old descriptions of planning errors to create new abstract structures. The learning method discussed is a one-trial, schema acquisition method, which is similar to DeJong's [DeJong 1983]. The method used involves taking schemas for planning situations that are found in an actual narrative situation, and using causal reasoning to construct a new schema which better characterizes the situation. This work is part of the MORRIS project at UCLA [Dyer 1983a]. The planning situations are represented using Thematic Abstraction Units (TAUs) [Dyer 1983b].

1 Introduction

In the real world, tasks cannot always be accomplished by using simple subgoal partitioning and recursive problem analysis. Both real world agents and narrative characters often must apply plans that require cooperation from other agents, adjust plans that conflict with an agent's concurrent goals, and manage plans that contribute simultaneously to more than one goal. A classification of real world plans is found in [Schank and Abelson 1977], and a taxonomy of goal/plan interactions can be found in [Wilensky 1978].

There are two reasons for looking at poor planning behavior. (1) Knowledge structures that encode poor planning can be used to critique plans and point-out possible weak

* This work was supported in part by grants from the Hughes Aircraft AI Center, the UCLA Center for Experimental Computer Science, and the Keck Foundation.

points. In this way they are similar in function to planning critics and meta-plans [Sussman 1973, Sacerdoti 1977, and Wilensky 1983]; however, they are much more varied than the fixed number of very general critics or meta-plans other researchers propose, and as this paper shows, they are learnable. (2) It is important to represent planning errors, not only to critique plans, but also so that in counter-planning situations [Carbonell, 1979], a planner can try to trick another agent into making a planning error. Any situation which is bad for a planner is generally a good situation into which to force an opponent.

Dyer [1983, 1981] showed how a class of planning errors could be represented by Thematic Abstraction Units (TAUs), and how these planning errors might be recognized in stories. This paper will present a representation for planning errors that also facilitates the combination of planning descriptions and use of thematic structures in planning.

The combination method requires an example narrative situation that contains a new planning error. The input example is conceptually analyzed to discover whether known planning errors are present in the example. In addition to the representations for TAUs, goals, and plans mentioned above, the examples here also rely on Schank's Conceptual Dependency theory [Schank 1972] for the output of the conceptual analysis. Planning errors in the story are examined to find out if they can be combined into new structures, and to establish the links between the component structures that make up the newly discovered planning error.

TAUs are used in planning by treating an agent's plans as if they were a story, and using the TAU recognition process to point out potential planning errors. Specific TAUs are indexed at memory nodes for plans, goals, and actions which are likely to be contained in those TAU.s The memory organization that supports this indexing is a dynamic memory similar to [Schank 1982], [Kolodner 1980], and [Lebowitz 1980]. Depending on the planning error detected, a particular set of heuristics is applied to try and fix the plan.

2 An Example Planning Situation

Here we will see a planning situation that contains three planning errors whose abstract descriptions are already known to the system. From this situation we will see how we can generate two specializations of planning errors, and one novel planning error construct.

The Fox and the Crow

> The Crow was sitting in the tree with a piece of cheese in her mouth. The Fox walked up to the bottom of the tree and said to the Crow, "Crow, what a beautiful voice you have; please sing for me." The Crow was very flattered and began to sing. When she did, the cheese dropped out of her mouth. The Fox grabbed the cheese and ran away laughing.

Note that this story can be looked at in two ways: (1) as an instance of bad planning on the part of the Crow and (2) as an instance of good planning on the part of the Fox. Here we will show how to recognize a planning error from its representation, how to learn new planning errors, and how to apply planning information to avoid goal failures. TAUs can also be turned around and used to find situations where other agents will make mistakes.

The first of the three planning errors we will discuss is the most basic. When the Crow sings, she does not realize that she is already using her mouth to hold the cheese. This planning error is characterized at an abstract level by TAU-CONF-ENABL (confused enablement). The full representation for TAU-CONF-ENABL is given in Figure 1 below.

Our representation of TAUs always defines them from the point of view of the poor planner. The slots 'mistake', 'consequence', and 'failed-goal' of the TAU memory structure provide pointers into a large conceptual graph of **PLAN/GOAL** interactions, which defines the meaning of the TAU.

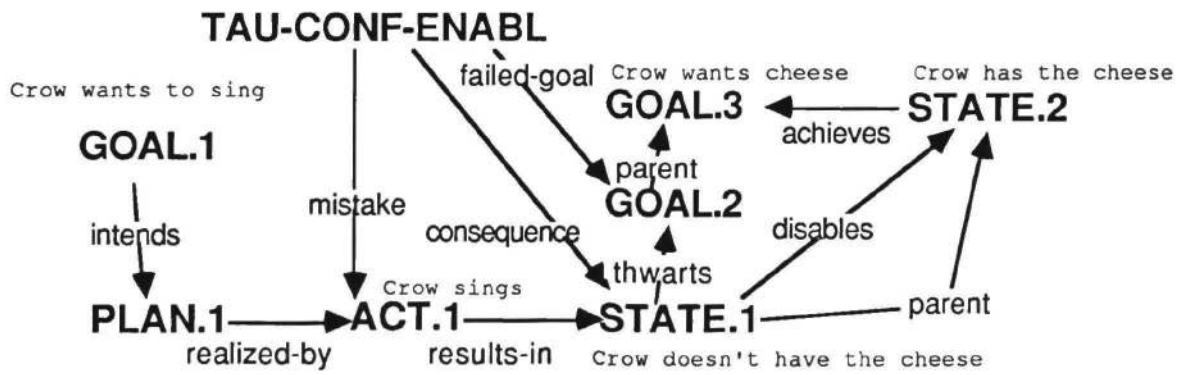


Figure 1

The 'parent' links indicate conceptual structures which are derived from other structures over time. Hence, **GOAL.2** is the same as **GOAL.3** except that its status is 'FAILED'. The rest of the links are derived from Dyer's I-links [1983a].

The abstract situation this structure characterizes is one where an agent has a goal, **GOAL.3**, which has failed, and where the goal was to preserve possession of some object. The cause of the goal failure is an act, **ACT.1**, which attempted to accomplish another goal, **GOAL.1**.

The processes of recognizing and indexing TAUs are covered more fully in [Dyer 1983] and [Dolan 1984]. Dolan [1984] gives a detailed explanation of the comprehension process and memory model that allows the recognition of TAUs in this format.

*The situation is taken from Aesop's fables; the version below is taken from Bewick's collection [Bewick 1973].

As we mentioned above, **The Fox and the Crow** instantiates two other TAUs: (1) TAU-VANITY is the planning error of allowing personal vanity to dictate plan choice; (2) TAU-ULTERIOR is the planning error of not considering another agent's possible motives before acting.

These TAUs can be combined to form new planning heuristics. There are two key problems in TAU acquisition:

1. How does a program know which TAUs to select and examine for combination attempts?

2. Once selected, how are TAUs actually combined to form new planning and indexing structures?

Both 1. and 2. are non-trivial. A sophisticated planner will have numerous stories indexed by multiple TAUs in memory. Attempting to combine TAUs arbitrarily would lead to combinatoric problems. Fortunately, memorable stories (such as Aesop's fables) are designed to give novel planning advice through illustrating planning errors. Thus, TAU selection can be governed by the following strategy:

```
+-----------------------------------------------+
| WHENEVER two TAUs share concepts in an observed |
| planning situation,                            |
| TRY to combine them to form a novel planning   |
| construct                                      |
+-----------------------------------------------+
```

This heuristic can only be applied *after* reading a story and thus the heuristic serves as a form of learning by example. The *comprehension of the story* thus dirrects the learning algorithm to the planning errors to combines and indicates which concepts are shared.

There are two ways to combine TAUs based on the way they share concepts:

(1) specialization,
(2) combination (chunking).

Recent work in specialization learning includes [DeJong 1983], [Lebowitz 1980], and [Kolodner 1980]. All three researchers formulate methods for creating new planning knowledge through specialization, but do not have a method extensible to chunking. Most research in learning by chunking has been in domains where there is no counterplanning [Laird 1984, Rosenbloom et al 1986].

A detailed discussion of the learning process for TAUs can be found in [Dolan and Dyer 1985]. Here we will simply outline the procedure and highlight the features of the representation that facilitate learning.

3 Creating New TAUs through Specialization

TAU-ULTERIOR represents the situation in which ACTOR1 tells ACTOR2 information that motivates ACTOR2 to perform an act which results in the disablement of one of ACTOR2's goals, while at the same time achieving a goal of ACTOR1's which ACTOR2 did not forsee. In the representation for TAU-ULTERIOR in Figure 2 below, outlined concepts are those which have ACTOR1 in the 'actor' slot.

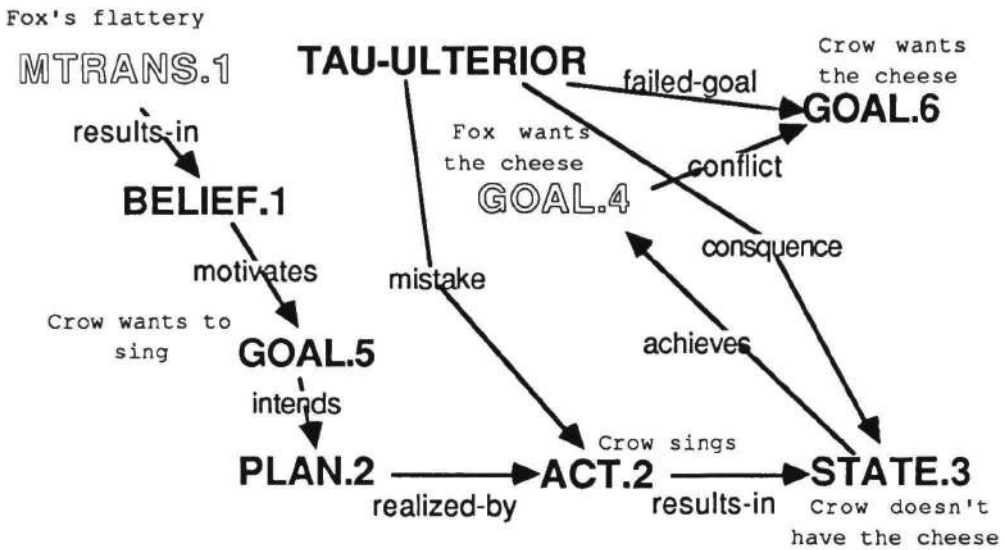


Figure 2

From the description of TAU-ULTERIOR we can see that in some sense it contains TAU-CONF-ENABL because all of the distinct concepts, those not linked by 'parent' links from TAU-CONF-ENABL, are present in TAU-ULTERIOR. The only concepts that TAU-ULTERIOR has that TAU-CONF-ENABLE does not are MTRANS.1, **BELIEF.1**, and GOAL.4. TAU-CONF-ENABL brings two relationships to this specialization: the realationship between **GOAL.2** and **GOAL.3** is that a goal which was achieved in the past has now failed; the relationship between **STATE.2** and **STATE.1** is that a current state disables a previous state. These relationships denote time dependencies not present in TAU-ULTERIOR.

All of the distinct components from TAU-CONF-ENABL are shared with TAU-ULTERIOR, but TAU-CONF-ENABL has additional constraints. We say that TAU-ULTERIOR contains TAU-CONF-ENABL. The containment relation allows us to create a new TAU, a specialization of TAU-ULTERIOR. The new TAU is formed by taking the extra constraints from TAU-CONF-ENABL ("extra" meaning those not already in TAU-ULTERIOR) and conjoining them with those in TAU-ULTERIOR. This TAU, which we will call TAU-MISDIRECTED-ENABL is used to characterize situations where ACTOR1 says something to ACTOR2 to deflect ACTOR2's attention away from an enablement condition on a current goal. TAU-MISDIRECTED-ENABL, shown in

Figure 3 below, is planning advice specific to possession goals and ulterior motives.

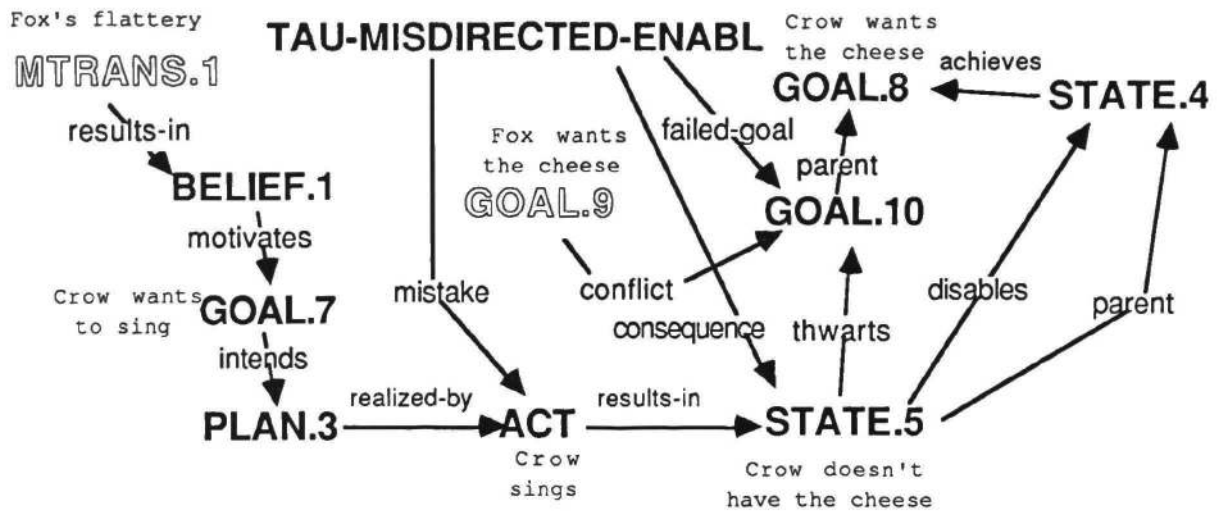


Figure 3

This TAU is also used to process stories such as the following one,

Joey's Waffle

> Joey was sitting at the breakfast table about to eat his waffle. His sister Mary said, "Look Joey, Halley's comet!". When Joey looked up Mary grabbed the waffle and started eating it.

Here the goal that failed for Joey was to maintain possession of the waffle. The enablement condition on that goal was to attend his eyes to the waffle and guard it. The condition was removed when he looked up.

In some sense, when one TAU contains another as with TAU-ULTERIOR and TAU-CONF-ENABL, the contained TAU mediates the application of the containing TAU. An example of how another TAU can mediate for TAU-ULTERIOR is,

Bogus Job Advice

> Fred was a post-doc working at the college. He told his boss, Henry, that there was a faculty position open at the university across town. Henry decided to give his notice at the college and apply for the job. When Henry left, Fred took his position at the college. As it turned out, Henry was not quite right for the new position and didn't get the job.

In this case, TAU-ULTERIOR is mediated by the TAU that encodes the adage, "Look before you leap." The way the two other TAUs mediate TAU-ULTERIOR is show in Figure 4 below. We see that this organization of TAUs points out the differences and similarities of the two stories on a thematic level.

4 Creating New TAUs through Combination

As we saw above, we can get a non-trivial specialization of a TAU by discovering containment in a particular situation. In general TAU combination, however, neither TAU contains the other. In these cases we must examine the relationships or constraints among the concepts *not* shared between the two TAUs. For an example, consider TAU-VANITY. This TAU represents a situation in which ACTOR2 believes he has a special skill and is thus motivated to have a goal (of "showing off" in the Fox and Crow story) that will interfere with pre-existing goals.

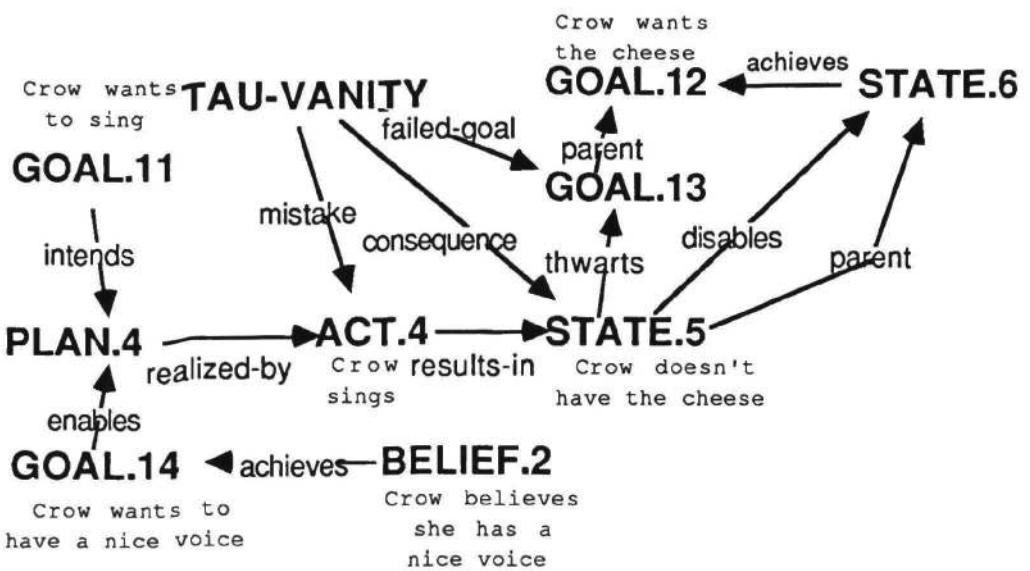


Figure 4

TAU-ULTERIOR and TAU-VANITY, as they are instantiated by **The Fox and the Crow,** share a number of concepts. The concepts that they do not share are MTRANS.1 and GOAL.4 from TAU-ULTERIOR and **GOAL.14** from TAU-VANITY.

The key to combining TAUs is finding the relationship between concepts that are not common to two TAUs being combined. Although **BELIEF.1** and **BELIEF.2** are instantiated by the same concept, they do not belong to identical parts of the graph and this fact provides the point at which we attach the two TAUs. The TAU that we get by combining TAU-VANITY and TAU-ULTERIOR is TAU-SUCKERED. A detailed example of this process is given in [Dolan and Dyer, 1985]. This TAU embodies the planning failure of allowing someone else to take advantage of one's dormant goals by providing one of the missing enablement conditions on that goal. In the case of **The Fox and the Crow** the dormant goal is the Crow's goal to show off. The missing enablement condition is a receptive audience. The Fox provides that audience and so tricks the Crow into defeating her standing goal of keeping the cheese.

TAU-SUCKERED is a new *counterplanning technique* which can be used by a planning program in situations where the appropriate constraints have been met. The

representation for TAU-SUCKERED is shown in Figure 5.



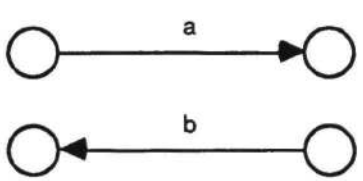
Figure 5

It is intuitively obvious to people TAU-SUCKERED represents the proper lesson to be learned by a reader who knows about vanity and physical enablements, but who has never seen this kind of trickery before.

5 Summary of TAU Learning

In both cases of TAU learning we have examined here, the representation played a central role in guiding the learning. In specialization, the presence of the 'parent' links allowed us to add extra constraints onto TAU-ULTERIOR. In the case of combination, the mapping between abstract concepts in the graph and concepts in the story showed us where to connect the two TAUs together.

The problem of learning a new TAU from two old ones, either by specialization or by combination, can be thought of as a graph union problem. The elements of the 'binding-spec' are nodes and the 'constraints' are arcs. The union of the graphs represents the intersection of the situations that conform to the constraints. In matching abstract graphs, however, we have a problem. If we want to take the union of two graphs below:



We have to make sure that the constraints 'a' and 'b' don't conflict with each other, making their union an inconsistent graph. That problem is solved here because we

have a situation, found in the story, which has all the constraints among the components of the story. Therefore we know that the new structure is a *possible* combination. This is why *learning from examples* is important because *the examples provide a model of what is actually possible.*

6 Using TAUs in Planning

Using TAUs in planning requires two phases: (1) discovering a potential planning error, and (2) taking corrective action. The first phase is accomplished in the same manner as TAU recognition in the task of comprehending stories. As the planner examines the state of the world and formulates various plans, the concepts are "played back" as if they were occuring in a story. When a TAU trigger of TAU-SUCKERED, such as MTRANS.3, the use of flattery, is encountered we have a potential planning error. Possible outcomes of the situation are examined to see if they meet the constraints of the TAU.

In these cases, where some outcome of the current situation might result in a goal failure, corrective action needs to be determined. In informal protocols taken from **The Fox and the Crow**, people seem to think that the Crow could have done three things: (1) put the cheese down and sang, (2) eaten the cheese and sang, or (3) refused to sing. Along with the structure present in TAU-CONF-ENABL, three general heuristics yield these two solutions. The first is,

```
+-----------------------------------------------+
| IF a new goal might cause a standing goal to fail |
| TRY removing a possible resource contention       |
| between plans for the two goals                   |
+-----------------------------------------------+
```

The standing goal is to maintain control over the chees. The new goal is the goal to sing. The common resource is the Crow's mouth. Since there is no way to achieve the new goal without using the Crow's mouth, the system would choose to find another way to support the cheese.

The second solution, that of eating the cheese on the spot, is actually an instance of removing the need for the standing goal. In that case the guiding heuristic is,

```
+-----------------------------------------------+
| IF a new goal might cause a standing goal to fail |
| TRY relieving the need for the standing goal      |
+-----------------------------------------------+
```

Here the system would see that the support for having the goal of controlling the cheese was to satisfy a hunger goal. If the hunger goal is satisfied, then the need to control the cheese goes away.

The function of TAUs such as TAU-SUCKERED is to point out goals that need to be

dropped. The goal conflict shown in TAU-SUCKERED points to the goal we need to examine. The heuristic,

```
+-----------------------------------------------+
| IF a potential goal conflict is found         |
| LOOK for a current plan which would thwart the |
| conflicting goal and reconsider the need support|
| for the current plan                          |
+-----------------------------------------------+
```

directs the program to look at 'thwarting' constraints for the ACT that could defeat the goal. On the other hand, the heuristic,

```
+-------------------------------------------------+
| IF a potential goal conflict is found           |
| LOOK for a current plan that would achieve the  |
| conflicting goal for the other actor, and reconsider |
| the support for the current plan                |
+-------------------------------------------------+
```

directs the program to look at 'achievement' constraints for the ACT which could defeat the goal. In the case of TAU-SUCKERED, both these heuristics give the same result.

Our hypothesis is that the number of these heuristics is small, but that a program with a large number of TAUs could correct most of its planning errors using a few heuristics, the knowledge contained in TAUs, and well chosen stories as input.

6 Progress and Future Work

The CRAM is under development as part of this research. Currently, CRAM is able to understand stories that are parsed by a phrasal parser using a conceptual memory implemented in a frame representation system. CRAM finds the planning errors in each story and characterizes them in terms of one or more TAUs. These TAUs are then used to index the story in memory for later retrieval. CRAM is also able to create new planning error descriptions based on the methods outlined in this paper.

The next step in the on-going development of the theory presented here is to give CRAM enough real-world knowledge so that it can understand a large number of stories, and to observe its performance as it learns more planning errors.

In addition, CRAM will be able to give advice to correct the character's planning errors.

7 Conclusions

The approach presented here allows both specialization-based learning and chunking-based learning of planning errors in multiple planning agent domains. The structures learned can be used both for critiquing plans and also for generating counterplanning advice.

References

Bewick, T., Illustrator, *Treasury of Aesop's Fables*, Avenel Books, New York, 1973.

Carbonell, J. G., *Subjective Understanding: a Computer Model of Belief Systems*, Yale Ph.D Dissertation, Technical Report 150, 1979.

DeJong, G., "Acquiring Schema Through Understanding and Generalizing Plans", in *Proceedings of the Eighth International Joint Conference on Aritifical Intelligence*, 1983.

Dolan, C. P., *Memory Based Processing for Cross Contextual Reminding: Reminding and Analogy Using Thematic Structures*, UCLA Computer Science Department, M.S. Thesis, 1984.

Dolan, C. P. and Dyer, M. G., "Learning Planning Heuristics through Observation", in *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, 1985.

Dyer, M. G., *In-Depth Understanding: A Computer Model of Integrated Processing for Narrative Comprehension*, The MIT Press, Cambridge, Mass., 1983.

Dyer, M. G., "Understanding Stories through Morals and Remindings", in *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, 1983.

Dyer, M. G., "The Role of TAUs in Narrative Processing", in *The Proceedings of the Third Annual Meeting of the Cognitive Science Society*, 1981.

Franklin, B., *Poor Richard: The Almanacks for the Years 1733-1758*, Heritage Press, New York, 1964.

Kolodner, J. L., *Retrieval and Organization Strategies in Conceptual Memory: A Computer Model*, Technical Report 187, Yale University, Department of Computer Science, Ph.D. Disseration, 1980.

Laird, J. E., Rosenbloom, P. S., and Newell, A., "Towards Chunking as a General Learning Mechanism", in *Proceeding of the National Conference on Artificial Intelligence*, 1984.

Lebowitz, M., *Generalization and Memory in an Integrated Understanding System*, Technical Report 186, Yale University, Department of Computer Science, Ph.D. Dissertation, 1980.

Meehan, J., *The Metanovel: Writing Stories by Computer,* Technical Report 74, Yale University, Computer Science Department, Ph.D. Dissertation, 1979.

Rosenbloom, P., et al. "Current Research on Learning in SOAR, in *Machine Learning: A Guide to Current Research*, MItchell, Carbonell, and Michalski (Eds), Kluwer Academic Publishers, 1986.

Sacerdoti, E., *A Structure for Plans and Behavior*, Elssvier North-Holland, 1977.

Schank, R. C., "Conceptual Dependency: A Theory of Natural Language Understanding", *Cognitive Psychology*, Vol 3, No 4, 1972.

Schank, R. C. and Abelson, R. P., *Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1977.

Schank, R. C., *Dynamic Memory: A Theory of Reminding and Learning in Computers and People*, Cambridge University Press, Cambridge, 1982.

Sussman G. J., *A Computational Model of Skill Aquisition*, AI Technical Report 297, AI Laboratory, MIT, Ph.D. Dissertation, 1973.

Wilensky, R., *Understanding Goal-Based Stories*, Technical Report 140, Yale University, Department of Computer Science, 1978, Ph.D. Dissertation.

Wilensky, R., *Plannning and Understanding*, Addison-Wesley, 1983.

The Law as a Learning System[1]

Edwina L. Rissland
Robert T. Collins[2]

Department of Computer and Information Science
University of Massachusetts

ABSTRACT

In this paper, we present three examples of the development of legal concepts and rules and analyze them in terms of AI learning methods, in particular, the candidate elimination algorithm. The three areas of legal doctrine are : (1) the doctrine of the liability of manufacturers to third party buyers; (2) the Fourth Amendment doctrine relating to stop-and-frisk searches; and (3) the doctrine of attractive nuisance. For each of the legal areas, we give a synopsis of the legal developments and show how it can be viewed from an AI point of view.

BACKGROUND

Work on machine learning [Carbonell et al., 1983] has typically concentrated on well-established, well-structured problem areas, like methods of integral calculus. In such domains, concepts and their interrelationships are already known ahead of time to the learning program. Thus the typical concept or rule refinement learning program already knows a concept description language over which it is going to form its generalizations and specializations.

There is nothing wrong with knowing the conceptual landscape before learning new material. However, there are other more fluid situations resting upon less finalized concept spaces in which a lot of interesting learning takes place. This is even true in traditional "neat" domains like mathematics [see Lakatos, 1976]. Furthermore, even in the case of learning in well-established and well-structured domains, the concept spaces are almost never truly static, almost always have some "bias", and are amenable to re-definition and re-structuring [Mitchell, 1983; Utgoff, 1983].

The law provides an example of a domain which is both fluid and "messy". Legal concepts and rules can change, and the concepts are "open-textured," that is, have no clearcut boundaries or hard-and-fast definitions. Sometimes both the doctrine (rules) and the underlying conceptual spaces are in flux at the same time.

In the sense that the legal system as a whole responds to new training instances (i.e., cases) and changes the performance of its adjudication task, we will think of the law as a learning system. The most obvious repository of its knowledge is its case base, which is published and accessible to all. In this paper we concentrate on how the law learns new rules and concepts as evidenced through discussion in published opinions.

[1]This work was supported in part by the Advanced Research Projects Agency of the Department of Defense, monitored by the Office of Naval Research under contract no. N00014-84-K-0017.

[2]Copyright ©1986. Edwina L. Rissland & Robert T. Collins. All rights reserved

The learning algorithm which we use to analyze the evolution of legal knowledge is Mitchell's candidate elimination algorithm (CEA) [Mitchell, 1983; *The Handbook of Artificial Intelligence, Vol 3*]. In essence the algorithm works to narrow the definition of a concept or range of applicability of a rule by squeezing in from very general and very specific instances. Positive instances cause CEA to generalize and negative ones cause CEA to specialize.

The CEA algorithm works by maintaining a "version space" of all descriptions which are consistent with the examples seen so far. The version space is represented by maintaining a maximally specific set S and a maximally general set G of concept descriptions. All other descriptions in the version space are implicitly contained within these two bounds. To refine the version space, CEA examines positive and negative "training instances". Given a positive instance of the concept, the set S is made more general to include descriptions matching it; set G is also made more specific to exclude descriptions covering negative instances. These positive and negative refinements push S and G closer together until they converge on the correct description.

The idea of describing the behavior of the legal system as a learning system using a CEA-like algorithm can be seen in an analysis by the legal scholar Max Radin :

> "The single instance is capable of generalization, and the generalization will not stop at any particular place, unless by negative decision, by a statement that a given situation is outside the genus, a subsequent court has deliberately attempted to stop it. Then the process begins all over again, because the excluded situation is itself capable of successive generalizations and we must know whether a large or small genus is to be excluded from (A). A Buick car is not in (A). Is an automobile truck? An electric machine? A hypothetical new type of car driven by more explosive mixtures than gasoline, and so on?"
>
> – Radin, 1933, at p. 209

One difference between application of CEA to the legal domain and the more usual ones is that in the law, the hierarchy of concepts and the range of a rule usually evolve at the same time.[3] This is so because of the non-static, open-textured nature of the law. In particular, one can change the effect of a rule by changing either (1) the rule itself such as through the addition of conjuncts to the pre-conditions or listing of exceptions to the rule, or (2) changing the meaning of the ingredient concepts. Thus in our analysis we interleave discussion of the evolution of a legal rule and its group of ingredient concepts. From an implementation standpoint, that means we have two CEA processes running cooperatively. We illustrate the development of rules and concepts in three example doctrines—products liability, search and seizure, and attractive nuisance. There is nothing special about these three; they are indicative of the growth of legal knowledge in general.

THE INHERENTLY DANGEROUS EXAMPLE
LEARNING A LEGAL CONCEPT

Our first example is a famous one that has been written of several times by legal scholars [Levi, 1949; Radin, 1933]. It concerns the development of the idea that a manufacturer of goods is liable (in

[3]Note: Even though we concentrate on "rules," Anglo-American law is based on cases and not rules. This is so even in statute law, for here the ingredient rules and concepts in the statute must be interpreted in light of specific cases (see [Levi, 1949] for an example).

case	date	item	finding
Thomas v. Winchester	1852	belladonna	liable
Loop v. Litchfield	1870	balance wheel	not liable
Losee v. Clute	1873	steam boiler	not liable
Devlin v. Smith	1882	painter's scaffold	liable
Torgesun v. Schultz	1908	bottle of aerated water	liable
Statler v. Ray Mfg. Co.	1909	coffee urn	liable

Table 1: **Summary of "Inherently Dangerous" Cases**

tort) to third parties injured by a defective product. For instance, if you buy a defective Buick and a front wheel falls off and you are hurt, the manufacturer of the car, Buick Motor Company, must compensate you for your injury even though you bought the car from an intermediate vendor like a car dealership (hence, you are called a "third" party), and not directly from Buick Motor Company. A closely related doctrine, which developed later and in a parallel way is the (contracts) doctrine of "implied warranty" by which is meant that there is an implicit warranty that goes along with a product to the effect that the product is warranted to safely do what it is supposed to do as long as it not abused. These doctrines are at the basis of our current protection of consumers.

This was not always the case. Originally, there was no relationship of liability between manufacturers and third parties : the law required "privity" of contract, that is, an immediate seller-purchaser relationship between manufacturer and injured. This could be summed up by the rule of thumb : *no privity, no liability.*

Our learning episode begins in 1852 with the case *Thomas and Wife v. Winchester* decided in the New York Court of Appeals. In this case, Mr. Thomas had his local druggist fill a prescription for dandelion extract for his ailing wife. The druggist filled the prescription from a jar which he had bought from another druggist who had bought it from the manufacturer, Winchester. Unfortunately, what was actually in the jar was belladonna, a substance looking somewhat similar but causing very different effects. As a result, Mrs. Thomas got quite sick and suffered "derangements of the mind"; she eventually recovered. Mr. and Mrs. Thomas sued Winchester, the manufacturer and packager of the mislabelled poison.

The Thomases won. The court justified making an exception to the "privity" rule because it believed that misbranded poisons like belladonna were so "imminently" or "inherently dangerous" that they deserved an exception. The court likened belladonna to a loaded gun in the hands of a child, an item so "inherently" dangerous that it was sure to cause injury to unsuspecting and innocent victims. Thus this landmark case established an exception to the privity rule. Subsequent cases would establish its scope.

For the period from 1852 until 1916, the New York Court of Appeals handed down decisions (see Table 1) concerning further exceptions to the privity rule. The exception started off with belladonna and ended up including painter's scaffolds, bottles of aerated water and coffee urns. It was not extended to cover items such as carriages or steam boilers which become dangerous only if defective.

In effect, the Court was defining the *inherently dangerous* category in a CEA-like manner by including or excluding examples on a case-by-case basis. That is, if an item allowed recovery for third person injury under the rubric of the inherently dangerous exception, it was in the class—i.e., was a *positive instance*— and if it didn't, it wasn't—a *negative instance*. Positive instances served both to flesh out the concept of inherently dangerous and added pressure to generalize it; negative instances narrowed it.

In this middle period of about fifty years, the Court was also steadfastly refusing to generalize the *inherently dangerous* exception to a more general class like *dangerous if defective*. However, in this period it was simultaneously pushing the boundaries of the *inherently dangerous* class (and thus the exception) to include some items like coffee urns and hairwash which seem to be more of the class of item that are *dangerous only if defective*. Nevertheless, the Court spoke as if it wasn't and summed up its position on the privity rule as follows :

> "One who manufactures articles inherently dangerous, e.g., poisons, dynamite, gunpowder, torpedoes, bottles of aerated water under pressure, is liable in tort to third parties.... On the other hand, one who manufactures articles dangerous only if defectively made, or installed, e.g., tables, chairs, pictures or mirrors hung on the walls, carriages, automobiles, and so on is not liable to third parties for injuries caused by them, except in cases of willful injury or fraud."
>
> – *Cadillac v. Johnson*, 221 F. 801 (1915), at p. 803

This case represents the last gasp of a court trying to hold onto its "old" interpretation of the concepts and the scope of the exception. Note it explicitly rules out carriages and automobiles. In the very next year these very things were let in.

In 1916, the New York Court of Appeals decided the landmark case of *MacPherson v. Buick* 217 N.Y. 382. In this case, MacPherson, a third party, was allowed to recover for injury caused by a defective Buick, an item *dangerous only if defective*. This is exactly the case that the Court had persistently (e.g., in *Cadillac* case above) said it wasn't going to let the exception cover. The case of a car or carriage had now flip-flopped from being a negative to being a positive exemplar for the privity exception. As a positive exemplar, it greatly expanded the exception. *MacPherson* also ended the unreasonable stretching of the *inherently dangerous* class and the need to talk fictions about being inherently dangerous to win recovery for an item dangerous only if defective.

As Judge (later to become Supreme Court Justice) Cardozo stated in this landmark opinion :

> "We hold then, that the principle of Thomas v. Winchester is not limited to poisons, explosives, and things of like nature, to things which in their normal operation are implements of destruction. If the nature of a thing is such that it is reasonably certain to place life and limb in peril when negligently made, it is then a thing of danger. Its nature gives warning of the consequences to be expected. If to the element of danger there is added knowledge that the thing will be used by persons other than the purchaser, and used without new tests, then, irrespective of contract, the manufacturer of this thing of danger is under duty to make it carefully. That is as far as we are required to go in this case. There must be knowledge of a danger, not merely possible, but probable. It is *possible* to use almost anything in a way that will make it dangerous

> if defective ... where danger is to be foreseen, a liability will follow ... We have put aside the notion that the duty to safeguard life and limb, when the consequences of negligence may be foreseen, grows out of contract and nothing else. We have put the source of the obligation were it ought to be. We have put its source in law."

This opinion places automobiles within the class of ***dangerous only if defective*** and grants recovery for injury by them. Thus the exception to the privity rule has been generalized. ***MacPherson*** also creates a somewhat finer concept hierarchy : it speaks of the class ***thing of danger***, which includes both things ***inherently dangerous*** and things ***foreseeably dangerous if defective***. See Figure 1.

Thus, after the episode beginning with *Thomas v. Winchester* decided in 1852 and ending with *MacPherson v. Buick* decided in 1916, the New York Court of Appeals has learned new concepts (e.g., ***inherently dangerous, dangerous only if defective***), placed them into a hierarchy of concepts, articulated an exception to a well-established rule, and re-classified a training instance for the rule from negative to positive (i.e., the car in *Cadillac* and that in *MacPherson*). Said another way, the law has learned the new ***liability to third party*** rule and its underlying concepts.

THE STOP AND FRISK RULE

Another example of the law elaborating an exception to a rule, and the concepts underlying it, is the development of the stop-and-frisk doctrine. This area of the law involves Constitutional issues surrounding the Fourth Amendment which states :

> The right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures, shall not be violated, and no Warrants shall issue, but upon probable cause, supported by Oath or affirmation, and particularly describing the place to be searched, and the persons or things to be seized.

The Amendment states that in order to search and seize a person or property (NB an arrest is a seizure of person), the police must have "probable cause" and a warrant.

Prior to the first stop-and-frisk case which came before the Supreme Court in 1968, there already were well-established exceptions to the necessity of a warrant for search and seizure but none to the necessity of probable cause. Examples of situations where police can perform a warrantless search are : (1) if a person consents, (2) if the search is incident to a person's arrest, and (3) if the police feel that evidence might be destroyed if the search is postponed. An example of a constitutionally approved warrantless seizure is the arrest of a suspect running away from the scene of a crime.

Thus, until this case, the requirements for a search were (1) probable cause, and (2) a search warrant or one of the recognized exceptions. The way that police are encouraged to behave according to these Constitutional standards is through the "exclusionary rule" which excludes illegally obtained evidence from being included at trial, or if admitted at trial, it permits a guilty verdict to be overturned.

The constitutionality of stop-and-frisk episodes, which we have all grown accustomed to seeing in TV renditions of police work, was laid down in the case of *Terry v. Ohio* 392 US 1 (1968). In this case, a police officer named McFadden observed three men, one of whom was Terry, apparently

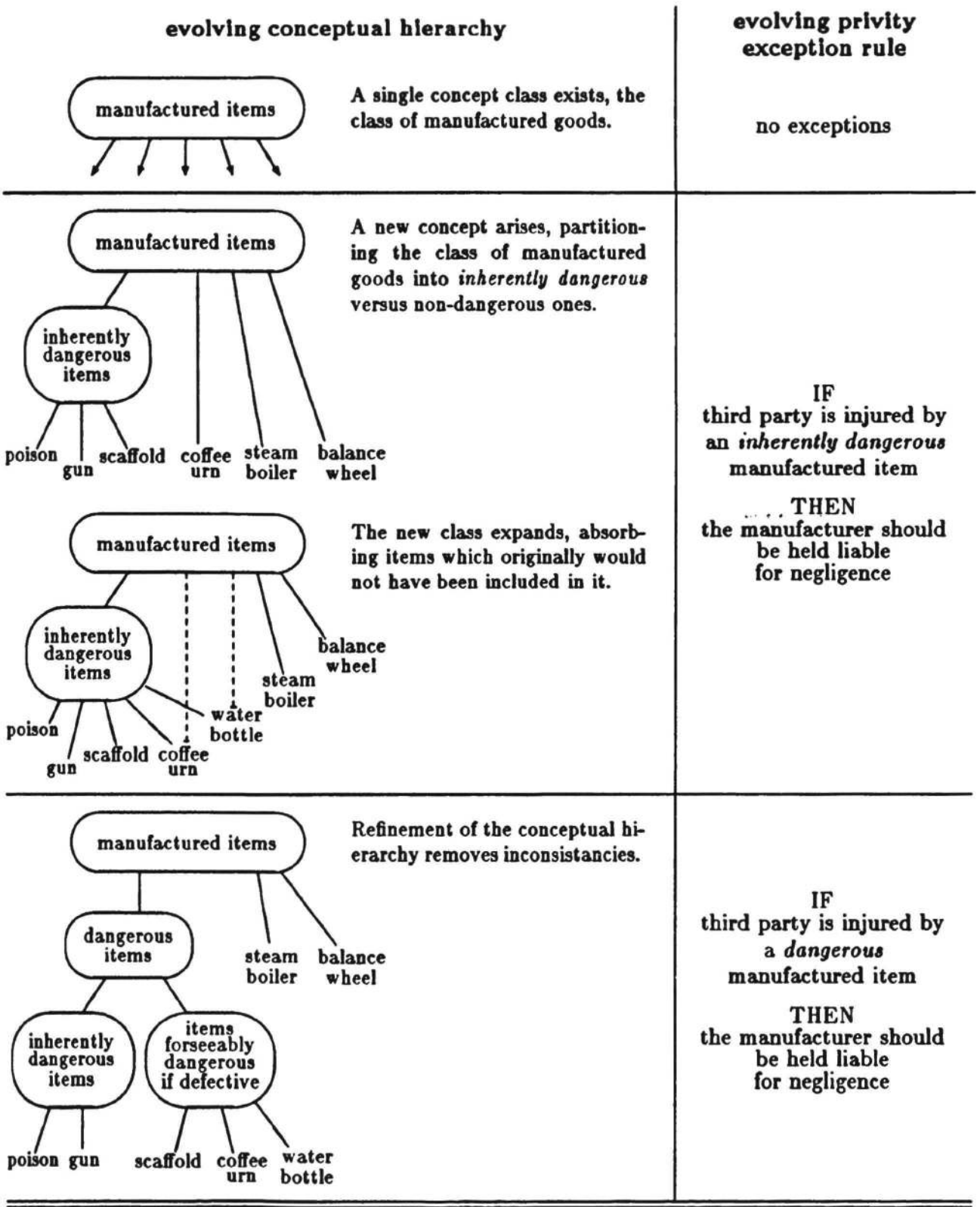


Figure 1: **The "Inherently Dangerous" Example**

"casing" a store in broad daylight in downtown Cleveland. McFadden's suspicions were aroused when Terry and his friends studied and walked past a store and then conferred with each other some two dozen times. McFadden suspected they were planning a robbery and since it was daytime, he feared that it would be an armed robbery and that the three men were armed. McFadden accosted and stopped them. When Terry mumbled something unintelligible to McFadden's inquiries, he grabbed Terry and patted down the outside of his clothing, felt a pistol, and subsequently removed it. He did the same to the second man, with the same result of finding a loaded weapon. The third man was "clean". Terry and his co-defendant were convicted of carrying a concealed weapon. They appealed, and the Supreme Court held it was an unreasonable search and seizure.

In the Supreme Court opinion, Chief Justice Warren immediately dismisses the prosecution's argument that Officer McFadden had probable cause (and thus its conclusion that the search and seizure was constitutionally correct according to the doctrine as it now stood). What he had was something less. Thus the constitutional question was whether one can perform a stop-and-frisk (a kind of arrest and search) without probable cause and under what conditions.

The Court said :

> "We merely hold today that where a police officer observes unusual conduct which leads him reasonably to conclude in the light of his experience that criminal activity may be afoot and that the persons with whom he is dealing may be armed and presently dangerous, where in the course of investigating this behavior he identifies himself as a policeman and makes reasonable inquiries, and where nothing in the initial stages of the encounter serves to dispel his reasonable fear for his own or others' safety, he is entitled for the protection of himself and others in the area to conduct a carefully limited search of the outer clothing of such persons in an attempt to discover weapons which might be used to assault him."
>
> – *Terry v. Ohio*, 392 U.S. 1 (1968), at p. 443

In other words, to reach the conclusion that *P can STOP and FRISK X*, five preconditions must be satisfied :

1. there exists *reasonable suspicion* of X's criminal activity;
2. there exists *reasonable belief* that X is armed and *dangerous*;
3. P *identifies* himself as a police officer;
4. P makes *reasonable inquiries*; and
5. there does not exist *dispelling evidence*.

This landmark decision for the first time recognized an exception to the requirement that Fourth Amendment searches and seizures must be based on probable cause.

What followed *Terry* were cases fleshing out the scope of the stop-and-frisk rule by discussing its ingredient concepts and the necessity of its preconditions. For instance, four years later in *Adams v. Williams* (1972), the Court allowed the stop-and-frisk exception to cover the somewhat dubious instance of an officer reaching through the driver's window to reach into the waistband of the driver and seize a gun—not exactly a classic example of stop-and-frisk. In *Adams*, the Court also seems to be trying to stretch the concept of *reasonable suspicion* in order to invoke the rule of *Terry*. In *Terry* the suspicion was based on the officer's own observations and his own experience from more than 30 years of detective work; In *Adams*, it is not based on the officer's own observation but on

a tip from an informant who had previously furnished unreliable information. In any case, *Adams* makes quite clear that the conditions necessary for a *Terry* stop are indeed less than probable cause.

Stretching the *reasonable suspicion* concept raises the specter of the stop-and-frisk exception swallowing up the established rules (especially vis-a-vis probable cause) of the Fourth Amendment. This is addressed by Justice Brennan in his dissent where he quotes the lower Appeals Court :

> "Terry v. Ohio was intended to free a police officer from the rigidity of a rule that would prevent his doing anything to a man reasonably suspected (our emphasis added) of being about to commit or having just committed a crime of violence, no matter how grave the problem or impelling the need for swift action, unless the officer had what a court would later determine to be probable cause for arrest. It was meant for the imminent danger or of harm recently perpetrated to persons or property, not the conventional ones of possessory offenses. If it is to be extended to the later at all, this should be only where observation by the officer himself or well authenticated information shows "that criminal activity may be afoot." ... I greatly fear that if the contrary view should be followed, *Terry* will have opened the sluicegates for serious and unintended erosion of the Fourth Amendment."
>
> – 407 U.S. 143 (1968), at p. 147

Justice Brennan emphasizes that *reasonable suspicion* is not the only condition necessary for a *Terry* stop and that for a suspicion to be reasonable it must be based on "good" information. He is also emphasizing that the second condition requiring the existence of *danger* is also necessary. It is well to note that much police work involves crimes of violence or possession, especially of drugs, and that the *Terry* exception was meant to handle certain situations growing out of the former and not as an investigatory tool—that is exactly what warrants are for.

In a separate dissent, Justice Marshall, with whom Justice Douglas joined, makes some meta-level comments on the evolution of this body of law as well as laying out his exception to it :

> "Four years have passed since we decided Terry v. Ohio, 392 U.S. 1 (1968), and its companion cases, Sibron v. New York and Peters v. New York, 392 U.S. 40 (1968). They were the first cases in which this Court explicitly recognized the concept of stop-and-frisk and squarely held that police officers may, under appropriate circumstances, stop and frisk persons suspected of criminal activity even though there is less than probable cause. This case marks our first opportunity to give some flesh to the bones of *Terry et al.* Unfortunately, the flesh provided by today's decision cannot possibly be made to fit on *Terry*'s skeletal framework."
>
> – 407 U.S. 143 (1968), at p. 148

To recap, what we have seen in *Terry* and *Adams* (and also of *Sibron* and *Peters* mentioned by Marshall), is the hewing out (in *Terry*) of a sizable exception to the Fourth Amendment requirement of probable cause, and essentially a carving out of an exception to the exception of *Terry* (in *Adams*) despite the dissenters' complaints. That is, after allowing stop-and-frisk in *Terry* when five preconditions are met, the Court then weakens and elevates the first *reasonable suspicion*

condition. So now the antecedent to the action stop-and-frisk rests primarily on the predicate of ***reasonable suspicion*** whose force has been weakened by stretching its meaning.

This second example of a learning episode is interesting to compare with our first. First of all, the stop-and-frisk episode feels like a more compact learning episode where the first did not. Whereas the courts took fifty or so years to spell out the inherently dangerous exception, the stop-and-frisk exception was well-articulated in one case. The emphasis in the inherently dangerous episode was on progressively refining the concept ***inherently dangerous*** and the exception to the privity rule, and the situation only gets well-settled in the final case. The emphasis in the stop-and-frisk episode is on arguing the importance of the predicate ***reasonable suspicion*** and spawning off a discussion much like the inherently dangerous one for the meaning of the concept of ***reasonable suspicion***. The next stage of the stop-and-frisk story, which we don't have space to recount, is to work out what is meant by a "stop", especially the parameters of its duration and place of occurrence. Thus both examples illustrate how a rule and the concepts underlying it evolve jointly. The tension in interpretation—and hence learning—seesaws between formulation of the rule and definition of its ingredient concepts.

THE ATTRACTIVE NUISANCE DOCTRINE
LEARNING THE RULE OF THE LEAD CASE

In our third and last example, we examine an episode from an area of trespass law : the doctrine of attractive nuisance. A longer exposition of this area can be found in [Collins, 1986]. This is another example of the creation of an exception to a well-settled rule. However, instead of focusing on the evolution of the doctrinal rule, we focus on the evolution of the rule of the lead case.

Doctrinal rules are what we have examined in the first two examples. They are rules which summarize the state of the case law as a whole. The rule of a case, on the other hand, refers to what an individual case stands for in the law. The difference is most apparent at the end of a learning episode; initially they may be identical. As example of the difference, the final doctrinal rule for liability of manufacturers to third parties concerns items which are dangerous if defective. The rule of the lead case, *Thomas v. Winchester*, says that only in the case of an item inherently dangerous is there liability.

Attractive nuisance doctrine deals with the liability of a landowner to young children who receive injuries while trespassing. Note this is an exception to the usual situation where a landowner owes no duties to trespassers, except to refrain from intentionally injuring them either in person or through the use of traps. This doctrine evolved through a line of cases in the manner we have already seen in our first two examples. Its lead case is *Sioux City and Pacific Railroad Company v. Stout* 84 U.S. 657 (1873).

What we want to concentrate on in this section is describing how the legal system comes to learn what *Stout* stands for, that is, what the rule of *Stout* is. In its most specific interpretation, the rule of *Stout* is a rule with antecedent conditions identical to the legal facts of the case, and with consequent identical to its legal conclusion. In discussing the rule of a case one does not usually want to state the rule in this most specific form for then it is sure to match almost no other case that comes before a court and thus cannot be applied to these other cases. On the other hand, if one jumps to generalities, especially in the antecedents, one might easily over-generalize. So the

problem becomes just what range of possibilities is to be stood for by the case. This is another CEA-like task. Again, subsequent cases will be the source of positive and negative exemplars.

This phenomenon of learning what a case stands for by examining successor cases was described by Karl Llewellyn :

> "Applying this two-faced doctrine of precedent to your work in a case class you get, it seems to me, some result as this : You read each case from the angle of its maximum value as precedent ... Contrariwise, you will also read each case for its minimum value as precedent, to set against the maximum ... The first question is, how much can this case be fairly made to stand for by a later court to whom the precedent is welcome? ... The second question is, how much is there in this case that cannot be got around, even by a later court that wishes to avoid it?"
>
> – Llewellyn, 1930, at p. 69

Llewellyn is clearly speaking of several things here such as the nature of precedent. He previously spoke of the nature of "loose" versus "strict" precedent; the former is reading a rule broadly to cover a large range of cases, the limiting maximally general situation being coverage of all cases (i.e., a rule with no antecedents), and the latter is reading narrowly to cover a small range of cases, the limiting maximally specific situation being coverage of only the case where the rule originally came from. That is, the rule of precedent can allow for a range of interpretations just as we see in version spaces.

In order to see how this process of interpretation through subsequent cases leads to the rule of a case, we look at how the rule of the *Stout* case evolved. A summary of the facts of *Stout* are :

> Henry Stout, who was a child of six years, was injured on a turntable belonging to the Sioux City and Pacific Railroad Company. The turntable was in an open space near two traveled roads but with few houses in its vicinity; it was about a quarter mile from the railroad's station house and was not enclosed or visibly separated from adjoining properties. Stout went to play on the turntable on the suggestion of two older boys, aged nine and ten, one of whom had previously played there. As the boys began to turn the turntable, Henry's foot was caught between the end of the rail on the turntable as it was revolving, and the end of the iron rail on the main track of the road, and was crushed.
>
> Railroad workers had previously seen boys playing on the turntable and had forbidden them from playing there. This was Stout's first visit to play there. The turntable had a broken latch at the time of the accident, although this latch was easily undone and whose main purpose was not to lock or guard the turntable but to keep it in position while being used.

The Supreme Court upheld the lower court's verdict for the plaintiff. It decided that the plaintiff was not himself negligent, either directly or through his trespass, because his tender age exempted him from the same degree of care required of an adult. Thus, the fact that Henry Stout was six years old is most relevant to the case and the language of the opinion suggests that this generalizes to all children of tender years. Further legal conclusions were : (1) the turntable was dangerous and likely to cause injury to children (evidenced by Stout's injury); (2) the defendant should have anticipated

children would resort to it (evidenced by the fact that even though it was located in a remote place, children had previously played there and had been observed by railroad employees); and (3) the defendant could have taken minor precautions to prevent the accident (e.g., fixed the broken latch, erected a fence).

To use CEA to learn the *rule of Stout*, we need to initialize our version space. See Figure 2. The most specific instance is just the facts and conclusion of *Stout* :

Antecedent : Henry Stout
who was six years old
received a crushed foot from
a dangerous railroad turntable
on SC&PRR property
on which he trespassed
and which SC&PRR could have anticipated
and whose accident was easily preventable

Consequent : SC&PRR is negligent and should have anticipated and prevented this incident.

We also need a most general formulation of the rule of *Stout*. To do this we will simply set all the attributes mentioned in the most specific rule to be the upper or most general numeric or symbolic value. The value hierarchies we use are somewhat arbitrary and are for illustrative purposes only. Thus from subsequent cases, what is to be learned is not only the ingredient conjuncts but also their ranges of values. This sort of top down learning has recently been studied by Fu and Buchanan [1985] in their *RL* program. In particular, all of our subsequent cases will be negative training instances and will serve to specialize the rule.

- Training Instance 1: *Smith v. Hopkins*, 120 F 921 (1902):

Facts: Plaintiff was an 18 year old high school student who was killed while on a picnic with his friends when he walked up on a pile of railroad ties placed on defendant railway's right of way and which collapsed.

Holding: Plaintiff lost. Court reasoned *Stout* did not apply here because plaintiff was 18 and of an understanding mind.

Rule Update: Specialize range of plaintiff's age from *any age* to *young*, the next most specific category excluding an 18 year old.[4]

- Training Instance 2: *Reidel et al. v. West Jersey RR Co.*, 177 F 374 (1910)

Facts: Plaintiff, seven or eight years old, was playing with other children near a railroad track and saw some flowers on the other side of the track. Access to the track was barred with a fence and a bolted gate. The children unbolted the gate and started across. While crossing, plaintiff fell on the "third rail" and was burned by the current.

Holding: For the defendant since, unlike *Stout*, the property was adequately guarded by a fence.

[4]We could also try to specialize *Stout* by adding a conjunct "not including railroad ties". We don't because in its opinion, the court pointed to the difference in the plaintiffs' ages as the basis to distinguish *Stout*.

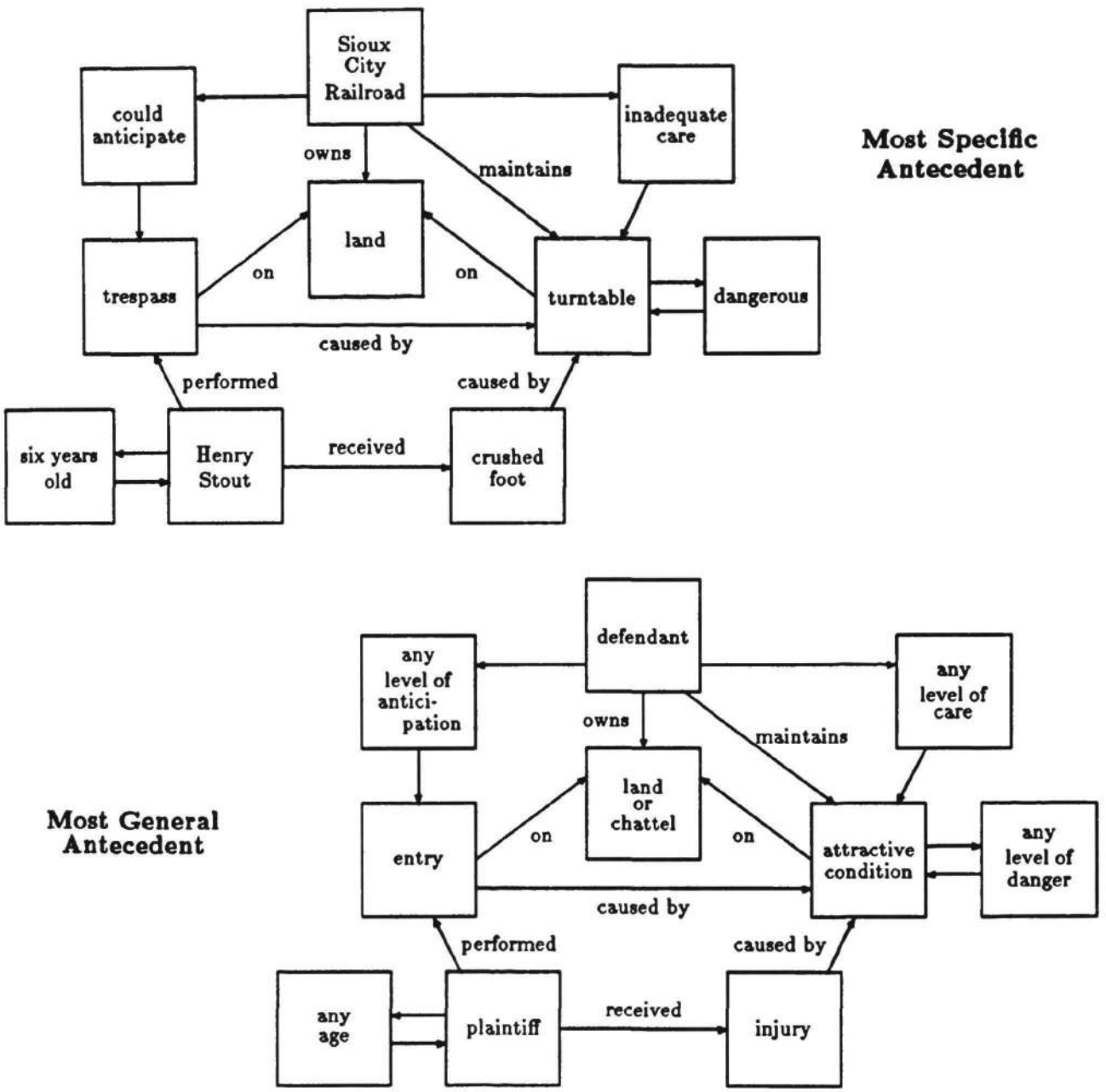


Figure 1: Initial Version Space for Antecedent of *Stout* Rule

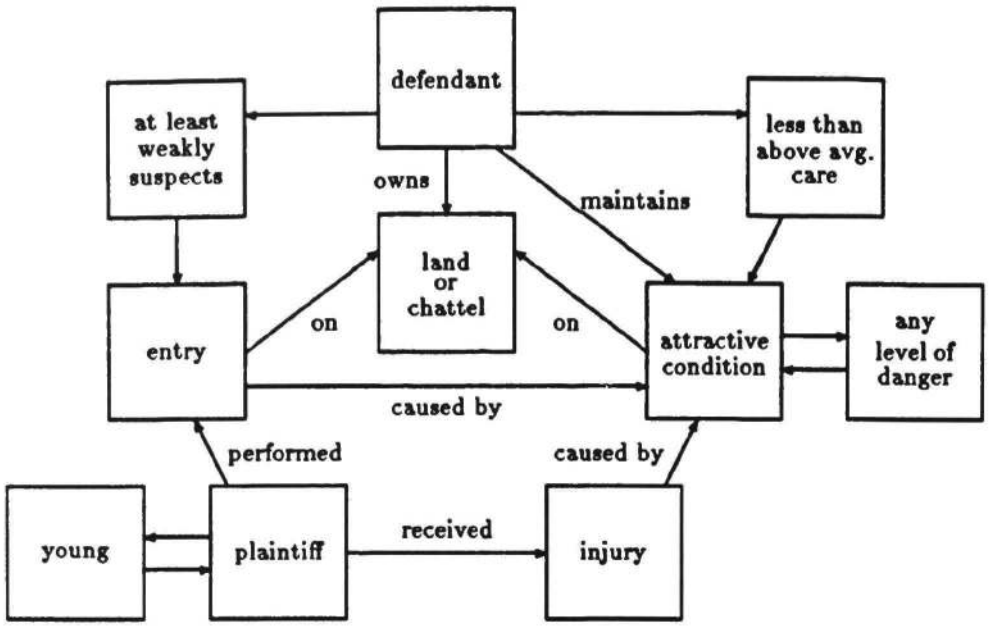


Figure 1: **Most General Final Antecedent of *Stout* Rule**

Update: Specialize *prevention* predicate to exclude above average care.

Training Instance 3: *Heller v. N.Y., N.H. & H.R. Co.*, 265 F 192 (1920).

Facts: Defendant owned and maintained an electrical wire passing close to a bridge abutment that was covered with a screen. The plaintiff, an eleven year old boy, climbed the abutment some twenty feet or more, climbed through a hole in the screen, reached out with a paint can some 29 inches, touched the wire, and was electrocuted.

Holding: For the defendant since the plaintiff's extraordinary act was not foreseeable.

Update: Specialize the range of anticipation predicate to exclude extraordinary unforeseeable situations.

Given these three training instances, the most general interpretation for the rule of *Stout* is now as shown in Figure 3 (the most specific rule is unchanged).

In summary, what we have tried to illustrate with our third example is how the possible interpretation of the rule of a case is refined through subsequent cases. In particular, we have shown how the rule of the case is specialized through subsequent, negative training instances.

CONCLUSION

What we have shown in this paper is how the law can be viewed as a system that learns from experience in a CEA-like manner. We presented three episodes of such learning. The first, the inherently dangerous example, shows how the law learns both a rule (actually an exception to a rule) and the underlying concept hierarchy simultaneously. The second, stop-and-frisk example,

shows how the law refines a rule (again, an exception to a rule) through discussing ingredient preconditions as well as (recursively) learning the concepts used in the rule. The third example shows how future (negative) exemplars are used to learn the rule of a case.

REFERENCES

Carbonell, J.G., Michalski, R.S., & Mitchell, T.M. (1983), "Machine Learning : A Historical and Methodological Analysis," *AI Magazine*, Fall.

Collins, R.T. (1986), "LEX Justice : A Version Space Approach to Legal Reasoning," M.S. Project, Department of Computer and Information Science, University of Massachusetts.

Fu, L., & Buchanan, B.G. (1985), "Learning Intermediate Concepts in Constructing a Hierarchical Knowledge Base," in *Proceedings IJCAI-9*, Los Angeles, California, August.

Lakatos, I. (1976), *Proofs and Refutations*, Cambridge University Press, London.

"Learning and Inductive Inference (Chapter 14)," *The Handbook of Artificial Intelligence, vol. III*, Cohen & Feigenbaum (Eds.), William Kaufmann, Inc. 1982.

Levi, E.H. (1949), *An Introduction to Legal Reasoning*, University of Chicago Press, Chicago.

Mitchell, T.M. (1983), "Learning and Problem Solving," in *Proceedings IJCAI-8*, Karlsruhe, West Germany, August.

Radin, M. (1933), "Case Law and Stare Decisis : Concerning *Prājudizienrecht in Amerika*," *Columbia Law Review*, vol. 33, p. 199.

Rissland, E.L. (1984), "Argument Moves and Hypotheticals," in *Proceedings First Annual Conference on Law and Technology*, University of Houston, Houston, Texas.

Rissland, E.L., & Ashley, K.D. (1984), "Hypotheticals as Heuristic Device," in *Proceedings AAAI-86*, Philadelphia, PA., August.

Utgoff, P.E. (1983), "Adjusting Bias in Concept Learning," in *Proceedings IJCAI-8*, Karlsruhe, West Germany, August.

Solving Puzzles with a Connectionist Network Using a Hill-Climbing Heuristic

Alan H. Kawamoto

Dept. of Psychology
Carnegie-Mellon University
Pittsburgh, PA 15217
(412) 268-3157

ABSTRACT

A connectionist network has been used to simulate the solution, using a hill-climbing heuristic, of the DOG- > CAT puzzle (changing 1 letter at a time, generate a sequence of 3-letter words beginning with *DOG* and ending with *CAT)* and a simpler variant of the 8-tile puzzle, the dog-cat-mouse (DCM) puzzle devised by Klahr (1985). Distributed representations have been used to represent the different possible states of the puzzles. These states are learned by the network and become local energy minima of the system. To simulate the sequence of states corresponding to a solution of the puzzle, the initial state of the network is set to the start state, and the goal state is presented to the network as a continuous input. A sequence of states is generated by habituation, a short-term modification of the connection strengths whenever all the elements in the network are maximally or minimally activated, and by exploiting the property that successive states comprising the solution are similar.

Puzzles and games have been studied extensively because they illustrate the importance of search in issues of problem solving. In the classic 8-puzzle, for example, 8 uniquely numbered tiles fit into a 3-by-3 matrix with one open space. To solve the puzzle, a tile adjacent to the open space is slid into that position, which, in turn, creates a new configuration with the open space now in an adjacent position. This continues until the goal state (a particular configuration of the tiles) is attained.

Although quite a large number of distinct configurations exist (9! = 362,880, excluding rotations and reflections), a few moves are sufficient to get to the goal state from any starting state. From any given configuration, there are 2, 3, or 4 possible successive states depending on whether the open space is at a corner, an edge, or the center, respectively. If moves back to the previous state are not allowed, even fewer possibilities arise. The set of configurations that are possible comprises the *problem space* or *state space*. The different ways of transforming one state to another are known as *operators*.

One approach to solve this puzzle involves the use of *search trees* (or *graphs)*. Each node in the tree represents a particular configuration of the puzzle, and arcs connect possible successive configurations. Beginning at the root of the tree, the start state, the tree can be systematically searched in a breadth-first (exploring all the nodes at one level before going to the next lower level) or depth-first (following a path for a fixed number of levels, then retracing that path and exploring alternative paths if the search fails) manner.

In contrast to the algorithmic solution described above, a simple heuristic can be used instead. One such heuristic, *hill-climbing,* consists of selecting that child of the current node that is closest to the goal state, then selecting the best child of that node while ignoring its siblings and offspring of those siblings. No memory of previous states considered, nor the path that led to the current state need be stored. Such methods are appealing because they are usually

faster (when successful) and seem to mirror at least some of the processes of human problem solvers.

One important characteristic of possible successive states that will be exploited in this study is that changes from one state to another are incremental, with successive states being as similar as they could possibly be without being identical. Note, however, that not all such minimally distinct states are possible successors of each other. In the 8-puzzle, for example, only those involving the movement of a tile from its current position to an adjacent empty position are possible successors, whereas configurations with two tiles switched are not.

In implementations to date, each state has been been represented as a distinct node, with arcs representing legal moves between two states. A distributed representation (where each state is represented as a pattern of activity over a set of features) can, however, more naturally exploit the hill-climbing heuristic. * There has been a lot of discussion of distributed representations (Anderson, Silverstein, Ritz, and Jones, 1977; Hinton, McClelland, and Rumelhart, 1986; Knapp and Anderson, 1984; McClelland and Rumelhart, 1985) by investigators using an approach known in some circles as parallel distributed processing, or more generally, connectionism, that explores parallel computing architectures. Such distributed representations have been shown to be important in issues of learning and generalization. This study shows that this representation scheme can be exploited in issues of problem solving as well.

To date, one drawback with a connectionist approach arises from the fact that once the network of elements settle into a state, the network remains in that state. Thus, criticisms exist that while a solution bound by a set of constraints can be found in parallel, serial behavior requiring transitions to successive states cannot be simulated by these networks.

Rumelhart, Smolensky, McClelland, and Hinton (1986) pointed out that the dilemna regarding sequences of states can be resolved by noting that the system can change with a change in the external environment, a change that can be effected in some cases by the experiencer (e.g., game playing). They also noted that these ideas can be extended in a way that allows these networks to simulate mental simulation.

There is, however, a class of stimuli that are *multistable* (e.g., the Necker cube). In these cases, a fixed stimulus can be perceived in two different configurations that alternate with each other. These stimuli suggest an alternative based on the notion of habituation, whereby continued activity of a pair of elements results in a short-term attenuation of that connection (Kawamoto & Anderson, 1985). This property will be used to simulate the solution, using "weak" methods, of two simple puzzles, the *DOG -> CAT* puzzle (i.e., change *DOG* to *CAT* by replacing one letter such that each letter triplet is a legal word) as well as a simpler variant of the 8-puzzle devised by Klahr (1985) known as the dog-cat-mouse (DCM) problem.

THE MODEL

The approach used here is based on the work of Anderson and his colleagues (Anderson, Silverstein, Ritz, & Jones, 1977; Anderson, 1983). Since this model has been described in detail elsewhere (Anderson, et al., 1977; Anderson, 1983; Kawamoto & Anderson, 1985), only the major points and more recent developments will be noted here. The use of a network of simple processing elements operating in parallel to simulate cognitive phenomenon is very similar to many others (see Feldman, 1985; McClelland & Rumelhart, 1986; Rumelhart & McClelland, 1986).

Network architecture

Auto-associative network. The principle network consisted of 216 elements, with each element forming connections to every other element in the network (hence the name, auto-

* The reader is cautioned to be aware that two closely related notions of hill-climbing are used in this paper. One is from the AI literature on search heursitics, and the other one, quite similar in spirit, arises from the connectionist literature in terms of searching for a particular state (energy minimum). In the framework of this paper, the former influences the selection of subsequent states, and the latter allows the network to find local energy minima. To minimize possible confusion, the terms *relaxing* or *settling* will be used in discussions of finding local energy minima.

associative). The activity of an element is represented as a real value ranging between -1.0 and + 1.0. These limits constrain the activity of the network within a high-dimensionality hypercube and is the basis for referring to this model as the "brain-state-in-a-box." The activity of an element changes with time, where time changes in discrete steps, such that subsequent activity of an element, $\mathbf{x}_i$, is simply the sum of the input, $\mathbf{s}_i$, some fraction, δ, of its activity at the previous iteration, and the activity of all the other elements, $\mathbf{x}_j$, weighted by the connection strength, ω_{ij}. That is,

$$\mathbf{x}_i(t+1) = LIMIT[\mathbf{s}_i + \delta \mathbf{x}_i(t) + \sum_j \omega_{ij}\, \mathbf{x}_j(t)], \quad (1)$$

where *LIMIT* constrains the activity to the range from -1.0 to + 1.0. The strengths of these connections are determined adaptively during the learning phase in the manner described in a later section. Since each element is connected with every other element, a feedback loop is formed and the state of the system continues to change until all the elements in the network are saturated (i.e., minimally or maximally activated).

Associative network. For the DCM puzzle only, the network will be supplemented with a second set of 216 elements whose activity represents the most recent puzzle configuration. Connections from *i* th element of this set of elements forms connections v_{ij} to the *j* th element of the first set of elements. The connections v_{ij} capture the constraint on legal moves.

Representation

In this approach, each state is represented as a pattern of features that are either on (a value of + 1.0) or off (a value of -1.0). For both the DOG -> CAT puzzle and the DCM puzzle, there are always a fixed number of components that comprise the configuration. In the first case, there are always only 3 letters, and in the second case, there are always a fixed number of positions and 3 tokens placed in those positions. A word or configuration is formed simply by concatenating patterns corresponding to letters, or tokens and their positions. A more detailed description of the representations will be postponed until the puzzles themselves are discussed.

Learning

An important aspect of modeling efforts of this type is the training involved in producing a network that successfully generates the desired output for a given input. This is achieved by modifying the connection strengths between pairs of elements to capture the correlations in their activities. One approach introduced by the early investigators of learning in networks (Rosenblatt, 1962; Widrow & Hoff, 1960) is to use "error-correction" schemes. Modification of the synaptic weights proceeds by minimizing (i.e., correcting) the error between the desired output and the actual output. For example, if the pattern **g** (representing a 3-letter word or a particular puzzle configuration) is to be learned, the difference between the correct value of the i *th* element of **g**, $\mathbf{g}_i$, and its actual value after a single iteration through the network, $\mathbf{g}_i'$, is used to determine the extent of the modification. Here, $\mathbf{g}_i$ is simply

$$\mathbf{g}_i' = LIMIT[\sum_j \omega_{ij}\, \mathbf{g}_j]. \quad (2)$$

After a learning trial, each connection strength ω_{ij} is modified by

$$\Delta\omega_{ij} = \eta(\mathbf{g}_i - \mathbf{g}_i')\mathbf{g}_j, \quad (3)$$

where η is a scalar learning constant.

For the DCM puzzle, the connections v_{ij} capture the constraint of allowable moves. In this case, the location of the open position of the puzzle always changes with each move. This has been implemented by associating the pattern representing the open position in each of the 4 different slots with possible successive open positions. These connections were modified according to a Hebbian learning scheme (Anderson, et al., 1977).

Energy

Although learning is generally an important consideration for this modeling approach, this particular aspect is not of primary importance in

this study. Rather, this simulation exploits the property that learning algorithms of the type used here lead to the input patterns becoming local energy minima (i.e., stable states) of the system (Golden, 1986; Hopfield, 1982). In these networks, each successive state, **x**, is more energetically favorable than the previous state. For the DOG -> CAT puzzle, an energy measure analogous to one used by Rumelhart, et al. (1986) is used,

$$E(t) = -\sum_i \sum_j \omega_{ij}\, \mathbf{x}_i(t)\mathbf{x}_j(t) - \sum_i \mathbf{x}_i\,(t)\mathbf{s}_i. \quad (4)$$

The hill-climbing search heuristic arises from the contribution of the stimulus, **s**, and the current activity, **x**. For the DCM puzzle. the additional constraint imposed by allowable transitions yields

$$E(t) = -\sum_i \sum_j \omega_{ij}\, \mathbf{x}_i(t)\mathbf{x}_j(t) - \sum_i \mathbf{x}_i\,(t)\mathbf{s}_i \quad (5)$$
$$-\sum_i \sum_j \nu_{ij}\, \mathbf{c}_j \mathbf{x}_i(t),$$

where **c** represents the most recent stable state.

Habituation

Since the elements at one layer are interconnected, a positive feedback loop is formed. Thus, once all the elements in the network are saturated, they tend to remain saturated. One solution to get the system out of this stable state is to *habituate* the system by a short-term modification of the connection strengths (Kawamoto & Anderson, 1985). Given the stable state **c**, a "corner" of the hypercube, the connections are modified by

$$\Delta\omega_{ij} = \gamma \mathbf{c}_i \mathbf{c}_j, \quad (6)$$

where γ is negative valued scalar constant. This modification takes place on each iteration the system is in the stable state. However, this modification is only temporary as the effect decreases exponentially with time.

Essentially, habituation results in a change in the energy landscape; i.e., the current stable state becomes less energetically favorable. At some point, that state no longer corresponds to a local energy minimum and the state of the network moves away from this point toward a new local energy minimum.

DOG- > CAT PUZZLE

Puzzle Description

The objective of this puzzle is to transform a given 3-letter word, *DOG,* to a different 3-letter word, *CAT,* by replacing a single letter of the current word on each successive move such that each new letter triplet forms a valid word.

Representation

In this simulation, the 30 3-letter words listed in Table 1 were learned by the network. Each letter is represented by a 72-dimensional random vector. These patterns are simply concatenated in the appropriate order to produce the pattern for each word.

Simulation

One aspect of this puzzle that makes it somewhat difficult is the large number of possible operators (25 possible letter replacements at each of 3 letter positions) for any given state, most of which lead to non-legal states (i.e., non-words). Rather than trying all possibilities, one strategy is to substitute a letter from the current word with the letter in the corresponding position from the goal, e.g., C in position 1. (These possiblities would all be "uphill" in terms of the hill-climbing heuristic.) This will limit initial consideration to just 3 very likely possibilities. Once a possibility has been generated, it can be tested to see whether or not the resulting triplet is a word or not. In contrast to this sequential method of considering different possibilities, the approach used here essentially considers all choices in parallel. Moreover, the model, in general, "considers" only legal words.

After the 30 words from Table 1 were learned, the network was tested to determine whether a sequence of stable states corresponding to a solution could be generated. The initial state of the network was set to *DOG,* and on each iteration, the pattern corresponding to the goal state *CAT,* was provided as input to the

Table 1. List of the 30 3-letter words learned by the network.

AIM	DOT	LIP
APE	DUB	MAD
ARC	EEL	PIN
BUN	EGO	ORE
CAR	GNU	ROB
CAT	GUM	SUB
COG	ILL	SUN
COT	IMP	TIP
DIG	IRE	URN
DOG	LED	USE

Table 2. A simulation of the solution DOG -> COG -> COT -> CAT.

0.	DOG
1.	DOG
20.	_OG
23.	_O_
36.	_OG
47.	COG
48.	COG
59.	CO_
78.	_O_
87.	COT
88.	COT
95.	C_T
98.	__T
111.	C_T
114.	CAT

Note: The underscores indicate that the subset of elements are not all saturated.

network. A selected sample of states during the iteration are displayed in Table 2. Whenever the network reaches a stable state, the connections are habituated in the manner described earlier.

The sequence of stable states is *DOG -> COG -> COT -> CAT*. While there are no characters common to both the starting state, *DOG*, and goal state, *CAT*, each successive state is always closer to the solution: i.e., the solution, in terms of heuristic search, is strictly uphill. Note that the model did not generate non-words such as *DAG*. This is a general tendency that arises because the lowest energy states of the network correspond to the words learned during the initial training period.*

There are actually two solutions to this puzzle. In addition to the solution generated here, an alternative is the sequence *DOG -> DOT -> COT -> CAT*. Both *DOT* and *COG* represent equally good states with respect to the goal state. In this particular solution, the nondeterminism has been resolved successfully and a legal stable state was generated.

When an impasse in problem solving is reached, headway can often be made by working backwards from the goal state toward the start state. In the AI literature, such an approach is used in searching simultaneously from the start state to the goal state and from the goal state to the start state. Here too, such an approach can be used, and the network successfully simulates the solution in the reverse direction, *CAT -> COT -> DOT -> DOG*. Note that here, the solution takes *DOT* as an intermediate state rather than *COG*.

Although there are false peaks in this problem, a problem does not arise because the initial state is at the base of the tallest peak. The next puzzle illustrates a case where the network begins at a false peak and gets down from it.

DCM PUZZLE

Puzzle Description

In this variant of the 8-puzzle, there are 3 different pieces, a dog, a cat, and a mouse, that must be placed in the configuration shown in Figure 1. A piece can be moved only to the open position, and only if there is a connection, indicated by the solid lines, from its current posi-

* There is, however, no guarantee that the network will not generate non-words because there are essentially only pair-wise connections between elements comprising a pair of letters. The probability of settling into states corresponding to words increases if there are additional units unique to a given word as proposed by Hinton (1981).

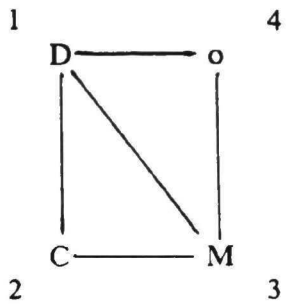


Figure 1. Configuration of the DCM puzzle. There are 4 positions and three tokens, with a token allowed to move only to the open position (if there is a connnection between the two locations).

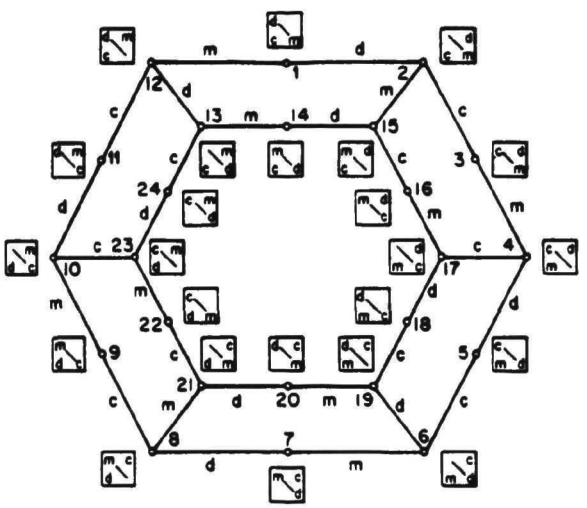


Figure 2. State space representation of the DCM puzzle. [From Klahr, 1985. Copyright 1985 by The Society for Research in Child Development. Reprinted by permission.]

tion to the open position.

Representation

Disregarding rotations and reflections, there are 24 possible configurations. A graph showing all possible states and legal transitions is shown in Figure 2.

Table 3 shows the representation of the 24 possible states of the puzzle. Each configuration is represented in the following way: The piece at each of the 4 positions *(D, C, M,* and *O)* representing Dog, Cat, Mouse, and the Open position, respectively) as well as each piece's position *(d, c,* and *m)* with a numerical subscript from 1 to 4 corresponding to its position), are indicated by a distinct slot.

Table 3. Representation of all possible configurations.

1	DCMO	$d_1c_2m_3$
2	OCMD	$d_4c_2m_3$
3	COMD	$d_4c_1m_3$
4	CMOD	$d_4c_1m_2$
5	CMDO	$d_3c_1m_2$
6	OMDC	$d_3c_4m_2$
7	MODC	$d_3c_4m_1$
8	MDOC	$d_2c_4m_1$
9	MDCO	$d_2c_3m_1$
10	ODCM	$d_2c_3m_4$
11	DOCM	$d_1c_3m_4$
12	DCOM	$d_1c_2m_4$
13	OCDM	$d_3c_2m_4$
14	MCDO	$d_3c_2m_1$
15	MCOD	$d_4c_2m_1$
16	MOCD	$d_4c_3m_1$
17	OMCD	$d_4c_3m_2$
18	DMCO	$d_1c_3m_2$
19	DMOC	$d_1c_4m_2$
20	DOMC	$d_1c_4m_3$
21	ODMC	$d_2c_4m_3$
22	CDMO	$d_2c_1m_3$
23	CDOM	$d_2c_1m_4$
24	CODM	$d_3c_1m_4$

Note: The number for each configuration corresponds to that depicted in Figure 2.

Simulation

Once the legal positions and constraints on the operators have been learned, the ability of the network to simulate the solution of the puzzle was observed. One example run is shown in Table 4. The start state is configuration 14, and the network successively reaches configurations 13, 12, and 1, the goal state.

During the course of running a number of simulations, it was observed that not all states the network settled into corresponded to a configuration. For example, two pieces were sometimes "moved" simultaneously to the open position, creating *2* new open positions and

Table 4. Simulation of solution of the DCM puzzle.

0	MCDO	$d_3c_2m_1$
1	MCDO	$d_3c_2m_1$
19	_CDO	$d_3c_2m_1$
42	_CD_	$d_3c_2m_1$
51	_CD_	d_3c_2__
56	OCD_	d_3c_2__
57	OCD_	$d_3c_2m_4$
58	OCDM	$d_3c_2m_4$
59	OCDM	$d_3c_2m_4$
66	_CDM	$d_3c_2m_4$
89	_C_M	$d_3c_2m_4$
95	_C_M	__c_2m_4
100	_C_M	__c_2__
105	_C_M	d_1c_2__
107	_COM	d_1c_2__
108	_COM	$d_1c_2m_4$
109	DCOM	$d_1c_2m_4$
110	DCOM	$d_1c_2m_4$
117	DC_M	$d_1c_2m_4$
149	DC_M	d_1c_2__
150	D___M	d_1c_2__
154	DC___	d_1c_2__
158	DC___	d_1c_2**
161	DC___	d_1c_2__
163	DC___	__c_2__
167	DC___	__c_2m_3
168	DC_O	__c_2m_3
169	DCMO	__c_2m_3
171	DCMO	$d_1c_2m_3$

Note. The underscores indicate that the corresponding set of units are not saturated. The asterisk indicates that although the set of units are saturated, they do not correspond to a defined pattern.

combination of the two pieces at the position that was formerly empty. This was a manifestation of the non-determinism involved in choosing a successive state and has been observed in young children (Klahr, 1985). In other cases, the equivalent of 2 moves were sometimes taken during one settling period.

In solving puzzles, one way to minimize the path length for a solution is to avoid backup, i.e., not returning to the previous state. Here, this constraint is imposed as a result of habituation. Previous states are less likely to be returned to simply because these states are no longer energetically favorable.

DISCUSSION

In this study, the solution of two puzzles using a hill-climbing search heuristic has been simulated within a connectionist framework. This is implemented by using a distributed representation to capture similarity of the states in the problem space. When the goal state is provided as an input, the more similar a state is to the goal state, the more energetically favorable that state is. Since the network settles into local energy minima, the states the network settles into are similar to the goal state. Once the network settles into a stable state, the connections are habituated. This changes the energy landscape and allows the network to move from the current state to a new stable state that is closer to the goal state.

Networks of this type can also go from one stable state to another (with a fixed input) if the activity of units are stochastic. In such cases, the probability that the network is in a particular state is determined from the Boltzmann distribution (Ackley, Hinton, & Sejnowski, 1985; Selman, 1985).

Although it has been demonstrated here that connectionist networks can simulate solution of puzzles, the behavior simulated is that of a naive problem solver. It would be nice if the network could also learn the *optimal* solution given a particular start state and goal state. An even more ambitious goal is for the network to generate a representation scheme that would allow it to generalize to configurations that were isomorphic. This would almost surely require the network to be able to train hidden units using methods studied by Ackley, et al. (1985) and Rumelhart, Hinton, and Williams (1985).

Acknowledgements

This work was completed while the author was supported by a grant from the Sloan Foun-

dation. The DOG -> CAT puzzle was pointed out by Jeff Shrager, and the DCM puzzle was brought to my attention by Jay McClelland. I would like to thank Jay McClelland for useful discussions of the problem, and Richard Golden for commenting on a draft of the manuscript.

References

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science,* **9,** 147-169.

Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review,* **84,** 413-451.

Anderson, J. A. (1983). Cognitive and psychological computation with neural models. *IEEE: Systems, Man, and Cybernetics.* **SMC-13,** 799-815.

Feldman, J. (Ed.). (1985). *Cognitive Science,* **9,** special issue on connectionism.

Golden, R. M. (1986). The "brain-state-in-a-box" neural model is a gradient descent algorithm." *Journal of Mathematical Psychology,* **30,** 73-80.

Hinton, G.E. (1981). Implementing semantic networks in parallel hardware. In G.E. Hinton & J.A. Anderson (Eds.), *Parallel Models of Associative Memory.* Hillsdale, NJ: Erlbaum.

Hinton, G.E., McClelland, J. L., & Rumelhart, D.E. (1986). Distributed representations. in D. R. Rumelhart & J. L. McClelland (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. v. 1: Foundations.* Cambridge, MA: Bransford Books MIT Press.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences USA,* **79,** 2554-2558.

Kawamoto, A. H., and Anderson, J. A. (1985). A Neural Network Model of Multistable Perception. *Acta Psychologica.* **59,** 35-65.

Klahr, D. (1985). Solving problems with ambiguous subgoal ordering: Preschooler's performance. *Child Developement,* **56.** 940-952.

Knapp, A., & Anderson, J. A. (1984). A signal averaging model for concept formation. *Journal of Experimental Psychology, Learning, Memory, and Cognition.* **10,** 616-637.

McClelland, J. L., and Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General.* **114,** 159-188.

McClelland, J. L., and Rumelhart, D. E. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, v. 2:* Cambridge, MA: Bradford Books MIT Press.

Rosenblatt, F. (1962). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms,* Washington: Spartan Press.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1985). Learning internal representations by error propogation. *ICS Report 8506.*

Rumelhart, D. E., and McClelland, J. L. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. v. 1: Foundations.* Cambridge, MA: Bradford Books/MIT Press.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). Parallel distributed processing models of schemata and sequential thought processes. in D.E. Rumelhart and J.L. McClelland (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. v. 1: Foundations.* Cambridge, MA: Bransford Books MIT Press.

Selman, B. (1985) Rule-Based Processing in a Connectionist System for Natural Language Understanding. *Technical*Report Computer Systems Research Institute, University of Toronto.

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

BoltzCONS: Reconciling Connectionism with the Recursive Nature of Stacks and Trees

David S. Touretzky

Computer Science Department
Carnegie-Mellon University

Abstract

Stacks and trees are implemented as distributed activity patterns in a simulated neural network called BoltzCONS. The BoltzCONS architecture employs three ideas from connectionist symbol processing -- coarse coded distributed memories, pullout networks, and variable binding spaces, that first appeared together in Touretzky and Hinton's neural net production system interpreter. In BoltzCONS, a distributed memory is used to store triples of symbols that encode cons cells, the building blocks of linked lists. Stacks and trees can then be represented as list structures. A pullout network and several variable binding spaces provide the machinery for associative retrieval of cons cells, which is central to BoltzCONS' operation. Retrieval is performed via the Boltzmann Machine simulated annealing algorithm, with Hopfield's energy measure serving to assess the results. The network's ability to recognize shallow energy minima as failed retrievals makes it possible to traverse binary trees of unbounded depth without maintaining a control stack. The implications of this work for cognitive science and connectionism are discussed.

1. Introduction

This paper begins with an assumption: that recursive symbol structures have a place in connectionist theories of cognition. A recursive structure is one whose components may be structures of the same type. Trees are recursive because their branches are trees; stacks are recursive because their tails (that which remains when the first element is removed) are stacks. The recursiveness of a data structure is independent of any algorithm; a recursive data structure may be manipulated by a nonrecursive algorithm, and recursive algorithms may operate on nonrecursive data structures, such as integers. This paper is not concerned directly with the feasibility or utility of recursive *algorithms*. Instead it will focus on the two most common recursive data structures: stacks and trees.

There are many potential uses of stacks in an intelligent system. At the level of conscious problem solving behavior, stacks could be used to keep track of goals while focusing temporarily on subgoals. In dialog or in narrative reading, stacks might be used to track conversation topics in the presence of digressions, parenthetical remarks, or interruptions. In these two examples stack manipulation occurs at a conscious or mentally accessible level. If there turn out to be recursive mental procedures at a *sub*conscious level of processing, then their implementation may rely on an internal (i.e., inaccessible to introspection) control stack. But even without recursion, a control stack might be useful for saving contextual information when a mental procedure must invoke several levels of subprocedure.

Mental trees may be harder to justify than stacks. In conventional AI programs trees are used to represent many sorts of things, such as syntactic structures (parse trees), decision procedures (discrimination nets), and goal/subgoal hierarchies (AND/OR graphs). Whether trees have any reality in the brain, or any essential role in connectionist theories of mind, is another matter, and for now it is an open question. But if connectionist systems could not represent tree structures, that would place a serious *a priori* constraint on the plausibility of connectionist theories.

To demonstrate that connectionism is capable of accomodating recursive structures, this paper presents BoltzCONS, a Lisp-inspired Boltzmann machine (Fahlman et al., 1983; Ackley et al., 1985). The ability to perform

associative retrievals in BoltzCONS and other connectionist architectures makes them unlike conventional von Neumann computers.

2. Stacks and Trees in a Distributed Memory

We begin by considering how stacks and trees might be represented in a distributed memory. In the following section we go on consider how they can be manipulated. Figure 1 shows a cons cell representation of the stack (A B C D) in what is almost conventional Lisp notation; the difference is that each cons cell in the figure has associated with it a unique symbol, or tag.[1] The figure can thereby be encoded as a set of triples of form (**tag car cdr**), as shown below:

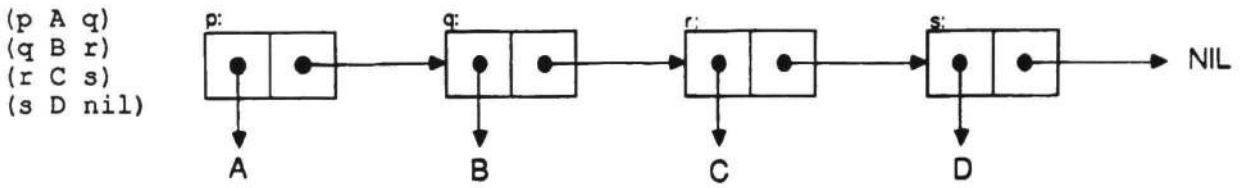


Figure 1: The stack (A B C D) represented as triples and as cons cells.

Trees can be encoded as triples in the same way. Figure 2a shows a binary tree, and figure 2b shows its cons cell representation. The translation from cons cells to triples is straightforward. But how can a connectionist network represent these triples in a computationally plausible way? We reject on efficiency grounds the extreme localist position that there be one unit for every possible triple, since we will only want to store a few triples at a time.

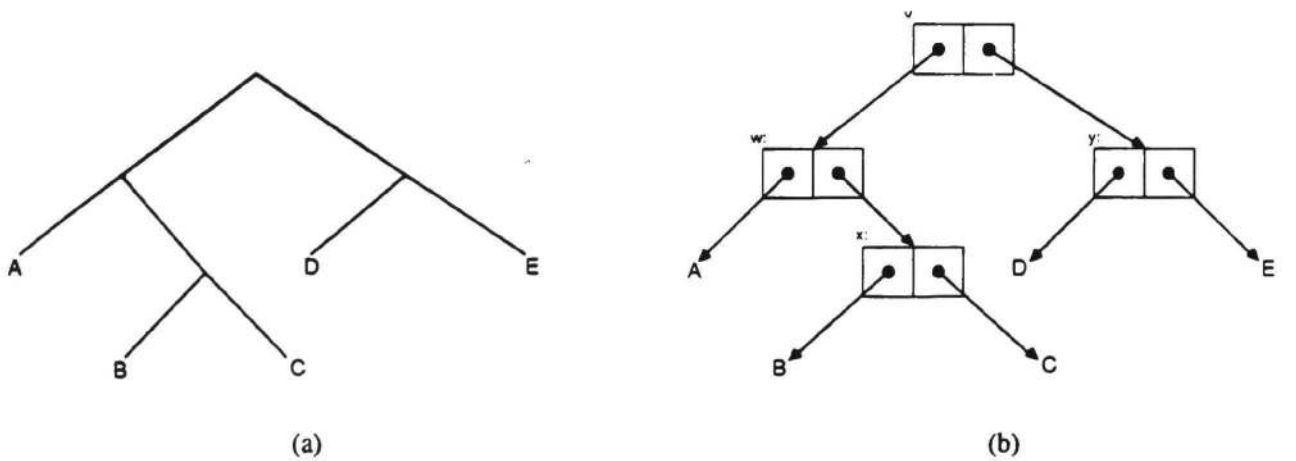


Figure 2: (a) a binary tree; (b) its cons cell representation.

In an earlier paper, Geoffrey Hinton and I described a coarse coded memory for triples of symbols which served as the working memory of a production system interpreter (Touretzky & Hinton, 1985). The same memory design is used in BoltzCONS for encoding cons cells. This scheme is called a distributed representation (Hinton et al., 1986) because each triple is represented by the collective activity of many units, while conversely, each unit contributes to the representation of many possible triples.

[1]Computer scientists might prefer the term "address" to "tag," but a distributed symbol memory such as the one used in BoltzCONS has neither sequential addresses nor discrete memory locations, so "address" would be misleading. Tags are unordered symbols, not integers.

Each memory unit in this representation has a randomly-generated receptive field table such as the one shown in figure 3, with three columns of six symbols each. The receptive field of the unit is defined by the crossproduct of the three columns, e.g., (F A P) would be recognized by the unit described in figure 3 but not (F J T), since J doesn't appear in the second column and T doesn't appear in the third. If the memory contains 2,000 units and there are 25 possible symbols, a triple will fall in the receptive field of $6^3/25^3 \times 2000$ or roughly 28 units, on average. This distributed representation is described as "coarse coded" because each receptive field is a $6 \times 6 \times 6$ slice out of the $25 \times 25 \times 25$ space of possible triples; the receptors can thus be said to be "coarsely tuned" along each of the three dimensions.

C	A	A
F	B	D
H	H	J
P	K	M
S	S	P
W	Z	R

Figure 3: A receptive field table

Starting with a twenty-five symbol alphabet, there are 25^3 or 15,625 possible triples one could potentially store in this memory. The number that can be reliably held there at one time depends on the number of units available. With 2,000 units, one can store roughly a dozen triples at once. As the alphabet size increases, the savings of a distributed representation over an extreme localist one grows rapidly.

To manually store a triple in this memory, we find its 28 or so receptors and turn them on. To delete a triple, we turn its receptors off. To test whether a triple is present, we examine the state of its receptors. If most of them (say, at least 70 percent) are on, we assume that the triple is present; if most are off we assume it is absent. These functions are embedded in the wiring pattern of the network; BoltzCONS does not actually search through receptive field tables when storing or retrieving triples.

The distributed memory has some interesting properties. It is immune to small amounts of noise, i.e., if a few units change state randomly there is no observable effect on the memory's contents. The performance of the memory degrades gradually as the memory fills up. Since each receptor is shared by 6^3 or 216 potential triples, there is some overlap in the representation of triples. As a consequence, if we store a triple in memory and then store and delete many other triples, the original triple will decay and eventually fade away unless it is refreshed by turning on all its receptors again. Finally, the memory exhibits a local blurring phenomenon, whereby if we store many closely related triples such as (F A A), (F A B), (F A C), etc., it becomes increasingly difficult to determine whether a related triple such as (F A L) is present in memory or not. However, it is still possible to tell whether an unrelated triple such as (G V Q) is present. Local blurring, like the gradual decay of triples due to deletions of other triples, is a consequence of the sharing of units in distributed representations.

3. The BoltzCONS Architecture

The architecture of BoltzCONS is shown in figure 4. Cons cells encoded as triples are stored in the space labeled Cons Memory, which contains 2,000 units. There is a one-one mapping of excitatory connections from Cons Memory units to units in the space labeled Cons Pullout; each active Cons Memory unit therefore tries to turn on the corresponding Cons Pullout unit. However, the Cons Pullout units are sufficiently mutually inhibitory so that only about 28 at a time can be on, i.e., just enough to represent one triple. Cons Pullout space is used to "pull out" one triple from the several that are stored in Cons Memory. The term "pullout network," first used by Michael Mozer (1984), is synonymous with the "clause spaces" described by Touretzky & Hinton (1985).

The three spaces labeled TAG, CAR, and CDR are winner-take-all networks (Feldman & Ballard, 1982) which settle into stable states representing one of the 25 symbols in BoltzCONS' alphabet. They are coarse coded versions of the bind spaces described in (Touretzky & Hinton, 1985). Since the spaces are coarse coded, units vote for sets of symbols rather than individual ones. The units all have bidirectional excitatory connections to appropriate units in Cons Pullout space. For example, a unit in CDR space that codes for the symbols **p** and **w** will have connections to a random subset of all the units in Cons Pullout space where either **p** or **w** appears in the third column of the unit's receptive field table.

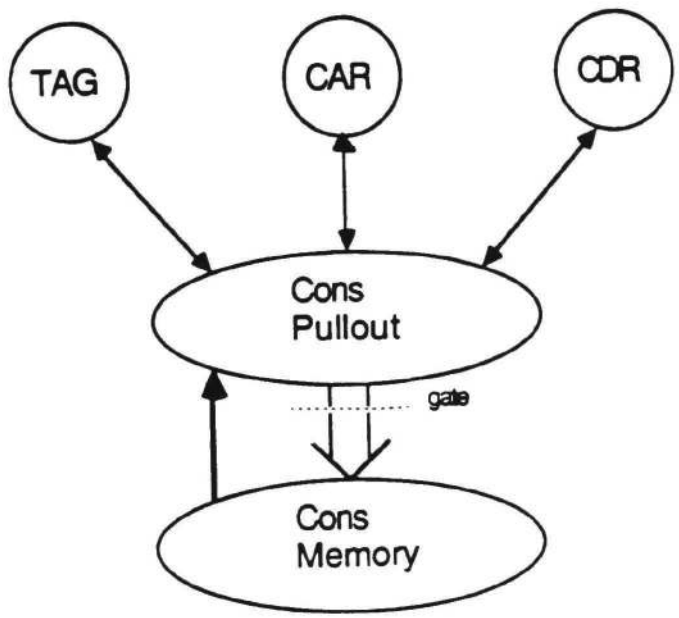


Figure 4: The architecture of BoltzCONS.

4. Manipulating Distributed Stacks

Let us assume the stack of figure 1 has been loaded into BoltzCONS' Cons Memory. The four cons cells in this figure are represented by an activity pattern. The top cell of the stack, encoded as the triple (**p A q**), is also represented by a pattern in Cons Pullout space, and its three component symbols are represented by patterns in TAG, CAR, and CDR spaces, respectively. Now suppose we wish to pop the stack. First, via a gated connection, the Cons Pullout units turn off their corresponding Cons Memory units, thereby deleting the triple (**p A q**) from Cons Memory. Next, the activity pattern in CDR space (denoting the symbol **q**) is transmitted to TAG space, which is then clamped. Finally, while TAG space supplies an excitatory stimulus to Cons Pullout space, a simulated annealing algorithm (Kirkpatrick et al., 1983) is run on the Cons Pullout, CAR and CDR spaces to effect an associative retrieval. The annealing algorithm, which is what makes BoltzCONS a Boltzmann Machine, causes BoltzCONS to search for a minimum energy state subject to the constraints imposed by units in TAG space and

Cons Memory. When it finds this stable, minimum energy state, the Cons Pullout network will have settled into a representation of the triple in Cons Memory whose **tag** component is **q**, namely **(q B r)**, and the CAR and CDR spaces will have settled into states **B** and **r**, respectively, representing the car and cdr components of the new top of the stack.

Pushing an element onto the stack is simpler than popping it, because a stack push does not require annealing. The current top cell of the stack, which after the the first pop would be **(q B r)**, is always represented in both Cons Pullout space and the TAG, CAR, and CDR spaces. To push the symbol **E** onto the stack at this point, we copy the contents of TAG space into CDR space, load the symbol **E** into CAR space, and load a new symbol, say **t**, into TAG space. Then we allow the TAG, CAR, and CDR spaces to independently apply excitation to units in Cons Pullout space. Their combined influence causes the triple **(t E q)** to appear there. (Recall that due to mutual inhibition only about 28 units at a time can be active in the pullout space; the units most likely to be active are those that receive support from the TAG, CAR, and CDR spaces simultaneously.) Finally, by opening a gated connection, Cons Pullout units are allowed to turn on their corresponding Cons Memory units, thus creating a new cons cell.

The use of associative retrieval allows BoltzCONS to perform certain additional stack operations not possible on a convenional machine. One of these is called "associative stack pop." The idea is that rather than popping the stack a fixed number of times, we may want to pop back to a particular state, e.g. the state where C is the top element. This can be accomplished in one step by clamping C into CAR space and running an associative retrieval. Note that in this case the elements "popped" from the stack in order to reach C are still present in Cons Memory; all we have done is move the stack pointer. Because these elements are not deleted, it is possible to restore them by sequentially "un-popping" the stack, again using associative retrieval. To perform one un-pop operation, we copy the symbol presently in TAG space into CDR space and then anneal with CDR space clamped.

5. Unbounded Depth Tree Traversal

One can do many more interesting things with trees than with stacks. Common tree problems include traversal, structural comparison (the Lisp EQUAL function), and terminal node comparison (the "samefringe" problem, often used to motivate coroutines.) Here we consider the problem of traversal. The goal is to find and display in correct order all the terminal nodes of a tree such as the one in figure 2b. The simplest Lisp solution is:

```
(defun traverse (tree)
  (cond ((atom tree) (print tree))
        (t (traverse (car tree))
           (traverse (cdr tree)))))
```

Unfortunately the above algorithm is recursive but not tail-recursive, so it requires a stack. One goal of this paper is to show how recursive structures can be manipulated without resorting to a control stack. (We cannot categorically deny the existence of such a stack in the brain, but our defense of the plausibility of connectionist models would be weakened if it depended on its existence.) The traversal problem cannot be solved on a conventional computer without such a stack.[2] Although we can replace the recursive algorithm with an iterative one, shown below, this merely forces us to build and manage the stack explicitly rather than relying on Lisp's internal control stack.

[2]This claim rests on two assumptions: that trees are represented as *one-way* linked lists as in Lisp, and that destructive operations on the tree are not permitted.

```
(defun iterative-traverse (tree &aux stack)
  (loop
    (cond ((consp tree)
           (push (cdr tree) stack)
           (setq tree (car tree)))
          (t (print tree)
             (if (null stack) (return nil)
                 (setq tree (pop stack)))))))
```

The iterative algorithm below, which uses associative retrieval, does not require a stack to execute. For conciseness, it is expressed using the abstract Lisp terminology of variables, pointers, and cons cells, but its implementation is actually in terms of BoltzCONS operations. We assume that the input tree contains at least one cons cell, and that the variable PTR, which points to the current position in the tree as we traverse it, starts out by pointing to the root.

1. If PTR points to a cons cell, set OLD to PTR, set PTR to its car, and go to step 1. Otherwise PTR must point to a symbol, so print it; then set PTR to cdr of OLD and go to step 2.
2. If PTR now points to a cons cell, go to step 1. Otherwise PTR must point to a symbol, so print it. Then set PTR to OLD and go to step 3.
3. If PTR points to the root, halt. Otherwise, use associative retrieval to search for a cell whose car is PTR. If the retrieval succeeds, make OLD point to that cell, set PTR to the cdr of OLD, and go to step 2.
4. Use associative retrieval to locate the cell whose cdr is PTR. Make PTR point to that cell and go to step 3.

The most important feature of this algorithm is that associative retrieval is used to follow pointers "backward," an operation that is not possible in Lisp. Since a cell may be pointed to by either the car or cdr of its parent cell in the tree, two retrieval attempts may be necessary in order to find the parent. Steps 3 and 4 accomplish this. Table 1 shows a trace of the algorithm during a traversal of the tree in figure 2b. The values shown for PTR and OLD are those immediately before the given step is executed. The places where the algorithm backs up over a pointer are indicated.

Step	PTR	OLD	Action or Comment
1	v		
1	w	v	
1	A	w	print "A"
2	x	w	
1	x	w	
1	B	x	print "B"
2	C	x	print "C"
3	x	x	*Associative retrieval on CAR fails,*
4	x	x	*but retrieval on CDR succeeds.*
3	w	x	*Retrieval on CAR succeeds.*
2	y	v	
1	D	y	print "D"
2	E	y	print "E"
3	y	y	*Associative retrieval on CAR fails,*
4	y	y	*but retrieval on CDR succeeds.*
3	v	y	halt

Table 1: Tree traversal algorithm applied to figure 2b.

Because we can follow pointers backward, there is no need to use a control stack for backtracking. In theory, then, the above iterative algorithm can be used to traverse any binary tree no matter how great its depth. In practice the only depth limitation derives from the size of the trees BoltzCONS can store, which in turn depends on the number of symbols in its alphabet and the number of triples that can be stored in Cons Memory.

6. Use of Hopfield's Energy Measure

Because BoltzCONS behaves as a Boltzmann Machine during simulated annealing[3], it acts to minimize an energy measure as it settles into a stable state. This energy measure was first proposed by Hopfield in an analogy to spin glass models in physics (Hopfield, 1982). If s_i is the state of the *ith* unit (either 0 or 1), θ_i is its threshold, and $w_{i,j}$ is the weight between it and the *jth* unit, then the energy of the network in a given state is:

$$E = \sum_i s_i \theta_i - \sum_{i<j} s_i s_j w_{i,j}$$

The energy measure can be used to detect whether an associative retrieval has succeeded. For example, suppose BoltzCONS is examining the cons cell labeled **x** in figure 2b, and the traversal algorithm now wants to find the parent cons cell. The parent will either have **x** in its car or in its cdr. If we clamp the symbol **x** into CAR space and run an annealing, the network will of course settle into an energy minimum, but in this case the energy will be high. This is because there is no cons cell in Cons Memory with **x** in its car, so there is no way for the network to find a good solution to the combination of constraints imposed by CAR space and Cons Memory space. If the energy of the stable state found by the annealing algorithm is above some empirically determined threshold, we know that the associative retrieval has failed.

When the first retrieval fails, BoltzCONS can try a retrieval with the symbol **x** clamped into CDR space instead. This time the retrieval will succeed, with the Cons Pullout units settling into a representation of the triple (**w A x**). The low energy of this state confirms that a valid cons cell has been retrieved.

Of course, we would prefer that the units in a connectionst model not be forced to measure global properties such as energy in order to function. In BoltzCONS, they don't. What actually happens is this: after the network has settled into a stable state, the thresholds of all the units are raised by a fixed amount[4], thus changing the energy landscape. If the network is in a deep energy minimum, representing a valid retrieval, its state will remain stable. However, if the network has settled into a shallow minimum, then when the thresholds change it will no longer be in a minmum energy state, but in a state from which it can reach a new local minimum by turning all its units off. So, the test for a successful associative retrieval consists of raising the thresholds, running the annealing algorithm for a few more steps at very low temperature (which makes the network act like a Hopfield net), and then checking to see whether any units remain active.

7. Controlling BoltzCONS

One important issue that has not been addressed in this paper is how BoltzCONS is to be controlled. Some external agent must send it commands to clamp a space, copy the state of one space into another, or begin an annealing. A simple sequence of these commands results in a stack push or pop; a more elaborate sequence can result in a tree traversal. For the work described here, the job of controlling BoltzCONS was handled by a piece of Lisp code. In addition to issuing control commands, the Lisp code was responsible for generating new tags for stack

[3]BoltzCONS is not a true Boltzmann Machine because it contains gated and asymmetric connections. However, during an annealing the network is equivalent to one that is a true Boltzmann Machine.

[4]This has the same effect as supplying all the units with some inhibitory bias.

push operations.

Recent work on transforming parse trees in a neural net led to the development of a connectionist control unit for BoltzCONS (Touretzky, 1986). This control unit is a modified version of Touretzky and Hinton's production system interpreter. Production rules describing a multi-step parse tree transformation contain right hand side actions that send commands and data to BoltzCONS. Executing these production rules in the proper sequence causes the parse tree (represented in BoltzCONS' Cons Memory) to be transformed in the desired way. A similar control unit, programmed with a different set of production rules, could be used to automate the steps involved in tree traversal.

8. Discussion

Turing machines, which as far as anyone knows can compute all computable functions, consist of an immutable finite state machine and a mutable external tape of unbounded length. Conventional von Neumann computers, being of fixed size, are merely large finite state machines, but it is usually more helpful to think of them as imperfect Turing machines. Doing so allows us to divide the computer into a finite state machine whose repertoire of states is relatively small, plus a tape (embedded in the computer's memory rather than being external) that is much larger in size, albeit finite. The linear address structure of a von Neumann machine's memory makes it natural to equate this memory with the Turing machine's tape.

Connectionists much prefer distributed memory models, both because they are more physiologically plausible and because they are more amenable to parallel processing. A distributed memory such as the one found in BoltzCONS offers neither discrete locations nor sequential addresses. Therefore the analogy between connectionist hardware and imperfect Turing machines is less straightforward (Pylyshyn, 1984). While there is no evidence that connectionist models can compute things Turing machines can't, to harp on this misses the real point, which is that connectionist models can compute things *in ways in which* Turing machines and von Neumann computers can't. The ability to rapidly solve complex constraint satisfaction problems through massive parallelism, such as when performing associative retrievals, is a crucial part of the connectionist program.

9. Conclusions

This work demonstrates that connectionist models can indeed represent and manipulate recursive symbolic structures. Furthermore, thanks to associative retrieval, these recursive structures can sometimes be manipulated iteratively without the use of stacks, while a conventional computer would require a stack. The associative retrieval capability also makes certain operations such as associative stack pop possible that cannot be done (in constant time) on a conventional computer.

Three techniques used in BoltzCONS -- coarse coded symbol memories, pullout networks, and bind spaces, were first put together in a neural network production system interpreter. They now appear to be generally useful devices for connectionist symbol processing, and are likely to turn up again in other connectionist applications. Coarse coded, distributed representations are useful for building complex symbol structures economically. Without pullout networks and bind spaces, though, we would have no way to manipulate them.

Acknowledgements

This material is based on work supported by a grant from the System Development Foundation, and by National Science Foundation Grant No. IST-8516330. I am grateful to Geoffrey Hinton, Mark Derthick, Jay McClelland, and David Plaut for useful discussions, and to Cindy Wood for help with the illustrations.

References

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985) A learning algorithm for Boltzmann Machines. *Cognitive Science* 9(1):147-169.

Fahlman, S. E., Hinton, G. E., & Sejnowski, T. J. (1983) Massively parallel architectures for AI: Netl, Thistle, and Boltzmann Machines. *Proceedings of AAAI-83*, Washington, DC, 109-113.

Feldman, J. A., & Ballard, D. H. (1982) Connectionist models and their properties. *Cognitive Science* 6:205-254.

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986) Distributed representations. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume I.* Bradford Books, Cambridge, MA.

Hopfield, J. J. (1982) Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences USA* 79:2554-2558.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983) Optimization by simulated annealing. *Science* 220:671-680.

Mozer, M. C. (1984) The perception of multiple objects: a parallel, distributed processing approach. Unpublished thesis proposal, Institute of Cognitive Science, University of California at San Diego.

Pylyshyn, Z. W. (1984) Why computation requires symbols. *Proceedings of the Sixth Annual Conference of the Cognitive Science Society*, 71-73.

Touretzky, D. S., & Hinton, G. E. (1985) Symbols among the neurons: details of a connectionist inference architecture. *Proceedings of IJCAI-85*, Los Angeles, CA.

Touretzky, D. S. (1986) Representing and transforming recursive objects in a neural network, or "trees *do* grow on Boltzmann machines." *Proceedings of the 1986 IEEE International Conference on Systems, Man, and Cybernetics*, Atlanta, GA.

Attractor dynamics and parallelism in a connectionist sequential machine

Michael I. Jordan
Department of Computer and Information Science
University of Massachusetts

ABSTRACT

Fluent human sequential behavior, such as that observed in speech production, is characterized by a high degree of parallelism, fuzzy boundaries, and insensitivity to perturbations. In this paper, I consider a theoretical treatment of sequential behavior which is based on data from speech production. A network is discussed which is essentially a sequential machine built out of connectionist components. The network relies on distributed representations and a high degree of parallelism at the level of the component processing units. These properties lead to parallelism at the level at which whole output vectors arise, and constraints must be imposed to make the performance of the network more sequential. The sequential trajectories that are realized by the network have dynamic properties that are analogous to those observed in networks with point attractors (Hopfield, 1982): learned trajectories generalize, and attractors such as limit cycles can arise.

INTRODUCTION

One of the arguments for "connectionist" or "parallel, distributed processing" networks has been that they have properties that seem to reflect processes at which humans are most naturally proficient (Hinton & Anderson, 1981; Rumelhart & McClelland, 1986). These properties include the ability to generalize from instances, the ability to deal with partial information, and insensitivity to noise. It has been suggested that it might be advisable to base theories on such primitives rather than on those associated with sequential, symbolic processing. These arguments have been made mostly in the context of models dealing with the interpretation of incoming data, or with mappings from one set of data to another. However, one need only consider the fluency of human speech to see that humans are also very good at certain kinds of sequential behavior. Furthermore, such behavior is often characterized by a high degree of parallelism, fuzzy boundaries, and insensitivity to perturbations — properties which are difficult to capture in a formalism in which the

underlying primitive is a sequential processor, but which are more natural in a connectionist system. In this paper, I consider the problem of parallelism in speech production and suggest a connectionist architecture that can exhibit behavior similar to that shown in speech. My approach is based on the recent work on learning by Rumelhart, Hinton, and Williams (1986) and is related to previous work by Henke (1966), Kohonen, Lehtio, and Oja (1981), and Rumelhart and Norman (1982).

COARTICULATION

Much of the complexity of describing sequential processes in speech production comes from the fact that speech gestures associated with nearby phonemes can overlap in time. Such overlap, or *coarticulation*, is ubiquitous in utterances and can be quite complex, given the many degrees of freedom of the speech apparatus. It is possible to see gestures that anticipate future phonemes, referred to as *forward coarticulation*, as well as perseveratory gestures, or *backward coarticulation*. The overall effect of coarticulation is to make the utterance more smooth by merging nearby phonemes and to allow speech to proceed faster than would otherwise be possible by taking advantage of opportunities for the parallel execution of movements.

Several studies have investigated coarticulation by recording articulator trajectories during utterances. Moll and Daniloff (1971) showed that in an utterance such as "freon", the velar opening for the nasal /*n*/ can begin as early as the first vowel, thereby nasalizing the vowels. [1] Benguerel and Cowan (1974) studied phrases such as "une sinistre structure," in which there is a string of the six consonants /*strstr*/ followed by the rounded vowel /*y*/. [2] They showed that lip-rounding for the /*y*/ can begin as early as the first /*s*/, an example of forward coarticulation over six phonemes.

One way to characterize these examples is to say that if certain degrees of freedom are not being used in the production of a particular sound, then they may anticipate or perseverate aspects of other phonemes in the utterance so that performance becomes more parallel. However, such a conception of coarticulation ignores the constraints which must be imposed on the parallelism. Certain anticipatory gestures, for example, would inflict too much change on the sound currently being produced, and there must therefore be a way to prevent such coarticulations while allowing others. In the case of "freon", for example, the velum is allowed to open during the production of the vowels because the language being spoken is English. In a language such as French, in which nasal vowels

[1] The velum is a muscular tissue that opens to allow air to pass between the pharynx and the nasal cavities.

[2] The vowel /*y*/ is the "u" in "tu", and is somewhat like pronouncing the English sound "ee" with rounded lips.

are different phonemically from non-nasal vowels, the velum would not be allowed to coarticulate. Thus the articulatory control system cannot blindly anticipate articulations, but must be sensitive to phonemic distinctions in the language being spoken.

The implementation of constraints on parallelism is complicated by the fact that the constraints cannot be encoded as relations between whole phonemes, but must be specific to particular phonemic structure. For example, in the case of */strstry/*, the *rounding* of the */y/* can be anticipated during the preceding consonants, but the *voicing* of the */y/* cannot, because that would change the phonemic identities of the consonants (for example, the */s/* would become a */z/*). Other features of the */y/*, such as those specifying tongue position, may be more or less constrained, depending on the particular allowable variations of the consonants. Again, such knowledge cannot come from consideration of strategies of articulation, but must reflect higher-level phonemic constraints.

Thus, speech presents an interesting control problem in which constraints of various kinds are imposed on the particular patternings of parallelism and sequentiality that can be obtained in an utterance. What I wish to discuss in the remainder of this paper is an approach to this problem based on connectionist mechanisms.

CONNECTIONIST NETWORKS

General discussions of connectionist networks can be found in Feldman and Ballard (1982) and Rumelhart and McClelland (1986). For present purposes, the main features of the networks that are relevant involve distributed representations, nonlinearities, and learning.

A one layer network with no recurrent connections computes a function from the vector of activation of its input units to the vector of activation of its output units. It is possible for such a network to learn to make associations between input vectors and output vectors. This can be done by an error-correcting learning rule that changes the weights coming in to each output unit in proportion to the difference between the actual output of that unit and the desired output (Widrow & Hoff, 1960).

An important property of such networks, which is due to the weighted sums that units form in determining their activations, is that similar input vectors tend to produce similar output vectors. Many connectionist approaches take advantage of this property by representing entities as distributed patterns of activation, so as to achieve a kind of automatic generalization between similar patterns (Hinton & Anderson, 1981).

A one layer network has only a single weight matrix and is restricted to linear mappings. By allowing more layers, with nonlinear activation functions on intermediate units, it is possible to implement arbitrary nonlinear mappings. Until recently, the learning rules in

these networks were restricted to the single layer case. However, several rules have now been developed for multilayer networks that allow essentially arbitrary associations to be formed (Ackley, Hinton, & Sejnowski, 1985; Barto & Anandan, 1985; Rumelhart, Hinton, & Williams, 1986). The back-propagation rule of Rumelhart, Hinton, and Williams has been used in simulations of the network discussed in this paper. The back-propagation rule is an error-correction procedure that generalizes the Widrow-Hoff rule. As before, errors are generated at the output units by comparing the actual outputs to the desired outputs, and these errors are used to change the weights of the output units. The errors are then propagated back into the network to provide intermediate units with error signals so that they can change their weights.

A NETWORK ARCHITECTURE FOR SEQUENTIAL PERFORMANCE

Let there be some sequence of *actions* $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, which are to be produced in order in the presence of a *plan* $\mathbf{p}$. Each action is a vector in a feature or parameter space, and the plan can be treated as an action produced by a higher level of the system. The plan is assumed to remain constant during the production of the sequence, and serves primarily to designate the particular sequence which is to be performed.

We wish to construct a network that can perform arbitrary sequences by taking a plan as input and producing the corresponding sequence. One approach is to explicitly represent the state of a sequential machine as an activation vector and to produce actions by evaluating a function from states to actions. At each moment in time, an action is chosen based on the current state $\mathbf{s}$, and the state is then updated to allow the next action to be chosen. Thus, there is a function f which determines the output action $\mathbf{x}_n$ at time n,

$$\mathbf{x}_n = f(\mathbf{s}_n, \mathbf{p})$$

and a function g which determines the state $\mathbf{s}_{n+1}$,

$$\mathbf{s}_{n+1} = g(\mathbf{s}_n, \mathbf{p}), \tag{1}$$

where both functions depend on the constant plan vector as well as the current state vector. Following the terminology of automata theory (Booth, 1969), f will be referred to as the *output* function, and g will be referred to as the *next-state* function. [3]

The basic network architecture is shown in Figure 1. The entities in the state equations — plans, states, and outputs — are all assumed to be represented as distributed patterns

[3] From the definition, it can be seen that the plan p plays the role of the input symbol in a sequential machine. The use of the term "plan" is to emphasize the assumption that p remains constant during the production of the sequence. That is, we are not allowed to assume temporal order in the input to the system.

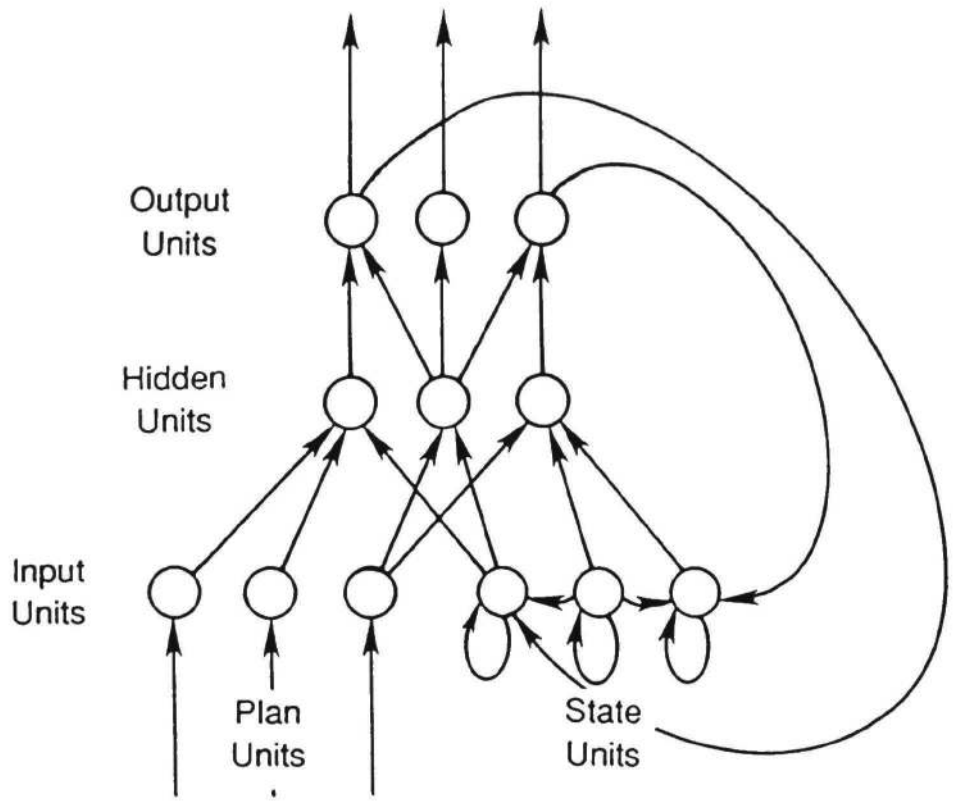


Figure 1: Basic network architecture. (Not all connections are shown).

of activation on separate pools of processing units. The plan units and the state units together serve as the input units for a multilayer network. This network implements the output function through weighted connections from the plan and state units to the output units. The output function is generally nonlinear, as will be discussed below, therefore it is also necessary to have hidden units in the path from the plan and state units to the output units. Finally, the next-state function is implemented with recurrent connections from the state units to themselves, and from the output units to the state units. This allows the current state to depend on the previous state and on the previous output (which is itself a function of the previous state and the plan).

The network can learn to produce sequences of actions by changing the weights in the network. Assume that the recurrent connections implementing the next-state function are given fixed values (particular choices for these values are discussed below). At each time step, an activation vector composed of the plan and the state is present on the input units, and an association can be learned from this input vector to a desired output vector. Clearly, one requirement for the network to be able to learn arbitrary sequences is that the next-state function produce distinguishable state vectors at each time step. It

is not necessary that these vectors be different between sequences, because the plan serves to distinguish the sequences. A second requirement is that there be no restrictions on the form of the associations that can be learned (such as a linearity restriction). This requirement is met by using the back-propagation learning rule. Note that the ability to learn arbitrary sequences does not imply that all sequences are equally easy to learn; indeed, the network will have more difficulty in learning and performing sequences when distinctions must be made similar states and plans.

A further requirement must be imposed on the next-state function so that the results on parallelism will hold: *State vectors at nearby points in time must be similar.* There are many ways to choose the recurrent connections so as to achieve this continuity property. One particular choice, which has been used in many of the simulations of the network, is based on a conception of the state as representing the *temporal context* of actions. Consider the case of a sequence with a repeated subsequence or a pair of sequences with a common subsequence. It seems appropriate, given the positive transfer which can occur in such situations as well as the phenomena of capture errors (Norman, 1981), that the state should be similar during the performance of similar subsequences. One way to achieve this is to define the state in terms of the actions being produced. However, the representation must provide an extensive enough temporal context so that there are no ambiguities in cases involving repeated subsequences. If the state were to be defined as a function of the last n outputs, for example, then the system would be unable to perform sequences with repeated subsequences of length n, or to distinguish between pairs of sequences with a common subsequence of length n. To avoid such problems, the state can be defined as an exponentially weighted average of past outputs, so that the arbitrarily distant past has some representation in the state, albeit with ever-diminishing strength. This representation of the state can be obtained if each output unit feeds back to a state unit with a weight of one, if each state unit feeds back to itself with a weight μ, and if the state units are linear. [4] In this case, the state at time n is given by

$$\begin{aligned} \mathbf{s}_n &= \mu \mathbf{s}_{n-1} + \mathbf{x}_{n-1} \\ &= \sum_{\tau=1}^{n-1} \mu^{\tau-1} \mathbf{x}_{n-\tau}. \end{aligned}$$

Since this representation of the state is an average, it tends to have the desired property that states nearby in time are similar. The similarity depends on the particular actions that are added in at each time step and on the value of μ. In general, however, with sufficiently large values of μ, the similarity extends forward and backward in time, growing weaker with increasing distance.

Other possible representations of the state are discussed in Jordan (1985). The major differences between different representations is in the particular metrics they induce on the

[4] The linearity assumption gives the state a simple interpretation and also gives the state units a more extended dynamic range, but is not essential for the operation of the network.

difficulty of learning and performing particular sequences and also the kinds of generalizations that can be made between sequences. It is also possible to consider learning of the next-state function. Indeed, the back-propagation algorithm applies to the case of recurrent networks, although in a more complex form, requiring units to store histories of their activations (Rumelhart, Hinton, & Williams, 1986). However, in the current framework, there is little to be gained by learning the next-state function; all the hard work can be done in learning the output function.

LEARNING AND PARALLELISM

The network as described thus far would appear to be strictly sequential: there is no overlap between neighboring actions. This is indeed the case and it is necessary to modify the form in which desired output vectors are specified to see that the network is in fact capable of highly parallel performance.

The form that desired output vectors are assumed to take is a generalization of the approach used in traditional error-correction schemes (Duda & Hart, 1973; Rosenblatt, 1961; Rumelhart, Hinton, & Williams, 1986; Widrow & Hoff, 1960). Rather than assuming that a value is specified for each output unit, it is assumed that in general there are *constraints* specified on the values of the output units. Constraints may specify a range of values which an output unit may have, a particular value, or no value at all. This latter case is referred to as a "don't-care condition." It is also possible to consider constraints which are defined among output units. For example, the sum of the activations of a set of units may be required to have a particular value.

Constraints enter into the learning process in the following way: If the activation of an output unit fits the constraints on that unit, then no error corrections are instigated from that unit. If, however, a constraint is not met, then the error is defined as a proportion of the degree to which that constraint is not met, and this error is used in the normal way to change weights towards a configuration in which the constraint is met. An example of this process is shown in Figure 2 for a desired output vector with three specified values and two don't-care conditions (represented by stars). As shown in the figure, errors are propagated from only those units where constraints are imposed. In the case of constraints among units, it is possible to impose constraints on units having fixed connections from the output units. Errors generated at these units are propagated back to the output units. This process is sketched in Figure 3, where the constraints $x_1 + x_2 = .6$ and $x_2 + x_3 = .4$ are imposed. Note that if many constraints are imposed on the same unit, the errors are simply added together, and the network will eventually find an activation value for the unit that satisfies all of the constraints (given that such a value exists).

Consider now the case in which desired output vectors specify values for only a single

$$\mathbf{x}_i = \begin{bmatrix} * \\ .7 \\ * \\ .3 \\ * \end{bmatrix}$$

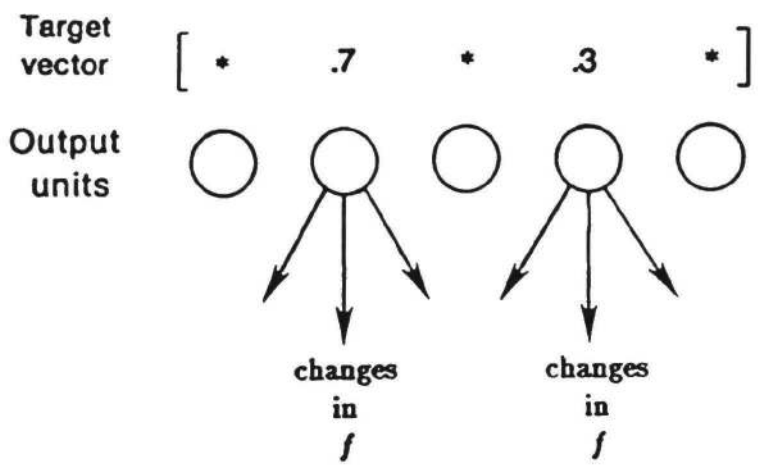


Figure 2: Learning with don't-care conditions.

output unit. This is essentially the case of local representations for actions, in which the network is essentially being instructed to activate its output units in a particular order. Suppose, for example, that a network with three output units is learning the sequence

$$\begin{bmatrix} 1 \\ * \\ * \end{bmatrix}, \begin{bmatrix} * \\ 1 \\ * \end{bmatrix}, \begin{bmatrix} * \\ * \\ 1 \end{bmatrix}$$

At each time step, errors are propagated from only a single output unit, so that activation of that unit becomes associated to the current state. Associations are learned from $\mathbf{s}_1$ to activation of the first output unit, from $\mathbf{s}_2$ to activation of the second output unit, and from $\mathbf{s}_3$ to activation of the third output unit. [5]

[5] where $\mathbf{s}_i$ denotes the activation of the state units at time i. I am ignoring the plan vector to simplify the exposition.

$$\mathbf{x}_i = \begin{bmatrix} x_1 + x_2 = .6 \\ x_2 + x_3 = .4 \end{bmatrix}$$

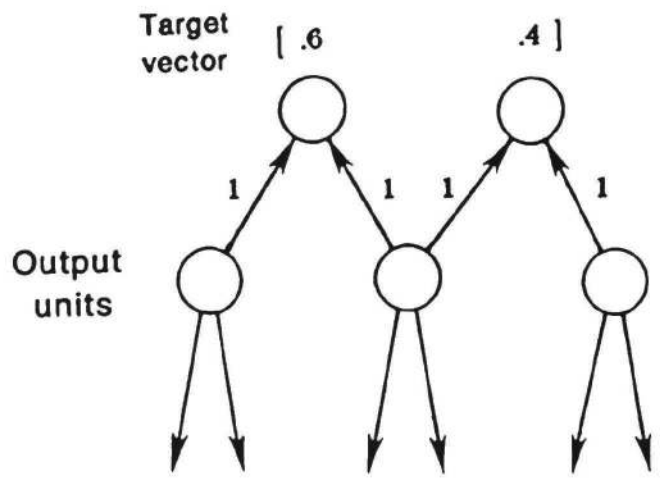


Figure 3: Learning with constraints among units.

After learning, the presence of $\mathbf{s}_1$ on the state units will activate the first output unit. It will also partially activate the second and third output units, even though no associations from $\mathbf{s}_1$ to these units have been learned. This occurs because $\mathbf{s}_1$ is similar to $\mathbf{s}_2$ and $\mathbf{s}_3$ (given the requirement made of the next-state function) and similar inputs tend to produce similar outputs in these networks. The associations made to $\mathbf{s}_2$ and $\mathbf{s}_3$ also generalize, so that after learning, the network will likely produce a sequence such as

$$\begin{bmatrix} 1 \\ .8 \\ .6 \end{bmatrix}, \begin{bmatrix} .8 \\ 1 \\ .8 \end{bmatrix}, \begin{bmatrix} .6 \\ .8 \\ 1 \end{bmatrix},$$

where at each time step, there are parallel activations of all output units. If the network is driving a set of articulators that must travel a certain distance, or have a certain inertia, then it will be possible to go faster with these parallel control signals than with signals where only one output unit can be active at a time.

The foregoing example is simply the least constrained case and further constraints can be added. Suppose, for example, that the second output unit is not allowed to be active during the first action. This can be encoded in the constraint vector for the first action so that the network is instructed to learn the sequence

$$\begin{bmatrix} 1 \\ 0 \\ * \end{bmatrix}, \begin{bmatrix} * \\ 1 \\ * \end{bmatrix}, \begin{bmatrix} * \\ * \\ 1 \end{bmatrix}.$$

After learning, the output sequence will likely be as follows:

$$\begin{bmatrix} 1 \\ 0 \\ .6 \end{bmatrix}, \begin{bmatrix} .8 \\ 1 \\ .8 \end{bmatrix}, \begin{bmatrix} .6 \\ .7 \\ 1 \end{bmatrix},$$

where the added constraint is now met. In this example, the network must block the generalization that is made from $\mathbf{s}_2$ to $\mathbf{s}_1$. In general, the ability to block generalizations in this manner implies the need for a nonlinear output function.

As further constraints are added, there are fewer generalizations across nearby states that are allowed, and performance becomes less parallel. Minimal parallelism will arise when neighboring actions specify conflicting values on all output units, in which case the performance will be strictly sequential. Maximal parallelism should be expected when neighboring actions specify values on non-overlapping sets of output units. Note that there is no need to invoke a special process to program in the parallelism. Essentially, the system generalizes naturally across similar state vectors, and given that state vectors nearby in time are similar, the generalizations act so as to spread actions in time. In most cases, it will be more difficult for the system to learn in the more sequential case when there are more constraints imposed on the system which block the generalizations. These observations are summarized in Figure 4, which shows the relationships between constraint vectors and parallelism.

ATTRACTOR DYNAMICS

The properties of the system that lead to parallel performance also make the system relatively insensitive to perturbations. Suppose that the system has learned a particular sequence and that during performance of the sequence the state is perturbed somewhat. Given that similar states tend to produce similar outputs, the output of the system will not be greatly different from the unperturbed case. This would suggest that the network will perform a sequence which is a "shifted" version of the learned sequence. However, a stronger property appears to hold: The learned sequences become attractors for nearby



Figure 4: Relationships between constraint and parallelism.

regions of the state space and perturbed trajectories return to the learned trajectories. This property is demonstrated in Figures 5 and 6. A network with two output units learned to follow a square in the two-dimensional space which corresponds to the activations of the output units. As shown in the figures, when the network was started at other points in the space, the trajectories moved toward the square. This occurred whether the trajectories began inside or outside the square, showing that the square is a *limit cycle* for the system. For a dynamical system to have limit cycles, it is necessary that the system be nonlinear (Hirsch & Smale, 1974), which further demonstrates the need for the output function to be nonlinear.

More globally, a network which has learned to produce several different cyclical sequences may have several regions of the state space which are attractor basins for the learned cycles. If the network is started in one of these basins, then the performed trajectory will approach the learned cycle, with the part of the cycle which first appears depending on where in the basin the network is started relative to the configuration of the cycle. The network can be regarded as a generalization of a content-addressable memory (cf. Hopfield, 1982) in which the memories correspond to cycles or other dynamic trajectories rather than static points.

Constraints on the output units in general define regions through which trajectories can pass. The network is free to choose a particular trajectory within the region, and this tends to be done in a way so as to avoid sharp changes in the trajectory. Whatever trajectory is chosen by the network, it will tend to generalize so as to become an attractor

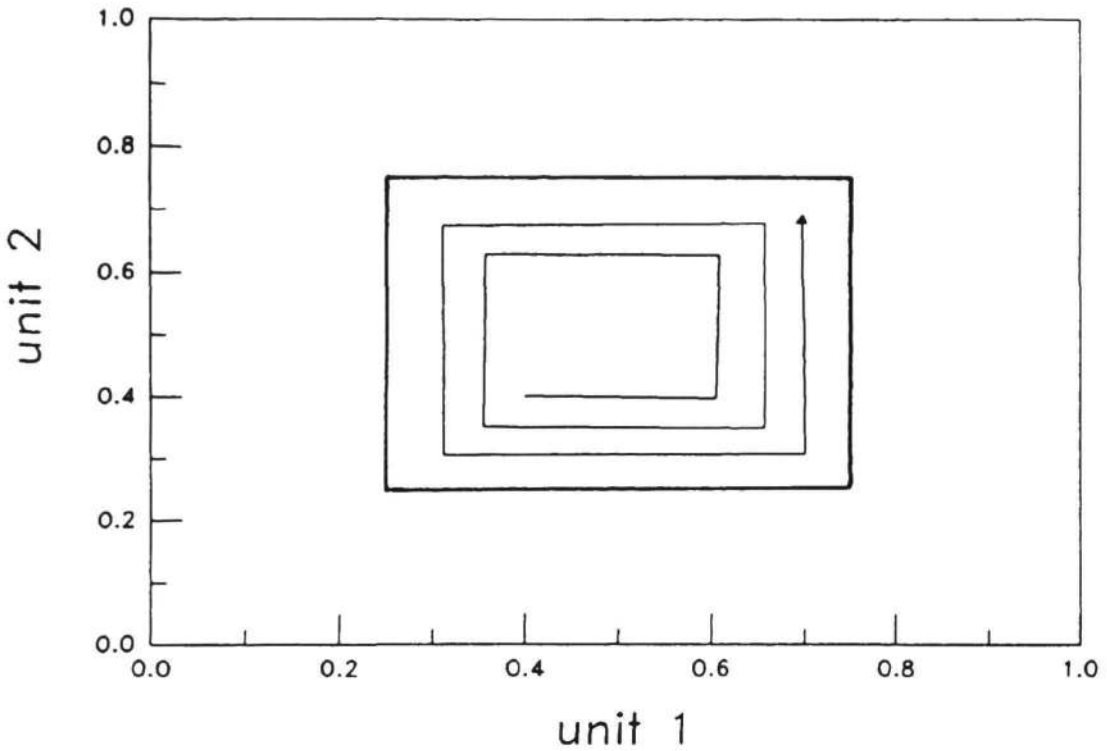


Figure 5: The activations of the two output units plotted with time as a parameter. The square is the trajectory that the network learned, and the spiral trajectory is the path that the network followed when started at the point (.4, .4).

for the surrounding space.

APPLICATIONS TO SPEECH PRODUCTION

In the case of speech, the constraint lists used by the learning process can be taken to encode knowledge about the phonetic structure of the language, and it is natural to identify these constraint lists with phonemes. Thus, in the current framework, the role of phonemes is to constrain the dynamical process that produces utterances by changing parameters of the process until the constraints are met. The constraints that define phonemes are themselves independent of context: They specify in what ways a phoneme can be altered by its context, without specifying values for particular contexts. During the learning process, parallel interactions between nearby phonemes can arise as long as they do not

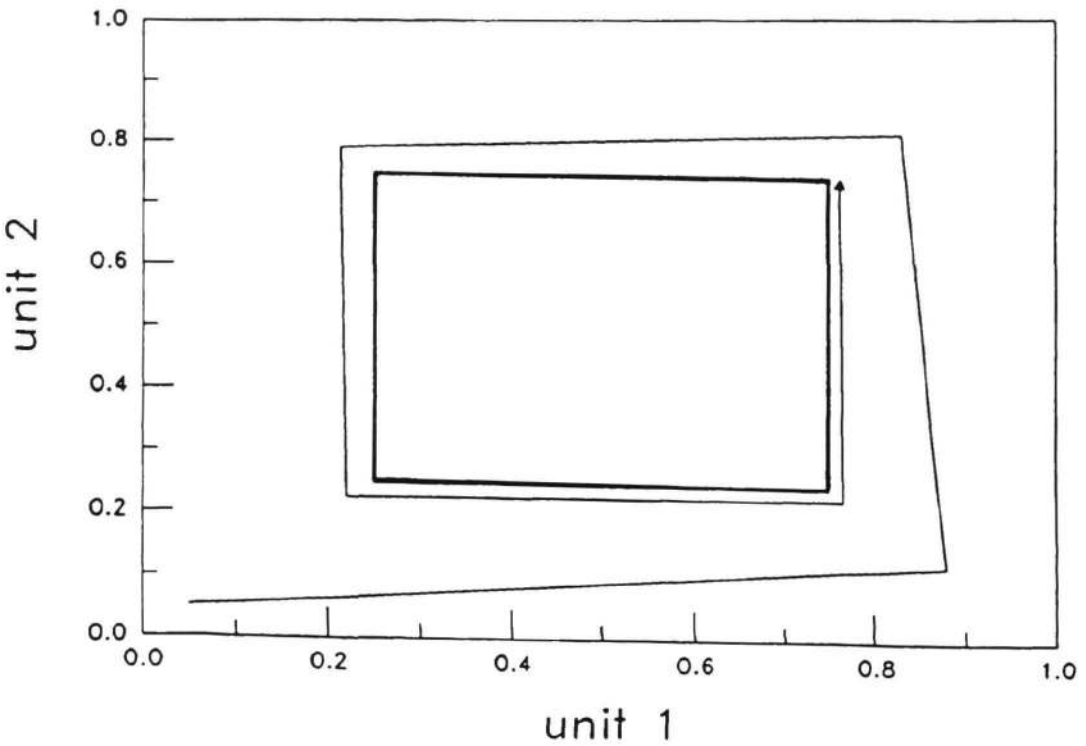


Figure 6: The activations of the two output units plotted with time as a parameter. The square is the trajectory that the network learned, and the spiral trajectory is the path that the network followed when started at the point (.05, .05).

violate the constraints.

I have elsewhere presented simulations that show that the network can mimic coarticulation data such as those presented earlier (Jordan, 1986). Several predictions were also made on the basis of these simulations. The simulations show that there can be non-adjacent interactions, so that, for example, the degree of anticipation of a feature can depend on what follows the feature. It is also the case that there is more coarticulation in the simulation over strings with homogeneous phonemic structure than over strings with heterogeneous phonemic structure.

Finally, it should be noted that it is consistent with the current approach to treat the state equations as discrete versions of a continuous process. In this case, the constraint vectors can still be applied at discrete epochs during learning. Thus, the approach would seem to have some promise for resolving some of the theoretical problems that arise at the interface between discrete phonemic representations and continuous articulatory processes

(Fowler, 1980).

DISCUSSION

One of the important problems that arises in the temporal domain is that there can be interactions both forward and backward in time. One approach to this problem is to represent actions explicitly in a spatial buffer, use relaxation techniques to allow interactions between buffer positions, and then map space into time by gating connections between actions (Feldman & Ballard, 1982). The present paper demonstrates a second approach. In the proposed network, there is no explicit representation of temporal order and no explicit representation of action sequences. There is only one set of output units for the network, therefore output vectors must arise as a dynamic process. Representing actions as distributed patterns on a common set of processing units has the virtue that partial activations can blend together in a simple way to produce the output of the system. Likewise, the representation of states as distributed patterns on a single set of units has the advantage that similarity between states has a natural functional representation in terms of the overlap of patterns. It is the similarity between nearby states that is responsible for interactions in time and this similarity has no time arrow associated with it, so that forward and backward interactions are both possible.

REFERENCES

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science, 9*, 147-169.

Barto, A. G. & Anandan, P. (1985). Pattern recognizing stochastic learning automata. *IEEE Transactions on Systems, Man, and Cybernetics, 15*, 360-375.

Benguerel, A.- P. & Cowan, H. A. (1974). Coarticulation of upper lip protusion in French. *Phonetica, 30*, 41-55.

Booth, T. L. (1967). *Sequential machines and automata theory.* New York: Wiley.

Duda, R. O. & Hart, P. E. (1973). *Pattern classification and scene analysis.* New York: Wiley.

Feldman, J. A. & Ballard, D. H. (1982). Connectionist models and their properties. *Cognitive Science, 6,* 205-254.

Fowler, C. A. (1980). Coarticulation and theories of extrinsic timing. *Journal of Phonetics, 8,* 113-133.

Henke, W. L. (1966). *Dynamic articulatory model of speech production using computer simulation.* Massachusetts Institute of Technology.

Hinton, G. E. & Anderson, J. A. (Eds.) (1981). *Parallel models of associative memory.* Hillsdale, NJ: Erlbaum.

Hirsch, M. W. & Smale, S. (1974). *Differential equations, dynamical systems and linear algebra.* New York: Academic Press.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Science, 79,* 2554-2558

Jordan, M. I. (1986). *Serial order: A parallel, distributed processing approach.* (Technical Report 8604). La Jolla, CA: University of California, San Diego, Institute for Cognitive Science.

Kohonen, T., Lehtio, P. & Oja, E. (1981). Storage and processing of information in distributed associative memory systems. In: G. E. Hinton and J. A. Anderson (Eds), *Parallel models of associative memory.* Hillsdale, NJ: Erlbaum.

Moll, K. L. & Daniloff, R. G. (1971). Investigation of the timing of velar movements during speech. *Journal of the Acoustical Society of America, 50,* 678-684.

Norman, D. A. (1981). A psychologist views human processing: Human errors and other phenomena suggest processing mechanisms. In: *Proceedings of the Seventh IJCAI.* Vancouver, BC.

Rosenblatt, F. (1961). *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms.* Washington, D.C.: Spartan Books.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal represen-

tations by error propagation. In: D. E. Rumelhart and J. L. McClelland (Eds.), ***Parallel distributed processing: Explorations in the microstructure of cognition, Vol.1: Foundations.*** Cambridge, MA: Bradford Books/MIT Press.

Rumelhart, D. E. & McClelland, J. L. (1986). ***Parallel distributed processing: Explorations in the microstructure of cognition, Vol.1: Foundations.*** Cambridge, MA: Bradford Books/MIT Press.

Rumelhart, D. E. & Norman, D. A. (1982). Simulating a skilled typist: A study of skilled cognitive-motor performance. *Cognitive Science, 6,* 1-36.

Widrow, B. & Hoff, M. E. (1960). Adaptive switching circuits. *WESCON Convention Record Part IV,* 96-104.

The Representation of Objects for Grasping

Thea Iberall
Laboratory for Perceptual Robotics
Department of Computer and Information Science
University of Massachusetts

ABSTRACT.

As the human hand reaches out to grasp an object, it preshapes into a shape suitable for the anticipated interaction. As it gets close to the object, it encloses it. Behavioral studies have shown interactions between grasping components and arm transport components. Biomechanical studies look at human hand postures, while neurophysiologists try to suggest how those postures are formed and controlled. In order to evaluate such studies, models are needed which can be used to predict grasping behavior characteristics from a few basic object properties. By looking at the action-perception cycle in primate hand movement, we hope to gain insight into neocortical organization, thus being able to suggest algorithms which could also be at work at all levels of human intelligence.

INTRODUCTION

> *"Our hands become extensions of the intellect, for by hand movements the dumb converse, with the specialized fingertips the blind read; and through the written word we learn from the past and transmit to the future."* – Sterling Bunnell

Over the last 30 years, artificial intelligence research has been trying to capture key aspects of higher cognitive thought [e.g., Shank & Colby 1973]. In the human brain, complex skills such as language, conceptual thinking, and planning have evolved due to the information processing capability of the neocortex [see, for example, Eccles 1977, Arbib 1972]. At the low end of the cerebral cortex, in terms of its distance from the periphery, sits motor cortex. In the neurophysiological literature, it has been noted that the control of fine, fractionated finger movements [Kuypers 1973] is permanently impaired in monkeys, apes, and man by ablation of motor cortex [Denny Brown 1966]. Also, Brinkman and Kuypers [1972] have shown that after a pyramidotomy (which cuts the direct pathway from motor cortex to the motoneurons), a monkey cannot grasp small objects between its fingers or make isolated movements of its wrist. As a way of gaining insight into neocortical organization, we therefore look at primate hand movement, with the goal of being able to

suggest algorithms which could also be at work at all levels of human intelligence.

We focus our studies on one specific type of hand movement, that of *prehension*, or the grasping of objects. As a movement, we note its complexity, which is in part due to it involving some level of object recognition. While objects present to the central nervous system (CNS) an infinite number of sensations, only a few properties seem to affect prehensile movements. The efficacy of the action perception paradigm [Neisser 1976, Arbib 1972] has been demonstrated through artificial intelligence research into attention focusing mechanisms for goal directed behavior [see, for example, Barr & Feigenbaum 1981]. We therefore direct our research to determining what properties are being perceived, and how the CNS might be translating them into goals for prehensile movements.

In terms of movement, joints between bones provide one or more degrees of freedom, or *dofs*. Active movement of the hand is caused by intrinsic muscles within the palm, and also by extrinsic muscles in the forearm, which send tendons through the wrist into the digits. When the human hand is used in prehension, particular styles of postures have been noted [Cutkosky and Wright 1986, Lyons 1985, Iberall & Lyons 1984, 1985, Landsmeer 1962, Napier 1956, Schlesinger 1919]. The classifications of Schlesinger [1919] and Napier [1956] continue to be frequently used in the literature. Schlesinger [1919] noted 6 types of postures, including palmar and tip prehension, cylindrical and spherical prehension, hook prehension, and lateral pinch. Napier defined precision grip and power grip as basic prehensile postures, with the hook grip as a non-prehensile posture for grasping. In our own first attempt at a prehensile classification [Iberall & Lyons 1984, 1985], we suggested six postures, using such terms as basic precision, basic power, basic precision/power, modified power, modified precision/power, and fortified precision/power. In that attempt, we were searching for a way to capture the fact that hand postures are not as discrete as most classifications suggest. Objects are not as regular as Schlesinger noted, and yet the hand can pick up a vast range of object shapes. Forces are subtly applied at multiple places around an object, indicating that more than one type of grip is being used at a time, instead of the either/or nature of these classifications. Even Napier stated that 'although in most prehensile activities either precision or power is the dominant characteristic, the two concepts are not mutually exclusive' [Napier 1956, p 906].

Stereotypical descriptions of this sort suggest that there are anatomical and physiological constraints in the hand-object interaction, which the CNS may in turn use. These could include stability constraints, such as postures involving stress limiting coadaption of articular surfaces [Kapandji 1982, Chao *et al* 1976], and leverage maximizing wrist positions [Hazelton *et al* 1975, Pryce 1980, etc]. Also at issue is the use of the frictional characteristics of the skin [Thomine 1982], and the location of tactile receptors [Johannson & Vallbo 1983], both of which are notably specialized at the finger pads [Glicenstein &

Dardour 1981].

However, while these classifications are useful as broad general descriptions, they do not provide a useful mechanism for describing the complete grasping process, from perceiving an object, to capturing and manipulating the object. We feel that a classification is needed which describes a prehensile posture in goal-directed terms; i.e., in terms of the number of forces needed in the task, in what direction they are to be applied, and at what strength must they be applied. We define the goal of prehension to be *the bringing to bear of functionally effective forces around an object for a given task given the hand's anatomical constraints.* The style of posture chosen by the CNS must then match the task requirements (the forces and *dofs* needed in the task) with the hand's capabilities (the forces and *dofs* available at the hand).

FUNCTIONAL ANALYSIS OF PREHENSION

Prehension involves applying opposition to the forces arising during interaction with an object, and stable prehension adds the requirement that opposition is applied by hand surfaces against other hand surfaces [Iberall, Bingham & Arbib 1985]. We call the applications of forces between hand surfaces *oppositions.* In order to apply those forces, the hand preshapes into a suitable posture during the reaching phase of the movement [Jeannerod 1981], thus setting up the required oppositions for the task. The fingers and thumb extend, creating a space within the hand, with characteristics which are highly task dependent. An obvious example of task dependency deals with object width: the larger the object width, the larger the maximum aperature between the thumb and fingers [Jeannerod 1981]. Enclosing the hand around an object works within that area defined by the preshaped hand, by increasingly flexing the fingers and thumb until they contact the object and can then apply the required force.

In studying these data, we ask ourselves what is the subject perceiving about the object that determines a particular hand posture as being functionally effective for the stated task. We call that perception the *task requirements*, and it is our goal to be able to describe prehensile movements in terms of those task requirements.

To gain some intuition for this analysis, we initially observed [Arbib, Iberall & Lyons 1985] the task of picking up different size mugs by their handles. We noted that three basic forces were being applied: one to provide a downward force from above the handle, one to provide an upward force from within the handle and, if necessary, a third force to stabilize the handle from below. We hypothesized that each of these functions could be represented as the task of a *virtual finger.* Real fingers move in conjunction within a virtual

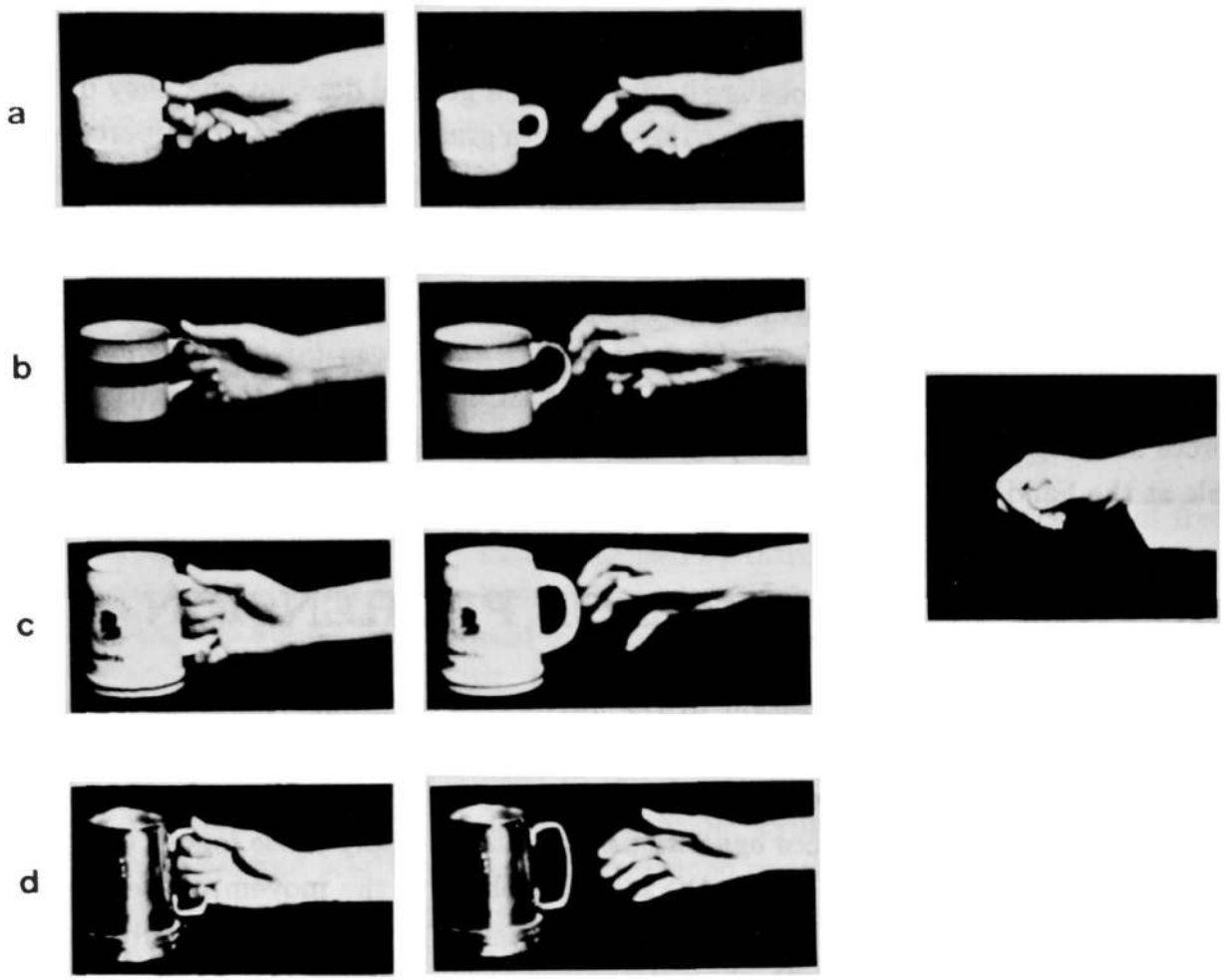


Figure 1: Grasping mug task. The number of real fingers which map into the second virtual finger varies, depending on the size of the handle.

finger, which has the same characteristics as real fingers. This both limits the *dofs* to those needed for a given task, and provides an organizing principle for task representation at higher levels in the CNS. It is then a subtask at a lower-level to perform the actual mapping to real fingers, making task implementation somewhat "tool-free".

Figure 1 shows an example of the use of virtual fingers. In the four examples, the thumb, mapped into the first virtual finger or VF1, provides a force from above the handle. For a teacup with a very small handle (Figure 1a), only one finger will fit inside the handle. During the preshaping, only the index finger is mapped into the second virtual finger VF2, which provides an upward force from within the handle, and which opposes the force applied by VF1. The rest of the fingers map into VF3, which provides support for cup stabilization. For a coffee mug of the kind pictured in Figure 1b, two fingers will fit

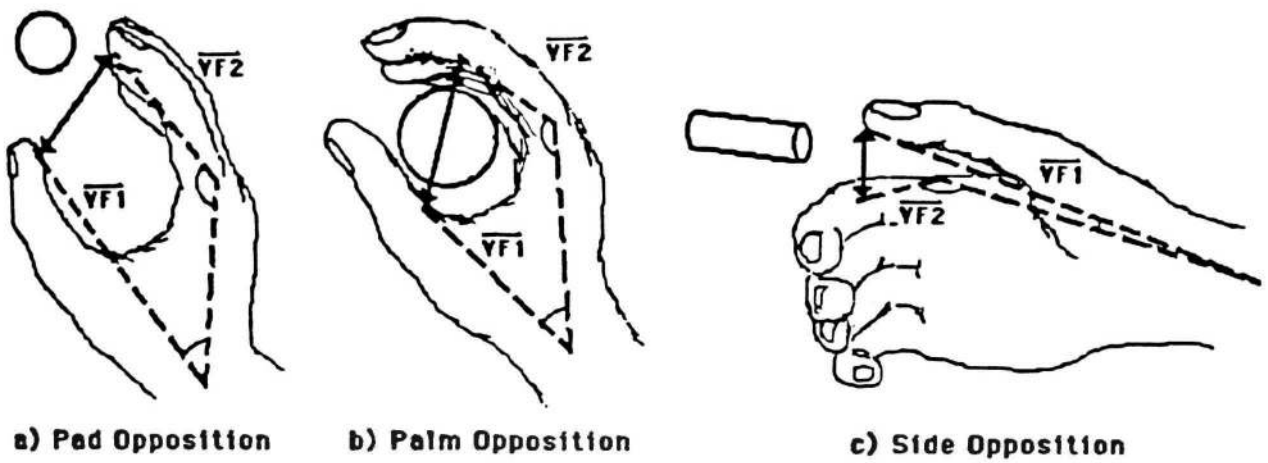


Figure 2: Basic oppositions (a) Pad opposition along axis generally parallel to palm (b) Palm opposition along axis generally normal to palm (c) Side opposition along axis generally transverse to palm. Solid lines show axis of opposition, dashed lines show virtual finger vectors $\overline{VF1}$ and $\overline{VF2}$.

within the handle to form VF2, with the other two fingers becoming VF3. For the mug of Figure 1c, VF2 will comprise three fingers, while Figure 1d demonstrates the case in which all four fingers are mapped into VF2, with an empty mapping to VF3.

We suggested in [Iberall, Bingham & Arbib 1985] that prehensile postures are constrained by the way the hand can apply opposing forces around an object for a given task, and noted that there are three basic ways that the hand can apply these oppositions. *Pad opposition* (Figure 2a) between the finger pads (VF2) and the thumb pad (VF1) occurs along an axis roughly parallel to the palm. Opposition at the digit pads offers greater flexibility in finely controlled manipulations of an object at the expense of stability and maximum force. These pads are highly specialized for prehension, in that they provide friction, due to the epidermal ridges, sticky excretions, and their ability to comply [Glicenstein & Dardour 1981]. They also have numerous tactile receptors, more so than most other parts of the body, thus giving the CNS much information about the object with which they come into contact [Johannson & Vallbo 1983]. *Palm opposition* (Figure 2b) between the digits (VF2) and the palm (VF1), sacrifices flexibility in favor of stability. Essentially the object is fixed in hand coordinates, along an axis roughly normal to the palm of the hand. Greater stability is achieved by a combination of factors [Thomine 1982]: the larger palmar surface area providing more friction, greater forces available from proximal finger areas [Hazelton *et al* 1975, Chao *et al* 1976], the ability to passively cancel

torques, and the use of the thumb against the fingers for adding force and changing the size of the palm. As more fingers are included in VF2, the magnitude of the applicable force increases. Finally, ***side opposition***, either between the thumb pad (VF1) and the side of the index finger (VF2) (Figure 2c), or else between the sides of the fingers, is a compromise between flexibility and stability. Its opposition axis occurs primarily along a transverse axis to the palm. The thumb, due to the orientation of its articular surfaces, has its pad oriented toward the sides of the other digits, giving side opposition using the thumb the extra frictional component. Side opposition between fingers is stronger (but less flexible) if the object is held proximally in the fingers, and weaker (but more flexible) if held more distally.

Each opposition has particular features which are related to the hand's anatomical and physiological structure, and thus define the nature of applicable forces and usable *dofs*. We suggest that the shape the hand takes on during the grasp reflects the use of one or more of these oppositions. In the grasp mug task, side opposition is used between the thumb and index (see Figure 1), and palm opposition occurs between the palmar surfaces of the fingers. Since the mug is grasped off-center from its center of mass, VF3 applies an opposition to the torque which brings the mug into the hand. In general, a complete functional description of each opposition would discuss the direction of the applied force, the magnitude of the applied force as related to the number of real fingers involved (in terms of minimal, maximal, and mean forces), the direction of available movement (*dofs*), the range of available movement, and the amount of control over those movements. A complete functional description of prehensile movement would show the how these oppositions are seen and used in the emergent postures.

OPPOSITION SPACE

Using the concept of virtual fingers, and noting how the hand can apply forces through oppositions, we defined *opposition space* [Iberall, Bingham & Arbib 1985] as the coordinates within which opposition can take place between virtual fingers; specifically, opposition space defines *the functional capabilities of the hand for executing stable grasps and object manipulations*. To completely characterize opposition space, we specify here:

1. a description of the task-relevant object features, or *task requirements*
2. a coordinate frame with relevant grasping parameters, or *opposition space*
3. a mapping between the task requirements and opposition space

Task Requirements

While it is not our goal to model how the CNS might do it, we suggest that perception of the object involves extracting a set of properties relevant to the person's *intent* for the use of the object in the task. The perceived task requirements, which help determine the shape of the hand, consist of both functional constraints and physical constraints. Functional constraints, coming from what must be accomplished in the task (the intent), include issues such as: don't drop the object, (stably) manipulate the object, and overcome anticipated forces. Physical constraints, based on object properties and hand properties, include such issues as properties of opposing surfaces, available forces, and available *dofs*. As a first approximation toward quantifying these, a working set of task requirements is as follows:

- a set of opposable surfaces
  - length of each opposable surface
  - radius of curvature of each surface
  - surface compliance
- an opposition vector between pairs of opposable surfaces
  - magnitude of opposition vector
  - orientation of opposition vector (relative to a vector between the opposition space origin and the center of the opposition vector)
- functional degrees of freedom about an opposition vector
  - direction of movement
  - range of movement
  - resolution of control
- a set of anticipated forces
  - object weight vector
  - inertial force vector
  - torques at the opposition vector

Objects have infinitely many properties, but we suggest that for prehensile movements, an object is perceived by its opposable surfaces, and only those relevant to the task. Between every two relevant opposable surfaces exists an opposition vector. The mug, for example,

having a cylindrical body and a curved cylindrical handle, has an infinite number of possible opposition vectors between the infinite number of points along its surfaces. But, due to its construction and due to the person's intent of drinking from the top opening, it provides at least three solutions in terms of task requirements:

1. an opposition vector, between body surfaces slightly above the center of mass, will allow the task-related rotation (i.e., drinking from the opening) of that opposition vector while stably overcoming the object's weight vector. By choosing the particular opposition vector which has its origin near the proximal surface of the mug and placing the thumb there, the object's (changing) weight vector will be opposed by the thumb as the mug is tilted to the lips.

2. an opposition vector, between the top and bottom body surfaces, will also allow the task-related rotation. It is, however, off the center of mass, and has smaller opposable surfaces, and is usually blocked by other obstacles, such as the table. (This of course focuses on the fact that other issues, missing in this first approximation, are at work; these include object temperature, obstacles, etc).

3. two opposition vectors, one through the top of the handle and one through the side of the handle, will also allow the task-related rotation. By choosing two particular opposition vectors, one with an origin on the top of the handle and one on the distal surface of the handle, the object's (changing) weight vector can be overcome by accounting for the torque caused by being off the center of mass. While these opposition vectors are off the center of mass and the surfaces smaller, they provide a solution to the temperature problem.

In our example of Figure 1, when the subject was told to 'Pick up the mug to drink out of it', she chose the third solution in each case.

Coordinate Frames

In order to now match the task requirements to the hand's capabilities, we define here a coordinate system for the three oppositions, as shown by the dashed lines in Figure 2. In pad opposition (Figure 2a), $\overline{VF1}$ describes where the thumb pad is, as it extends and flexes through its 'workspace', relative to the palm. The virtual finger vector $\overline{VF2}$ similarly describes where the finger pad is, relative to the palm. These vectors increase and decrease in length and orientation as the thumb and fingers extend and flex. The dark line indicates the opposition vector along which the opposition is occuring. For palm opposition (Figure 2b), the virtual finger vectors are initially shorter, ending on the palmar

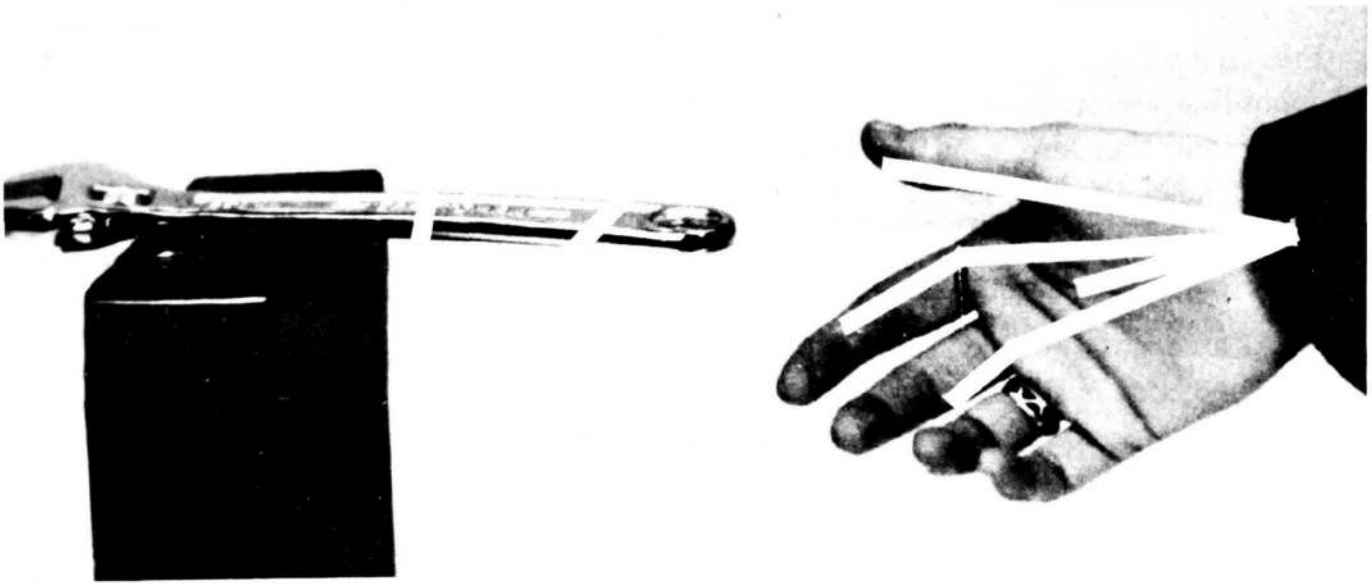


Figure 3: Composite grasp involving palm and side opposition. Grasping medium-sized wrench involves both power and control components.

surfaces where the opposition is taking place. In side opposition (Figure 2c), $\overline{VF2}$ stops at the proximal interphalangeal joint of the index finger.

An example of these coordinate systems being used in a composite grip is seen in Figure 3. The object is perceived with two opposition vectors, one matching to the thumb (VF1) and index finger (VF2) in side opposition along the transverse axis, and the other matching the palm (VF1) and the four fingers (VF2) in palm opposition.

Mapping from Task Requirements to Opposition Space

The general problem that the CNS has in an action/perception paradigm [Neisser 1976, Arbib 1972] such as grasping an object, is in transforming perceived object properties into useful motor spaces. We use schema theory [Arbib 1981, Neisser 1976] to describe perceptual structures and units of motor control. To focus attention, perceptual schemas in the CNS must somehow extract relevant information about that object, in this case, a reasonable opposition vector (or set of opposition vectors) for the particular task, transformed from retinotopic coordinates into a three dimensional space. The Preshape motor schema will map the current hand posture (which also must be perceived) into a suitable opposition space, from where the Enclose schema will bring the hand actually around the

object.

This problem can be defined as an inverse kinematic problem, in the sense that the Preshape schema must determine the correct angles for $\overline{VF1}$ and $\overline{VF2}$ given a particular opposition vector. However, the inverse kinematics do not necessarily give a unique solution, and therefore, a further analysis of the goals is needed. We can also ask what is the intended direction for the application of the forces (i.e., what is a solution of the task requirements which accounts for stably overcoming the anticipated forces), and what are the forces available from various hand postures.

As an example toward looking at the first question, we can look at what happens in a pad opposition involving the thumb (VF1) and index finger (VF2). The intended direction of force application is to apply forces through the index finger and thumb pads along the opposition vector. As a first approximation, we can just look at the direction of a normal through the index finger and one through the thumb pad, and see how they vary as the finger and thumb move within their anatomically-constrained 'workspace'. We measured the workspace of the index finger pad and that of the thumb pad for one subject, and from sample data points, we did a linear interpolation to produce the plot seen in Figure 4. Each arrow shows the direction of a normal with respect to the pad, when the thumb vector $\overline{VF1}$ (on the left side of the figure) or the index finger vector $\overline{VF2}$ (on the right side of the figure) is at a given angle and length. The dots are areas outside the workspaces. The abscissa values are in opposite direction, thus graphically showing how the normals become colinear but opposite at various locations in the workspaces.

Such a description of the workspaces of VF1 and VF2 provides a first attempt at a functional analysis of hand capability. If, somewhere in the CNS (in cortical motor and/or premotor areas), such a representation of the hand existed, it would act as both a motor map, providing a goal for lower level systems to achieve through proprioceptive feedback, and also as a cognitive map, providing a description of the hand's functionality in that part of opposition space. The Preshape Schema could look into this map for an orientation of the pad normals which would satisfy the magnitude and orientation of the opposition vector. However, the hand must be positioned at a distance and orientation from the opposition vector, along an approach vector, where that solution exists. There are, however, an infinite number of approach vectors, in terms of the distance at which the maximum aperature between VF1 and VF2 occurs and the approach angle at which it happens. How might the Preshape Schema decide which solution is best, or at least reasonable?

The selection of this solution can be done using a competition model involving computations in a neural network [Amari & Arbib 1977]. Of all the given inputs to such a network, the node receiving the maximum input is selected. In our case, for a given op-

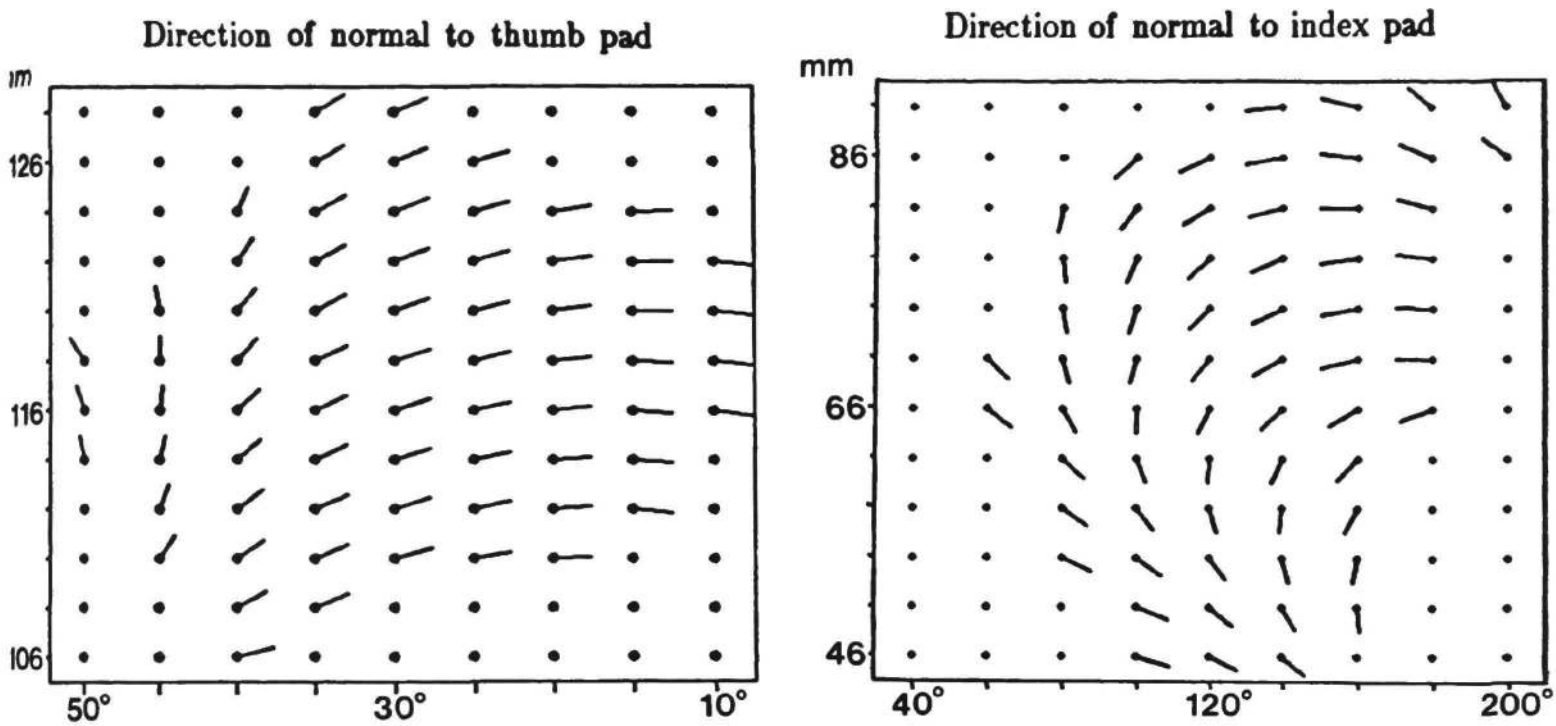


Figure 4: Direction profiles for pad opposition of thumb pad normals (left figure) and index finger normals (right figure). The abscissas, opposite in direction, represents the angle between $\overline{VF1}$ or $\overline{VF2}$ and the palm. On the ordinate is the length of the virtual finger vector.

position vector magnitude, all possible distances and orientations of the opposition vector relative to the opposition space origin would compete in an excitation/inhibition network. Distances and orientations outside the range of opposition space would die out. VF1 and VF2 configurations which correspond to colinear pad normals would reinforce each other's excitation level, thus allowing peaks in the network model to grow. A competition model of this type is currently being developed, and should show how the mapping of task requirements into opposition space constrains various aspects of prehensile movements, such as the approach vector the arm takes in reaching.

The actual posture chosen by the Preshape Schema would also incorporate a small buffer zone, the size of which is dependent on many task features such as Fitt's Law (for a review see [Keele 1981]). This issue, however, is beyond the scope of this paper.

The second question looks at what forces are available from various hand postures. In this broader issue of mapping from task requirements to opposition space, we have yet

to determine how do interactions of task requirements lead to interactions of oppositions and to real-to-virtual finger mappings, producing observed postures. We suggest that perceptual schemas extract an opposition vector at each place where an opposition is needed in a task, along with all the necessary task requirements. As was noted, the grasp mug task is an example of multiple oppositions.

As an example toward answering this second question, we look at the task that was seen in Figure 3, where the person is reaching for a medium-sized wrench. The fact that there are two opposition vectors anticipates the need for two types of oppositions, thus capturing the two characteristics of the task (i.e., power and control). The lengths of opposable surfaces for each of the opposition vectors are longer than the hand is wide, and so do not present a virtual finger size constraint. The width of the opposition vectors are not particularly big, indicating that the object is not particularly heavy. The anticipated task torque (tightening the nut) indicates the need for palm opposition, using as many fingers as possible in VF2. However, during the first part of the task (fitting the jaws of the wrench around the nut), some control is needed. This, along with the fact that only coarse resolution of control over the task *dofs* (turning the wrench) is needed, indicates that side opposition would be more useful than pad opposition for the second opposition, with the index finger mapped into VF2 and the thumb into VF1. This allows all four fingers to map into VF2 and the palm into VF1 for the palm opposition.

CONCLUSION

Research in artificial intelligence has demonstrated the efficacy of attention focusing mechanisms for goal directed behavior. In terms of natural intelligence, the CNS is presented with an infinite number of sensations when dealing with even one object. In order to successfully interact with such an object, only information specific to the interaction is desireable. Important perceptions for prehensile movements involve: what surfaces can be grasped given a particular intention, what are the characteristics of those surfaces, and what will happen when they are grasped. In attempting to capture this information in a computational model of prehensile movements, we call these features the *task requirements* and list them, in a first approximation, as a set of opposable surfaces, an opposition vector between pairs of opposable surfaces, the functional degrees of freedom about an opposition vector, and a set of anticipated forces arising relative to an opposition vector.

The human hand, having anatomical and physical constraints, applies oppositions around objects along three basic axes using the fingers in groupings called virtual fingers. Prehensile postures reflect the combined use of one or more of these oppositions. Each type of opposition has particular functional capabilities, involving the direction and

magnitude of the applied force, the direction and range of available movement, and the amount of control possible. We call the collective set of these oppositions the ***opposition space*** of the hand, and use the coordinate system described here to describe the prehensile postures of opposition space. The goal of a preshape schema in the CNS is to map task requirements into opposition space so that functionally effective forces can be brought to bear around a perceived object for a given task. In order to apply those forces, the hand must be positioned at a distance and orientation from the opposition vector where a solution in opposition space exists. A competition model of neural networks would allow possible kinematic solutions to compete, with a solution that is within opposition space being selected when the model converges. Such a model for mapping perceived task requirements into hand functionality presents a mechanism for explaining goal-directed movement in somatotopic, distributed processing terms, while also allowing for the eventual incorporation of learning algorithms.

REFERENCES

Amari, S. & Arbib, M.A. (1977): Competition and Cooperation in Neural Nets. In J. Metzler (ed), *Systems Neuroscience*, NY: Academic Press, 119-165.

Arbib, M.A. (1972): *The Metaphorical Brain. An Introduction to Cybernetics as Artificial Intelligence and Brain Theory*, NY: Wiley-Interscience.

Arbib, M. A. (1981): Perceptual Structures and Distributed Motor Control. In V.B. Brooks (ed.), *Handbook of Physiology–The Nervous System, II. Motor Control*, Bethesda: Amer Phys Soc, 1449-1480.

Arbib, M.A., Iberall, T., & Lyons, D. (1985): Coordinated Control Programs for Movements of the Hand. In A.W. Goodwin & I. Darian-Smith (eds), *Hand Function and the Neocortex*, Berlin: Springer Verlag, 111-129.

Barr, A. & Feigenbaum, E.A. (1981): *The Handbook of Artificial Intelligence, vol I.* Los Altos: William Kaufmann.

Brinkman, J. & Kuypers, H.G.J. (1972): Splitbrain Monkeys: Cerebral Control of Ipsilateral and Contralateral Arm, Hand, and Finger Movements. *Science*, 176(4034): 536-539.

Bunnell, S. (1944): *Surgery of the Hand.* Phil: J.B. Lippincott.

Chao E.Y., Opgrande J,D. & Axmear, F.E. (1976): Three Dimensional Force Analysis of Finger Joints in Selected Isometric Hand Functions. *J Biomechanics*, 9:387-396.

Cutkosky, M.R. & Wright, P.K. (1986): Modeling Manufacturing Grips and Correlations with the Design of Robotic Hands. *Proc 1986 IEEE Intl Conf on Robotics and Automation*, San Francisco, Calif, Apr 7-10, 1533-1539.

Denny-Brown, D. (1966): *The Cerebral Control of Movement*, Liverpool: Liverpool Univ. Press.

Eccles, J.C. (1977): *The Understanding of the Brain, Second Edition*, NY: McGraw-Hill.

Glicenstein J., & Dardour J.C. (1981): The Pulp: Anatomy and Physiology. In R. Tubiana (ed), *The Hand, vol. 1*, Phila: W.B. Saunders, 116-120.

Hazelton, F.T., Smidt, G.L., Flatt, A.E. & Stephens, R.I. (1975): The Influence of Wrist Position on the Force Produced by the Finger Flexors, *J. Biomechanics*, 8: 301-306.

Iberall, T., Bingham G. & Arbib, M.A. (1985): Opposition Space as a Structuring Concept for the Analysis of Skilled Hand Movements, *COINS TR 85-19*, Dept of Computer & Info Science, Univ of Mass, Amherst, Mass. To be published in *Exper Brain Research Supplmt.*

Iberall, T. & Lyons, D. (1984): Towards Perceptual Robotics. *Proc 1984 IEEE Intl Conf Systems, Man, Cybernetics*, Halifax, Nova Scotia, Oct. 9-12, 147-157.

Iberall, T. & Lyons, D. (1985): Perceptual Robotics: Towards a Language for the Integration of Sensation and Perception in a Dextrous Robot Hand. In S. Chang (ed), *Languages for Automation*, NY: Plenum Press, 415-427.

Jeannerod, M. (1981): Intersegmental Coordination During Reaching at Natural Visual Objects. In J. Long & A. Baddeley (eds), *Attention and Performance, IX*, Hillsdale: Erlbaum, 153-168.

Johannson, R.S. & Vallbo, A.B. (1983): Tactile Sensory Coding in the Glabrous Skin of the Human Hand, *Trends in Neuroscience*, 6(1): 27-37.

Kapandji I.A. (1982): *The Physiology of the Joints. Vol One. Upper Limb*, Fifth Edition, Edinburgh: Churchill Livingstone.

Keele, S.W. (1981): Behavioral Analysis of Movement. In V.B. Brooks (ed), *Handbook of Physiology–The Nervous System, II. Motor Control*, Bethesda: Amer Phys Soc.

Kuypers, J.G.J.M. (1973): The Anatomical Organization of the Desceding Pathways and their Contributions to Motor Control Especially in Primates. In J.E. Desmedt (ed), *New Developments in Electromyology and Clinical Neurophysiology*, 3, Basel: Kruger, 38-68.

Landsmeer, J.M.F. (1962): Power Grip and Precision Handling, *Ann Rheum Dis*, 21: 164-170.

Lyons, D. (1985): A Simple Set of Grasps for a Dextrous Hand, *Proc 1985 IEEE Intl Conf on Robotics and Automation*, St. Louis, Missouri, Mar 25-28, 588-593.

Napier, J. (1956): The Prehensile Movements of the Human Hand, *J Bone Jt Surgery*, 38B(4): 902-913.

Neisser, U. (1976): *Cognition and Reality.* San Francisco: W.H. Freeman.

Pryce, J.C. (1980): The Wrist Position Between Neutral and Ulnar Deviation that Facilitates the Maximum Power Grip Strength, *J Biomechanics*, 13: 505-511.

Schank, R. & Colby, K., eds (1973): *Computer Models of Thought and Language*, San Francisco: W. H. Freeman.

Schlesinger, G. (1919): Der Mechanische Aufbau der Kunstlichen Glieder. In M. Borchardt *et al* (eds), *Ersatzglieder und Arbeitshilfen*, Berlin: Springer.

Thomine (1981): The Skin of the Hand. In R. Tubiana (ed), *The Hand, vol 1*, Phila: W.B. Saunders, 107-115.

Tubiana, R. (1981): The Architecture and Functions of the Hand. In R. Tubiana (ed), *The Hand, vol 1*, Phila: W.B. Saunders, 19-93.

A Connectionist Learning Model for 3–Dimensional Mental Rotation, Zoom, and Pan

Bartlett W. Mel

Artificial Intelligence Group
Coordinated Sciences Laboratory
University of Illinois

Abstract

A connectionist architecture is applied to the problem of 3–D visual representation. The **VI**sual **P**erception System (VIPS) is organized as a flat, retinotopically mapped array of 16K simple processors, each of which is driven by a coarsely–tuned binocular feature detector. By moving through its environment and observing how the visual field changes from state to state for various kinds of motion, VIPS learns to run internal simulations of 3–D visual experiences, e.g. mental rotations of unfamiliar objects. Unlike traditional approaches to visual representation, VIPS learns to perform 3–D visual transformations purely from visual–motor experience, without actually constructing an explicit 3–D model of the visual scene. Instead, the third dimension is represented *implicitly* in the knowledge as to how the pattern of activation on its flat sheet of binocularly–driven processors will shift about as VIPS moves, or only imagines moving, through space. VIPS is argued to be more compatible with a variety of phenomena from the psychology of 3–D perception than previous vision systems, particularly with respect to development, plasticity, and stability of perception, as well as the analogical, linear–time mental rotation phenomena.

1. Introduction

A fascinating puzzle in the study of visual perception lies in the mechanisms by which a compelling impression of 3–dimensionality is achieved via two, 2–dimensional sensory ports. Within the field of computer vision, the most common approach to the problem of 3–D vision has been to develop sophisticated algorithms that use the 2–dimensional retinal images to construct explicit, 3–dimensional models of the visual scene [e.g. Marr & Nishihara, 1978].

The **VI**sual **P**erception System (VIPS) described herein is a connectionist architecture that embodies a new and very different approach to 3–D visual representation. Inspired by the architecture of the brain, which consists to a remarkable degree of 2–dimensional, topographic maps of the sensory and motor modalities, VIPS demonstrates how a flat, retinotopically mapped sheet of simple processors, each driven by a coarsely coded binocular feature detector, can to a large degree achieve the *effect* of a 3–D representation without actually constructing an explicit 3–D model of the visual scene. Furthermore, VIPS is argued to be more compatible with a variety of phenomena from the psychology of 3–D perception than previous vision systems, particularly with respect

This work was originally supported by Thinking Machines Corporation, Boston MA, and subsequently by a Hewlett–Packard/AEA Fellowship.

to development, plasticity, and stability of perception, as well as the analogical, linear-time mental rotation phenomena.

Two tenets illustrate the nature of the VIPS approach to 3–D vision. The first holds that the faculty for running *internal simulations* of visual events, or *envisionment*, is of central importance to an intelligent vision system. More concretely, VIPS learns to drive state sequences on its internal, binocular visual map that approximate the state sequences on the *same* map that would be driven *externally*, by the retinae, were VIPS actually moving through its environment and observing the visual changes. A 3–dimensional visual experience is therefore represented as a *sequence* of states on an internal binocular visual map, where each state codes for the instantaneous 3–D surface layout in the visual field. VIPS' representational repertoire consists of any mostly–continuous transformation of the binocular visual array, that is, any transformation in which *most* of the low–level features in the visual field follow smooth trajectories from state to state in a motion sequence. This is to say that in each state, only a small fraction of the features in the visual field should appear or disappear abruptly at an occluding contour. Included in this class of transformations are *zoom*, *pan*, and *rotation* of the visual field along or about any axis in 3 dimensions, as well as arbitrary combinations of these, and optionally, with different sub–portions of the visual field undergoing different transformations. Visual changes brought about by autonomously moving bodies fall within the representational power of VIPS, but cannot yet be learned or predicted. This limitation is discussed in the concluding section with respect to future stages in VIPS' design.

The second tenet holds that no *a priori* knowledge of 3–D visual transformations need be built in. Rather, VIPS is based on the philosophy that an extremely rich source of knowledge about the 3–D world is available to an intelligent vision system that can act on the environment through its "musculature", and observe and record the resulting changes in its sensory stream. In just this way, VIPS learns the relationship between its own state of motion, and the resulting, highly predictable way in which the visual scene changes through time. In lay terms, VIPS embodies a "learn–by–doing" approach to 3–D vision, as dependent on its own state of activity as on the content of the incoming sensory stream.

While VIPS and its constituent elements are being proposed elsewhere as an abstract model for a visual/motor association cortex in the brain [Mel, 1986a], it will be described here only in its capacity as a connectionist architecture applied to the problem of 3–D visual representation, with particular attention to its compatibility with a variety of phenomena in the psychology of 3–D visual perception.

2. The Organization of VIPS

2.1. Connectivity

Structurally, VIPS consists of 16K simple processors, or Contextrons, organized in a flat, retinotopically mapped grid, with each processor receiving its dominant input, and therefore its "meaning", from a coarsely tuned binocular feature detector that responds optimally to a short, physical micro–feature at one of 4 orientations and 4 depths in the visual field. This type of visual feature detector is purposely analogous in design to the binocular cells in the mammalian visual cortex [e.g. Hubel & Wiesel, 1979], where the depth is given by the horizontal disparity between the Contextron's left and right–eye receptive fields. In addition to this powerful retinally–derived input, a Contextron receives a weighted, excitatory feedback connection from each of its neighbors within some fixed radius, allowing the *current* state of activity in its neighborhood to influence its own *next* state. Finally, each Contextron receives a motion–context input from VIPS' controlling "motor center".

2.2. The Contextron

The Contextron is the simple connectionist processor out of which VIPS is built, essentially performing the function of a multiplexer (fig. 1). In each state, a context field selects some subset of a Contextron's

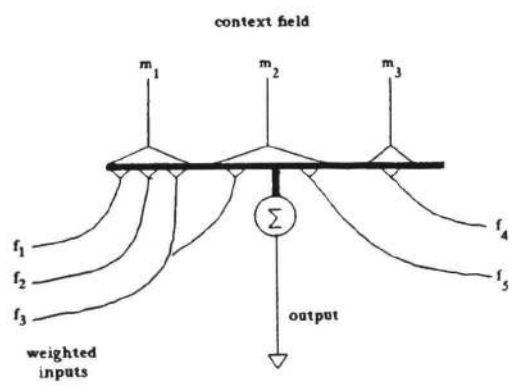


Figure 1. The Contextron.

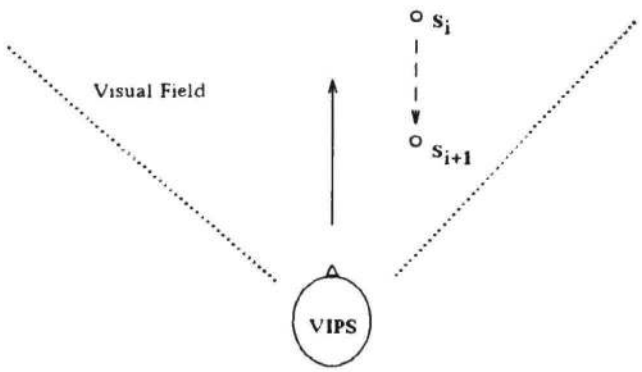
weighted inputs to be gated to the output[1]. Its function as a multiplexer is most easily understood for the degenerate case in which the context field enables only a *single* weighted input. In this case, the Contextron simply gates the single input directly to the output through a weighting factor. In general, however, a context field will usually enable not one, but some small number of weighted inputs. Crucially, each of the enabled inputs in a given context state is signalling the presence of the *same* physical event E, where the weight for each input is a measure of its historical reliability in reporting E. The Contextron can therefore be thought of as taking a weighted "poll" of its selected inputs in order to determine the probability that event E occured. In VIPS, the subset of weighted inputs that is enabled in a given context state will tend to originate from a localized *cluster* of the Contextron's neighbors, and will be seen to signal the imminent incursion of a physical micro-feature into its own receptive field.

2.3. Theory of Operation

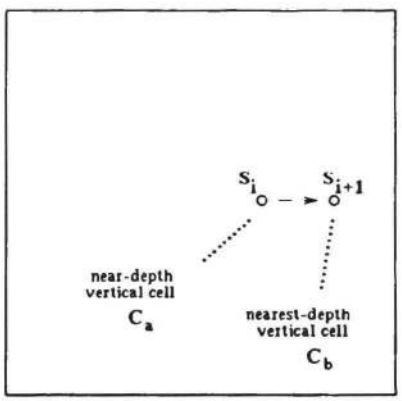
The operation of VIPS depends most directly on the assumption that within a given motion-context, such as "move forward", *most* of the physical features in the visual field will follow smooth, predictable trajectories relative to the moving observer. This is identical to the continuity of flow assumption detailed by Marr [1982].

Consider F, a short, vertically oriented feature in the near-right visual field, and C_a, the Contextron that is most strongly activated by F. If VIPS is moving smoothly forward in space, then the same physical feature that is exciting C_a in state i will move deterministically into a new position relative to VIPS, depending only on the rate of forward motion (fig. 2a). In state $i+1$ therefore, the *same* physical feature will be somewhat less distant and further to the right in the visual field, and will therefore excite a different Contextron, C_b (fig. 2b). Assuming the feature has not had time to travel far across the retina in a single state in the motion sequence, then C_b will be a relatively close neighbor to C_a, within the radius of the feedback paths. Under the continuity-of-optic-flow assumption, it is clear that in the context of forward motion, C_a is an excellent predictor of the next state of C_b. This fact is reflected by a heavily weighted feedback connection from C_a to C_b, for the context of forward motion (fig. 2c). As a useful abstraction, C_a may be given the special title *Probabilistic Physical Predecessor* (PPP) of

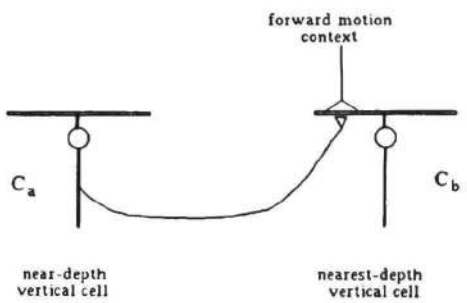
[1] In the current implementation, the output is computed as a simple weighted sum of the selected inputs, but the particular transfer function is not of immediate importance to the model.



(a) Visual change with forward motion.



(b) Internal binocular state.



(c) Resulting feedback connection.

Figure 2. (a) As VIPS moves forward, a vertical feature in its near-right visual field moves closer and further to the right. (b) This same vertical feature stimulates a vertically oriented *near*-cell in state i, followed by a vertically oriented *nearest*-cell in state $i+1$. (c) Since C_a tends to predict the activity of C_b in the context of forward motion, a strong feedback connection will develop for that context.

C_k for this "forward" motion–context, though because the feature detectors driving VIPS are coarsely tuned, a small *population* of Contextrons clustered around the PPP will also develop connections to C_k. In the context of *reverse* motion, C_k should receive strongly weighted connections from the opposite neighborhood, that is one that codes for a physical feature even further to the right in the visual field and still nearer.

For each motion–context, some *different* subset of C_k's neighbors will tend to predict, one state in advance, the onset of its own activation. This reflects the determinism of the trajectories of physical features moving in 3–D space, when the motion state of the observer is known.

2.4. The Contextron Learning Rule

We have seen the desired final distribution of weights for a Contextron, in which each motion–context enables only the weighted input connections from those neighbors that tend to predict the Contextron's own activation, one state in advance. In order to justify this final distribution of weights, we describe here the Contextron learning rule, a variant of the Perceptron learning rule [Rosenblatt, 1958], and tracing its roots to a learning rule proposed by Hebb [1949] for neurons:

> "If a neuron, A, is near enough to another, B, to have any possibility of firing it, and it does take part in firing it on one occasion...the probability is increased that when A fires next B will fire as a result."

Within VIPS, the weight modification rule takes the following form: whenever the feedback input from a neighbor is repeatedly in temporal agreement with the *retinal* input to a Contextron, which acts as a "teacher", then the weight for this neighbor's feedback input is selectively increased for the motion–context currently in effect. Feedback inputs that are *uncorrelated* in their activity with the retinal input are punished, down to a minimum weight of zero. The details of the weight modification algorithm are ommitted here, as they are essentially identical to iterative weight modification rules described elsewhere [e.g. Rosenblatt, 1958].

The physical significance of this rule is as follows: if the current–state activity of some neighbor, as conveyed by its radially projecting feedback path, is repeatedly "felt" to arrive at the same time as the *retinal* stimulation to a Contextron, then the retinal input becomes, in a sense, informationally redundant. Ultimately therefore, a Contextron is able to compute its next state solely on the basis of the current state of its neighbors, the primitive capacity that enables VIPS to internally simulate visual events without benefit of the retina. No negative weights are needed on the assumption that visual micro–features will not in general act consistently as *negative* evidence for other visual micro–features, from state to state in a motion sequence.

The motivation behind the continuity–of–optic–flow assumption can now be made more clear. VIPS can represent exactly those global visual transformations that are both smooth enough and slow enough that each Contextron can know (with high probability) its next state as a function of current state of activity in its fixed local neighborhood.

3. Current Status of VIPS Implementation

3.1. A Simple Visual/Motor Environment

A standard graphics package is used to generate left– and right–eye views of simple 3–D objects on the 32 by 32 bit retinae (fig. 3). VIPS "moves" by issuing a motor command, which has the dual function of setting up the motion–context input to the field of Contextrons, as well as stimulating its virtual "musculature", having the desired side–effect of producing motion–related changes to the binocular visual display. Thus if VIPS were to issue the command "circle object to left", it would "see" the binocularly depicted 3–D object rotating to the right in depth, in 10° intervals about a vertical axis in the center of the visual field.

3.2. Visual–Motor Learning Phase

Since each VIPS motion–context enables a *logically* distinct (though perhaps *physically* overlapping) set of feedback connections for each Contextron, the single type of circular motion described above has been implemented as a test of system principles[2].

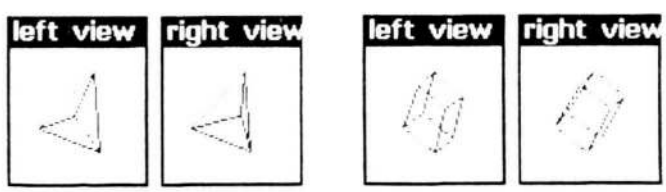


Figure 3. Examples of binocularly displayed VIPS objects.

[2] Pragmatically, the number of possible motion–contexts for each Contextron is limited by the number of resolvable PPP's for that Contextron, a point that is developed further in [Mel, 1986b].

During this phase, each Contextron learns to anticipate retinal stimulation on the basis of the current state of its relevant neighbors. In actively circling a number of randomly constructed and situated parallelograms, the pattern of weighted input connections to each Contextron is sharpened in the following way: input weights from neighbors that only spuriously concur with a Contextron's retinal input will eventually decay to zero, while inputs from neighbors that reflect real, physical predecessors in a motion sequence will become strengthed over time. In this way, each Contextron in VIPS begins to preferentially develop feedback connections from its PPP. Figure 4 shows the pattern of input weights to a particular far-depth/oblique cell that developed through training with 60 total views (e.g. 5 objects through 12 rotation steps each). The figure shows that this particular Contextron developed heavily weighted input connections exclusively from its neighbors of similar orientation, lying slightly to the right and slightly less distant on the average, and lined up in a curious, obliquely-oriented macro-pattern around the PPP. This macro-pattern was unexpected but has the following explanation: whenever the PPP is active, so will a continuous line of *other* Contextrons of similar orientation and depth be active in general, since the objects seen by VIPS tend to be composed of lines much longer than the receptive field of single Contextrons. These may be referred to as the physical correlates of the PPP. An interesting consequence of this elongated pattern of feedback connections is that a Contextron's PPP need not *itself* be present in the current state in order for the Contextron to fire in the next state--instead, VIPS has a tendency to "see" an absent feature if its presence is strongly implied by its physical correlates. The possible relationship of this effect to the perception of *subjective contours* has yet to be investigated.

3.2.1. Envisionment Phase

The second phase of VIPS' operation is the phase of *internal simulation* or *envisionment*, and runs concurrently with phase 1--but only becomes accurate after sufficient phase 1 training. In phase 2, let us assume the array of Contextron's in VIPS is excited into some initial state of activation by the retina, when confronted by a novel 3-D object viewed from an arbitrary perspective. This internal visual state of the Contextron array may be thought of as a "mental image". (Graphic representations of several mental images can be seen in figure 6, each actually a 16K element vector of activation levels over the field of Contextrons, where the intensity at each pixel represents the combined activation levels from cells of 4 orientations at 4

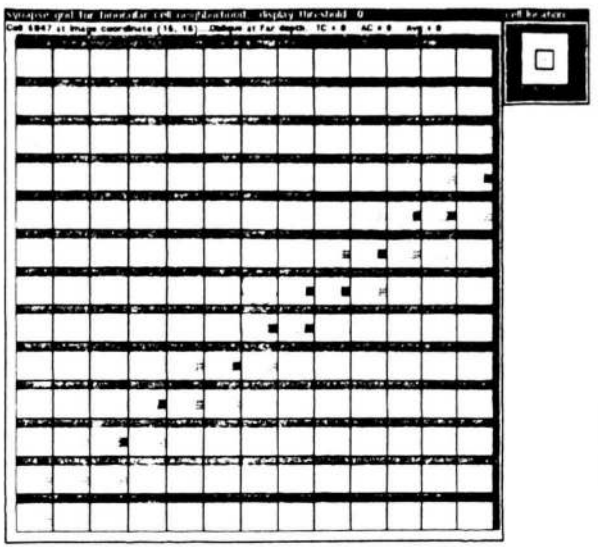


Figure 4. One Contextron's pattern of weighted inputs from its neighbors.

depths with receptive fields centered on that pixel.) By issuing a motion-context to its Contextron's and temporarily inhibiting the retinal pathway, VIPS can transform (e.g. rotate, zoom, or pan) this mental image through time in an *approximation* to the internal state sequence that would be driven by the retinae, were VIPS *actually* moving through its environment and "seeing" the changes. This internal simulation can be successful since each Contextron has learned during phase 1 to compute its *probable* next state of activation solely on the basis of the current state of its neighbors--a level of activation that is normally confirmed by the retinae during a *real* motion sequence. An internal simulation will continue, state by state, until the random error introduced at each state transition renders the mental image unrecognizable.

It was first desired to establish VIPS' capacity to represent 3-D visual transformations under optimal conditions. To this end, VIPS was trained from scratch with a *single* visual-motor sequence, i.e. circling a wire-frame rectangle to the left through 70°, and was simply asked to reproduce the sequence internally. In this configuration, VIPS simply acted as a sensory "tape-recorder", and performed with a high degree of accuracy. Figure 5 depicts both the ideal, retinally driven sequence and the subsequent, internally simulated sequence, with a correlation coefficient reflecting the increasing difference between the two sequences. After 7 steps in the sequence (70 degrees), VIPS maintained a correlation of 0.74 to the ideal retinally produced state. For purposes of comparison, two adjacent states in a visual sequence are typically

correlated at only 0.25, this being a measure of the difference between two views of the same object separated by 10 degrees. This first test therefore indicates that VIPS *is* in fact capable of accurately representing 3–D transformations.

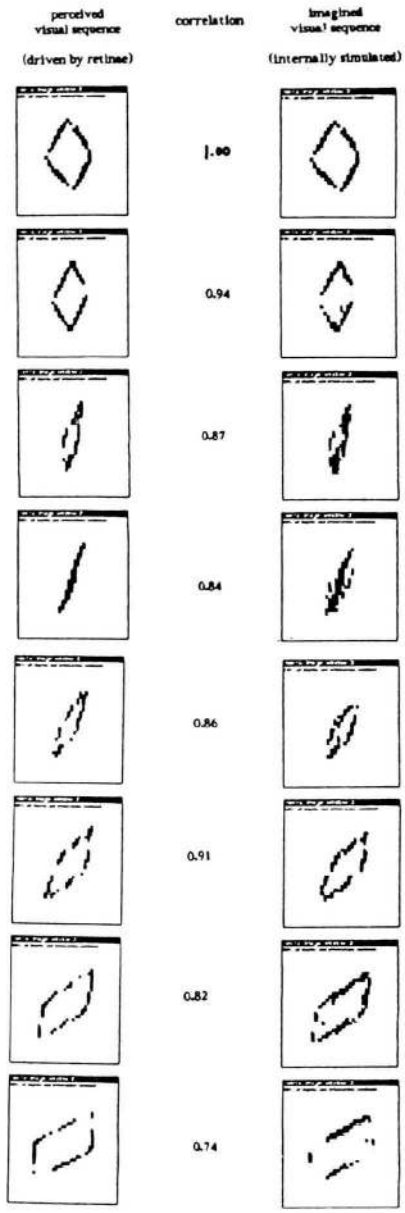


Figure 5.

The results of a more comprehensive training run in the same motion–context appear in Figure 6, which shows the gradual improvement in VIPS' ability to mentally rotate a novel object (a wire–frame pyramid) through one step (i.e. 10°). The left–hand column represents the same, ideal, retinally produced coding in each case, while the right–hand column represents the internally transformed coding after one step. The absolute accuracy of the simulations is still unspectacular, but improving as the implementation is refined.

4. VIPS: Psychological Issues

The idea that visual perception is fundamentally an interactive visual–motor process extending through time, is a rather old one. Whereas the prevailing philosophy in the fields of artificial intelligence and computer vision have focused on the algorithmic extraction of invariant features for the recognition or classification of static visual images, many authors in the psychology of visual perception have emphasized the dynamic character of vision [Miles, 1931; Koffka, 1935; Gibson et al., 1959; Wallach & O'Connell, 1953; Ullman, 1979]. Gibson [1979] argues forcefully that optical motion is the true substance of visual perception, that vision fundamentally involves the extraction of meaning from a continuously changing retinal image, and that optical rest is but a limited special case.

Others have stressed more explicitly that visual–motor *interactions* are the essential ingredient in the perception of space and in the development of such perception [Berkeley, 1709; Washburn, 1916; Helmholtz, 1925; Piaget, 1956; Held & Hein, 1963; Gyr et al., 1979]. Piaget [1952] gives a detailed account for the development of a child's conception of space in terms of the coordination of early motor and visual schemas, an accurate if metaphorical description of VIPS' learning phase. He further describes the child's transition from the *sensory–motor* to the *pre–operational* period as the time when the child learns to take his motor schemas "underground", allowing "abbreviated" movement to drive visual imagery. Again, an apt metaphor for VIPS' envisionment phase of operation.

Metaphor aside, three general areas from the experimental psychology of visual perception seem to support the VIPS account for 3–D visual representation.

4.1. Perceptual Stability

The intimate link between visual and motor processes in perception is perhaps most evident in the long tradition of work in *perceptual stability* [e.g. Held,

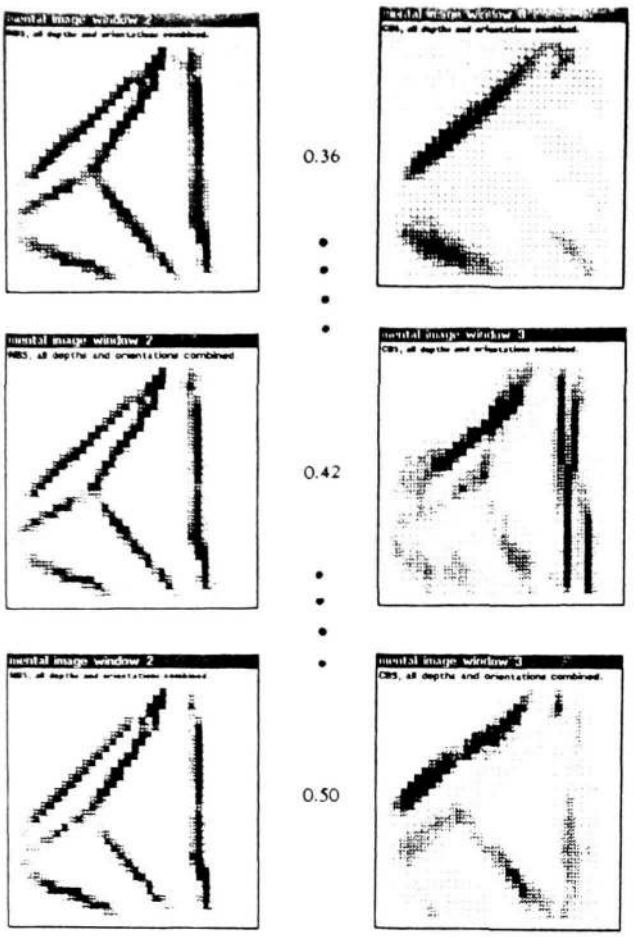


Figure 6. Gradual improvement in 10° mental rotation with training.

1965; Epstein, 1977; Wallach, 1985]. The question addressed by these workers is as follows: Given that the retinal images change with great rapidity during normal visual behavior, what accounts for the subjective stability of perception? The clearest answer to emerge from a large body of experimental work holds that the brain allows active movements of the eyes, head, and body to generate precise expectations of change in the visual field. To the extent that the internal predictions *concur* in real-time with the actual changes in the visual field, the visual environment is subjectively perceived as *stable*, while a lack of concurrence between internal prediction and perception is an indication of movement or change in the environment. In one of its modes of operation, VIPS can be described as just such an "on-line" prediction mechanism, where its internal motion-context field indirectly causes VIPS to "move", while simultaneously maintaining each Contextron in the appropriate context in which to compute its own probable next state of activation. VIPS has yet to be put to the task of identifying parts of its internal visual map for which the locally predicted next-state of each cell is *not* in agreement with the incoming retinal signal, though it is ideally suited to perform this kind of computation. The identification and analysis of autonomously moving bodies in the visual field is an area of particular interest for future work and is discussed briefly in the concluding section.

4.2. Development and Plasticity

Interestingly, the brain's predictive mechanisms for perceptual stability are *plastic*. That is, if the relationship between movements of the self and the associated patterns of visual change is suddenly altered, artificially or otherwise, the predictive mechanisms will at first make incorrect predictions, generally by reporting a stationary environment to move in disconcerting ways. As the subject of such a sudden alteration continues to move about however, he gradually learns the *new* visual-motor relationship, until his internal predictions coincide once again with the stream of retinal input, at which time the visual environment is reported to be stable once again. This illustrates the fact that the stability of perception is much less a function of the absolute rate-of-change in the visual field, than it is of the degree of correspondence between predicted visual change and that which is actually reported by the retinae. Stratton [1896] was the first to experiment with this effect, demonstrating that perceptual stability could be *reattained* in less than a week, even when the visual field was turned upside down with inverting lenses.

Entirely in keeping with the notion that the development of 3-D perception involves the learning of visual-motor relationships, neither development in the young nor such plasticity effects in the adult are observed, in general, in the absence of *active* motor participation on the part of the perceiver. A startling example of this fact is witnessed in an experiment of Held & Hein [1963]. Two cats were yoked together on opposite sides of a circular treadmill, one of the cats walking under its own power, the other being passively carried through the same circular trajectory in a basket (fig. 7). Thus, while both cats shared essentially identical visual experience, only the active cat had the opportunity to learn a consistent relationship between its own activity and the resulting changes in its visual array. On subsequent testing, the active cat appeared normal, while the passive cat was determined to be deficient in a variety of 3-D spatial tasks. A second developmental study with related results is the celebrated visual-cliff study of Gibson & Walk [1960], in which the initial hypothesis and ultimate conclusion both stated that infants of a variety of species begin to

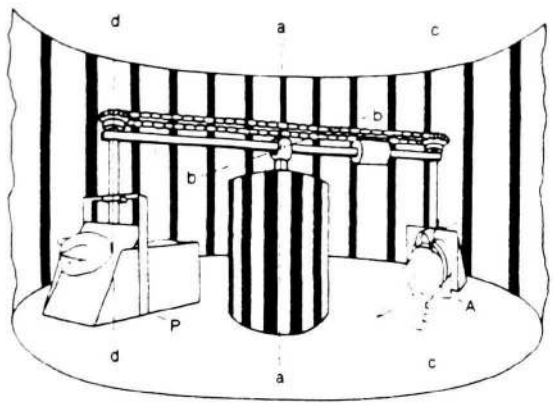


Figure 7. Active vs. passive visual change during development. (From "Movement-produced stimulation in the development of visually guided behavior", by R. Held & A. Hein. In *Journal of Comparative and Physiological Psychology*, 1963, *56*, 872-876. Copyright 1963 by the American Psychological Association.

appreciate the behavioral significance of the third dimension about the time they begin to move under their own power.

VIPS displays both of these effects of development and plasticity: motor activity must be correlated with visual change for successful development of 3-D appropriate behavior, and if the visual-motor relationship is ever altered, the simple weight modification rule will adaptively maintain VIPS' predictive function in as close agreement with the retinal stream as possible. From a systems point of view, this state of affairs is extremely advantageous, since the system designer need not anticipate a *particular* set of visual-motor transformations by building in the appropriate set of special-purpose algorithms [e.g. Hinton, 1981; Kosslyn & Schwartz, 1977; Funt, 1983]. Instead, only the *continuity-of-optic-flow* condition on the visual transformations need be assumed *a priori*, and to the extent that the visual world is well behaved in this respect, VIPS can learn (and relearn) to reliably predict 3-D, movement-induced visual change.

4.3. Mental Rotation

A extensive and elegant body of work on the chronometry of mental rotation, zoom, pan, folding, and apparent motion [Shepard & Metzler, 1971; Robins & Shepard, 1977; Shepard, 1978; Kosslyn, 1980; Shepard & Cooper, 1982; Finke & Pinker, 1983] elucidates a further quality of the internal representations of 3-D objects in human subjects with which VIPS is compatible. The most striking characteristic common to all of these mental transformations has been that the time necessary to carry out the mental transformation grows with its spatial extent, often linearly. Thus, for example, it takes twice as long to mentally rotate an object through 60° as 30°. While this result may seem intuitive, it is very troublesome for those theories of vision that argue for an internal, *view-independent* representation of a 3-D object, for which no mental rotation should be necessary [e.g. Marr, 1982]. On the basis of the chronometry and other data, Shepard [1979] draws the following conclusion on the nature of the internal visual representations:

> "The mental transformation is carried out over a path that is the internal analog of the corresponding physical transformation of the external object...By analogical or analog process I mean just this: a process in which the intermediate internal states have a natural one-to-one correspondence to appropriate intermediate states in the external world."

VIPS has exactly these analogical properties, where the intermediate states during an internal simulation correspond to intermediate visual states, and the time necessary to carry out a transformation is indeed a linear function of its extent. Kosslyn & Schwartz' influential model [1977] captures this same property for a variety of 2-D image plane transformations, but gives no account for the development or plasticity effects, and more importantly, does not deal with 3-D visual transformations. Funt [1983] models stepwise 3-D mental rotations with a built-in spherical shell of processors, but points out that the model cannot account for stepwise transformations of any other kind.

5. Conclusions and Future Directions

The VIPS architecture has been proposed as a more "natural" approach to the representation of the 3-D visual world and its transformations. Starting only with a set of simple, built-in binocular feature detectors driving a regular grid of radially interconnected Contextrons, VIPS learns to envision the visual consequences of an arbitrary motion in its repertoire. VIPS' knowledge of the third dimension is not embodied in explicit 3-D models of the visual scene, as is the common *modus operandi* in computer vision systems. Instead, the knowledge lies in the pattern of modifiable weights that determine how the pattern of activation

over its internal binocular map of the 3–D visual world shifts about during each kind of motion.

Beyond simply improving the performance of the current implementation, there are three directions currently planned for VIPS that reflect its significant limitations.

Firstly, VIPS has a disturbing "out–of–sight, out–of–mind" quality, with no representation of surfaces hidden from view. This limitation highlights an important future direction for this project. Currently, the initial visual state that begins every VIPS simulation is restricted to come from the retinae. If instead, VIPS were configured as one of several distinct fields of Contextrons representing a variety of modalities, then a visual state could be induced in VIPS via a parallel projection from one of the other *internal* fields, supplanting the function of the retinae. The induced visual state could, among other things, be a "remembered" view of the backside of an object. Clearly, however, the complex and varied processes underlying full object–recognition remain far from specified within this model, and are topics for future work.

Secondly, VIPS cannot currently "understand" autonomously produced motion in its visual field. On the other hand, VIPS is capable of *representing* any smooth transformations of the binocular visual field, including those produced by autonomously moving bodies. One possible solution to this problem is a process by which VIPS can "fiddle" with its internal motion contexts until its Contextrons are predicting, as well as is possible, the autonomously produced visual changes.

Most importantly however, to be truly useful, VIPS must be embedded in a larger system that has some form of representation of plans and goals. Currently, VIPS is *capable* of running internal simulations of 3–D visual transformations, but embodies no knowledge as to *why* a particular simulation might be useful for some particular goal, such as bringing a novel view of a 3–D object into registration with another for comparison. VIPS is therefore a useful tool to a system that wishes to run particular internal simulations in service of its goals. This remains a rich and unexplored area for further research.

References

Attneave, F. Representation of physical space. In A. W. Melton & E. Martin, (Eds.), *Coding Processes in human memory*. Washington, D.C., 1972.

Berkeley, *An essay towards a new theory of vision*. 1709 (see any modern edition).

Epstein, W. *Stability and Constancy in Visual Perception: Mechanisms and Processes*. New York: John Wiley & Sons, 1977.

Feldman, J. A. Four frames suffice: A provisional model of vision and space. *The Behavioral and Brain Sciences*, 1985, (in print).

Finke, R. A. & Pinker, S. Directional Scanning of Remembered Visual Patterns. *Journal of Experimental Psychology: Language, Memory, and Cognition*, 1983, *9(3)*, 398–410.

Funt, B. V. A parallel process model of mental rotation. *Cognitive Science*, 1983, *7(1)*, 67–93.

Gibson, E. J. & Walk, R. D. The visual cliff, *Scientific American*, 1960, *202*, 64–71.

Gibson, J. J. *The ecological approach to visual perception* Boston, MA: Houghton Mifflin, 1979.

Gibson, E.J., Gibson, J.J., Smith, O.W., & Flock, H. Motion parallax as a determinant of perceived depth. *J. Exp. Psychol.*, 1959, *8*, 40–51.

Gibson, E.J. & Walk, R.D. The visual cliff, *Scientific American*, 1960, *202*, 64–71.

Gyr, J., Willey, R., & Henry, A. Motor–Sensory feedback and geometry of visual space: an attempted replication. *Behavioral and Brain Sciences*, 1979, *2*, 59–94.

Hebb, D. O. *The organization of Behavior*. New York: Wiley, 1949.

Held, R. & Hein, A. Movement produced stimulation in the development of visually guided behavior. *Journal of Comparative and Physiological Psychology*, 1963, *56*, 872–876.

Held, R. Plasticity in sensory–motor systems. *Scientific American*, 1965, *213*, 84–94.

Helmholtz, H.v. *Handbook of physiological optics*, vol. 3. New York: Optical Society of America, 1925. (Translated by J.P.C. Southall).

Hinton, G. E. Shape representation in parallel systems. In *Proceedings of the 7th International Conference on Artificial Intelligence, 2,*, 1088–1096. Vancouver BC, Canada, 1981.

Hubel, D.H., & Wiesel, T. N. Brain mechanisms of vision. *Scientific American*, 1979, *241*, 150–162.

Koffka, K. *Principles of gestalt psychology*. New York: Harcourt, 1935.

Kosslyn, S. & Schwartz, S. Simulating visual imagery. *Cognitive Science*, 1977, *1*, 265–295.

Kosslyn, S.M. 1980 *Image and Mind*. Cambridge, MA: Harvard University Press, 1980.

Marr, D. *Vision*. San Francisco: Freeman Press, 1982.

Marr, D. & Nishihara, H. K. Representation and recognition of the spatial organization of three–dimensional shapes. *Proc. R. Soc. Lond.* B, 269–294, 1978.

Mel, B.W. Model for the structure and behavior of a hypothetical visual–motor association cortex. Working Paper #69, 1986a. Artificial Intelligence Group, Coordinated Sciences Lab,

University of Illinois, C-U.

Mel, B.W. A connectionist model for simple visual function. Working Paper #70, 1986b. Artificial Intelligence Group, Coordinated Sciences Lab, University of Illinois, C-U.

Miles, W.R. Movement in interpretations of the silhouette of a revolving fan. *Am. J. Psycholol.*, 1931, *43*, 392–404.

Minsky, M. A framework for representing knowledge. In *The psychology of computer vision*, P. H. Winston, (Ed.), New York: McGraw-Hill, 1975

Piaget, J., & Inhelder, B. *The child's conception of space.* New York: Humanities Press, 1956.

Robins, C. & Shepard, R. N. Spatio-temporal probing of apparent rotational movement. *Perception and Psychophysics*, 1977, *22*, 12–18.

Rosenblatt, F. *Principles of neurodynamics. Perceptrons and the theory of brain mechanisms.* Washington, D.C.: Spartan Books, 1961.

Shepard, R. N. The mental image. *American Psychologist*, 1978, *33*, 125–137.

Shepard, R.N. & Cooper, L. *Mental images and their transformations.* Cambridge, MA: MIT Press, 1982.

Shepard, R.N., & Metzler, J. Mental rotation of three-dimensional objects. *Science*, 1971, *171*, 701–703.

Stratton, G.M. Vision without inversion of the retinal image. *Psychol. Rev.*, 1896, *3*, 611–617.

Ullman, S. *The interpretation of visual motion.* Cambridge: MIT Press, 1979.

Wallach, H. *Scientific American*, 1985.

Wallach, H. & O'Connell, D. N. The kinetic depth effect. *Journal of Experimental Psychology*, 1953, *45(4)*, 205–217.

Washburn, Margaret F. *Movement and Mental Imagery.* Boston & New York: Houghton Mifflin Company, 1916.

Seeing-More-Than-is-There: A Probe of Retinoid Networks

ARNOLD TREHUB and ALEXANDER POLLATSEK
Department of Psychology
University of Massachusetts, Amherst

ABSTRACT

Dynamic properties of a model post-retinal, short-term visual store called a retinoid are presented. The retinoid model generates specific predictions about visual perception in several variants of the seeing-more-than-is-there (SMTT) paradigm. Three experiments were conducted to test for the predicted effects. Performance curves are presented and discussed. A new visual illusion of motion is shown to follow from the properties of the retinoid model.

INTRODUCTION

A fundamental question in the study of human vision concerns the integration of disparate foveal stimuli into a proper unified representation of a larger real-world scene. Some kind of post-retinal mechanism which can register and appropriately combine inputs from a series of visual saccades or a continuous visual scan, seems to be required. To account for this central process, as well as for a variety of other high-level visual operations, a theoretical neuronal mechanism, called a retinoid, has been proposed and described in detail (Trehub, 1977, in press).

If a target figure is moved back and forth behind a narrow window in an otherwise opaque surface, much more of the figure can be seen at a given instant than is physically present within the window. This phenomenon was reported as long ago as 1862 by Zollner, and by Helmholtz in 1867. It has been called the seeing-more-than-is-there effect (McCloskey and Watkins, 1978). The phenomenon seems to depend upon a post-retinal integrating mechanism like a retinoid to give the subject the perception of a visual pattern which is much larger than the small "slices" of the figure actually stimulating the retina. In the laboratory, one can exercise systematic control over a number of stimulus and contextual parameters for the seeing-more-than-is-there (SMTT) effect. Thus SMTT may be a particularly good paradigm for probing the dynamic properties of retinoid networks. In this paper, we review the basic structure and general operating characteristics of the retinoid model, and test some implications of the theory using the SMTT procedure.

Theoretical Model

A retinoid is a postulated brain mechanism that can organize successive retinocentric visual patterns into a unified egocentric or allocentric representation of object space. The network is called a retinoid because, like the retina, it represents information in visual space and projects afferents to higher visual centers. Retinoids serve as visual "scratch pads" with spatially organized information stored as short-term memory. The mechanism of storage is assumed to be a retinotopically

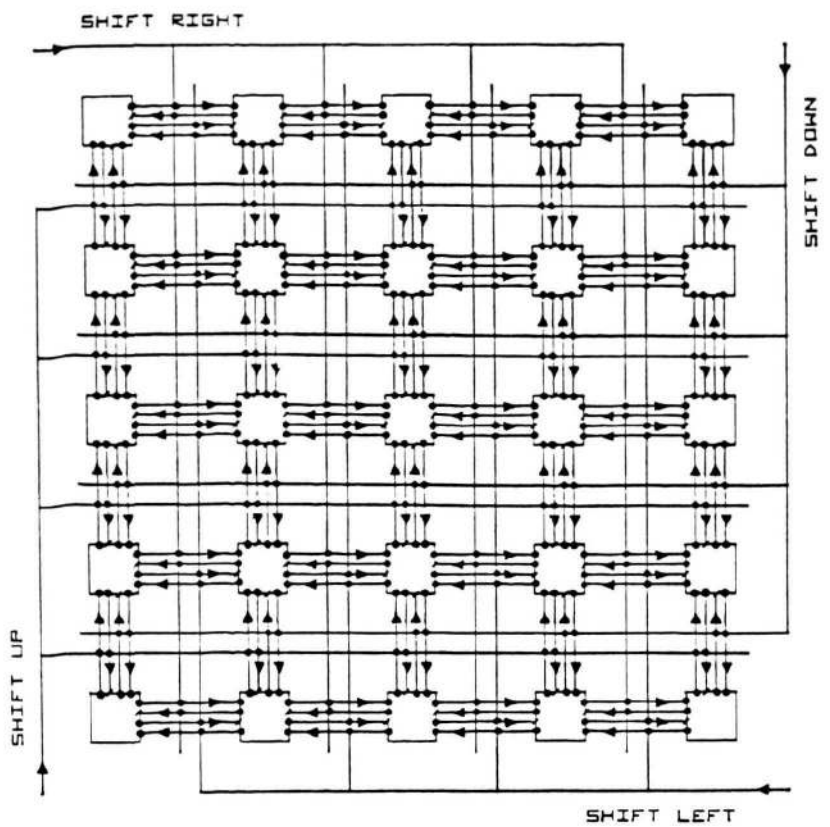


Figure 1. Translation retinoid. Large squares represent autaptic cells serving short-term memory. Small filled triangles represent interneurons. Shift-control cells are designated by direction of effect.

organized array of excitatory autaptic neurons. Neurons of this type have at least one of their axon collaterals in recurrent excitatory synapse with their own cell body or dendrite (Shepherd, 1974). In order for an active autaptic neuron to refire itself, there must be other concurrent excitatory input to it. Thus the sheet of autaptic neurons in the retinoid can represent in its sustained firing pattern any transitory organized input for as long as diffuse priming excitation (arousal) is sustained. If the priming background input is removed or sufficiently reduced then the captured pattern on the retinoid is "erased".

Figure 1 shows a retinoid of autaptic neurons connected by a grid structure of excitatory and inhibitory interneurons. Axon collaterals of shift-control neurons are in excitatory synapse with selected groups of interneurons. Any momentary input from an afferent visual array to its homologous autaptic retinoid cells will evoke sustained firing of the retinoid targets if there is a sufficient level of diffuse tonic excitation. Thus any retinal stimulus induces a comparable retinoid pattern of spatially organized discharge. Each active autaptic neuron induces a

sub-threshold, priming excitatory postsynaptic potential (EPSP) in each of the eight contiguous interneurons capable of eliciting excitatory and inhibitory potentials (IPSP) in their targeted autaptic cells. If a primed interneuron receives an increment of excitation from one of the shift-control cells, that interneuron will fire and send a spike input to its target cell. The rules are: (a) if an autaptic cells which is not discharging (i.e., off) receives an EPSP from an interneuron, it will fire (turn on); (b) if an autaptic cell which is on receives an IPSP from an interneuron, it will turn off unless it receives simultaneously an EPSP from another interneuron, in which case it will remain on; (c) if diffuse tonic excitation to the retinoid falls below a critical level, all cells in the retinoid turn off.

If we imagine a standing pattern of excitation evoked on a retinoid, this "captured" pattern can be spatially translated in any direction by appropriate pulses (discharge spikes) from the shift-control command cells. For example, each pulse from the shift-right line will transfer standing activity from any active autaptic cell to an adjacent autaptic cell on its right and, at the same time, erase the activity of the previously active donor (on the left) unless that cell is also receiving transferred excitation from an active autaptic cell to its immediate left. Thus successive command pulses from the shift-control cell will move the entire retinoid pattern to the right in successive increments of a single autaptic cell. The more rapid the pulses, the more rapid the pattern will move; the longer the shift-command pulse train is sustained, the greater will be the distance through which the pattern is moved. Appropriate sequences of shift right/left, shift up/down can move the pattern of cell activity to any position on the retinoid surface. The model assumes that automatic translation of retinoid patterns occurs when shift-control neurons are driven by eye and/or head deviations from the normal foveal axis (the line of sight of the fovea when the eyes are straight ahead, head unturned, shoulders square with the body), and by the discharge of motion detecting cells in the lower-level visual mechanisms.

Explaining SMTT

We theorize that the SMTT phenomenon occurs when the component pattern segments, which are directly observed through the narrow aperture, are swept across the retinoid surface in synchrony with signals from the motion-detector cells. This explanation depends on: (a) the detection of the velocity and direction of the stimulus pattern on the basis of information available in the screen aperture; (b) the short-term memory properties of the retinoid that enable it to assemble a unified representation of the stimulus from the sequence of partial inputs that are registered in its aperture region and then shifted step-wise across adjacent cells.

Figure 2 shows, in schematic form, the significant factors in the SMTT paradigm. The stimulus pattern which is depicted is a triangle moving to the right with its leading edge just entering the aperture in an otherwise occluding screen. The symbols in the schematic figure are to

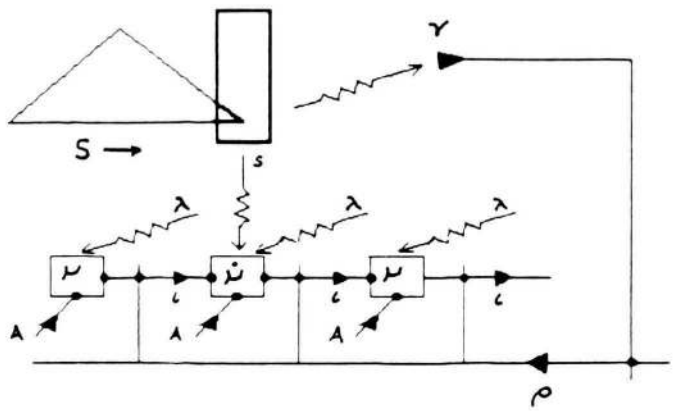


Figure 2. Schematic showing significant elements and parameters in SMTT paradigm.

be interpreted as follows.

S = input from stimuli in aperture

$\dot{\mu}$ = autaptic cell in the aperture field excited by stimulus segments

μ = autaptic cell beyond aperture field

i = excitatory interneuron

ρ = retinoid shift-control cell (shift-right)

γ = motion detector (senses velocity of motion to right)

A = diffuse input from arousal

λ = diffuse input from ambient illumination

Let us now consider how some of the factors described above affect the rate of spatial translation over the retinoid network. It should be noted that the subset of autaptic cells ($\underline{\dot{\mu}}$) "looking at" the stimulus segments receive their input from: (1) stimulus excitation ($\underline{S}$); (2) arousal excitation ($\underline{A}$); (3) ambient illumination ($\underline{\lambda}$); (4) excitation from active interneurons ($\underline{i}$). Since neurons integrate all EPSP input over time, the slope of integration and, thus, the frequency ($\underline{F}_{\dot{\mu}}$) of cell discharge in $\underline{\dot{\mu}}$ will be a positive monotonic function of the dwell-time of $\underline{S}$ on $\underline{\dot{\mu}}$, the magnitudes of $\underline{A}$ and $\underline{\lambda}$, and the discharge frequency of $\underline{i}$. For the subset of autaptic cells ($\underline{\mu}$) outside of the direct stimulus (aperture) field, their discharge frequency ($\underline{F}_{\mu}$) will be a positive monotonic function of input from their associated interneurons

$\underline{i}$ and the magnitudes of $\underline{A}$ and $\underline{\lambda}$. These relationships may be expressed as follows.

$$F_{\dot{\mu}} \propto (S, A, \lambda, i) \quad (1)$$

$$F_{\mu} \propto (A, \lambda, i) \quad (2)$$

The discharge frequency ($F_{\underline{i}}$) for any interneuron ($\underline{i}$) is a positive monotonic function of the input pulse rate from its shift-control driving cell ($\underline{\rho}$) and the spike frequency of its donor autaptic cell ($\underline{\dot{\mu}}$ or $\underline{\mu}$). Thus

$$F_i \propto (\rho, \dot{\mu}) \quad \text{or} \quad F_i \propto (\rho, \mu) \quad (3)$$

Given any arbitrary gating pulse rate from a shift-control cell (say $\underline{\rho}$), variation in the actual velocity of translation of an excitation pattern across the surface of a retinoid will be a positive function of the discharge frequency ($\underline{F}$) in those retinoid cells carrying the pattern. This follows from the fact that the latency of discharge in each successive cell on the translation route of the image is inversely related to the $\underline{F}$ of its donor cell. If the translation velocity of the retinoid image is equal to the velocity of the stimulus pattern, then it is assumed that the perceived proportions of the pattern will correspond to the veridical proportions of the stimulus. If, however, translation velocity on the retinoid is less than stimulus velocity, then the stimulus pattern will be perceived as shorter along its axis of motion in relation to its other dimensions.

GENERAL METHOD

Implications that follow from the assumptions about the dynamics of retinoid networks as presented above, were tested in three SMTT experiments. In particular, we probed the effects of stimulus velocity, aperture width, and ambient illumination.

Subjects

Four subjects participated in each of three experiments. The subjects, the two authors of this paper and two graduate student volunteers, were unpaid. The student subjects were unaware of the theoretical background and other details of this study until after their data were collected.

Apparatus

Visual stimuli were presented on the screen of a CRT which was partially masked by a sheet of gray poster-board with a narrow, vertically oriented aperture in its center. The height of the aperture was fixed and was greater than the height of the stimulus pattern; the width of

the aperture was adjustable and always much less than the width of the stimulus, so that only a small segment of the stimulus pattern could be seen by the subject at any given moment. A hand-held controller was provided the subject that had a potentiometer knob which allowed him to vary the rate of stimulus motion across the face of the CRT from 1.0 degrees/second to a maximum velocity of 12.5 degrees/second at the viewing distance used in this study. The controller also had a button which, when pressed by the subject, signalled that a perceptual criterion was met. After instructions were read to the subject, all experimental trials were controlled by an Apple II+ computer. Stimulus presentation, recording and analysis of the subjects' responses were performed automatically according to a pre-programmed protocol.

EXPERIMENT 1

Hypothesis I

Given a horizontally oscillating figure in the SMTT paradigm, the perceived width of the figure will contract as stimulus velocity increases.

This follows from the fact that dwell-time of $\underline{S}$ on $\underline{\dot{\mu}}$, hence $\underline{F}_{\dot{\mu}}$, will decrease with an increase of stimulus velocity, causing the rate of translation over the retinoid to decrease relative to stimulus velocity.

Hypothesis II

For any given stimulus velocity, the perceived width of the figure will increase as the aperture width increases.

This follows from the fact that the proportion $\underline{\dot{\mu}}/\underline{\mu}$ will increase as the aperture for direct stimulation increases. Since $F_{\dot{\mu}} > F_{\mu}$, the velocity of translation over the retinoid will be greater for the larger aperture.

Procedure

The subject was seated at a distance of approximately 58 cm from the CRT. A laterally oscillating triangle on the face of the CRT was exposed to the subject through the narrow aperture in the masking screen. The triangle disappeared behind the screen on both left and right excursions. The height of the triangle was always 0.7 degrees of visual angle, whereas the base of the triangle varied randomly on each trial within a range of 0.8 degrees to 1.9 degrees in visual angle. Incident illumination on the CRT mask was set at approximately 8.0 foot-candles. The subject was instructed to fixate the center of the aperture on each stimulus presentation, and to adjust the knob on the controller until the width of the perceived triangle was equal to its height. As soon as this oc-

curred, the subject was to press the controller button which recorded the rate setting, cleared the screen, and terminated the trial.

Data were collected from each subject under the two aperture conditions in five 20-trial blocks for each condition. Two subjects were tested on five blocks with the 0.1 degree aperture, followed by five blocks with the 0.2 degree aperture. The other two subjects viewed the stimulus first through the 0.2 degree aperture followed by the 0.1 degree aperture. During each block of trials, there was an interstimulus interval of five seconds following each response. A rest interval of 90 seconds was given after each block of trials.

Results

In this experiment, the independent variables were: (1) the height/width ratios of the stimulus patterns; (2) the width of the aperture. The dependent variable was the stimulus velocity set by the subject to give the perception of height/width equivalence. At each velocity setting required to meet the criterion of equivalence, the ratio of perceived-width/true-width is expressed as a contraction ratio. Figure 3 shows the results plotted separately for each of the four subjects. Hypothesis I is confirmed by the response curves obtained from each subject under the two aperture conditions. Using a binomial test, the confidence levels for each of the four subjects are as follows: (AP) $p<.025$; (JH) $p<.025$; (JP) $p<.005$; (AT) $p<.005$. Thus it can be concluded that in the SMTT paradigm, perceived width contracts as stimulus velocity increases.

To test Hypothesis II, we analyzed the stimulus velocity settings selected by the subject to achieve the contraction ratios required for the perception of height/width equivalence for the 0.1 degree vs. the 0.2 degree aperture. If the subject sees a wider triangle given a wider aperture, then he must increase the velocity of the stimulus to perceive equivalence. Examination of the differences in response under the two aperture conditions for each subject reveals that Hypothesis II is confirmed. Using a binomial test, the confidence levels for each of the four subjects are as follows: (AP) $p<.005$; (JH) $p<.005$; (JP) $p<.005$; (AT) $p<.005$. Thus it can be concluded that in the SMTT paradigm, perceived width increases as aperture width increases.

EXPERIMENT 2

Hypothesis III

For a given aperture, the perceived width of the figure will increase as ambient illumination increases.

This follows from the assumption that ambient illumination increases both $\underline{F}\,\dot{\mu}$ and $\underline{F}\,\mu$. The velocity of translation over the retinoid should be greater with higher general illumination.

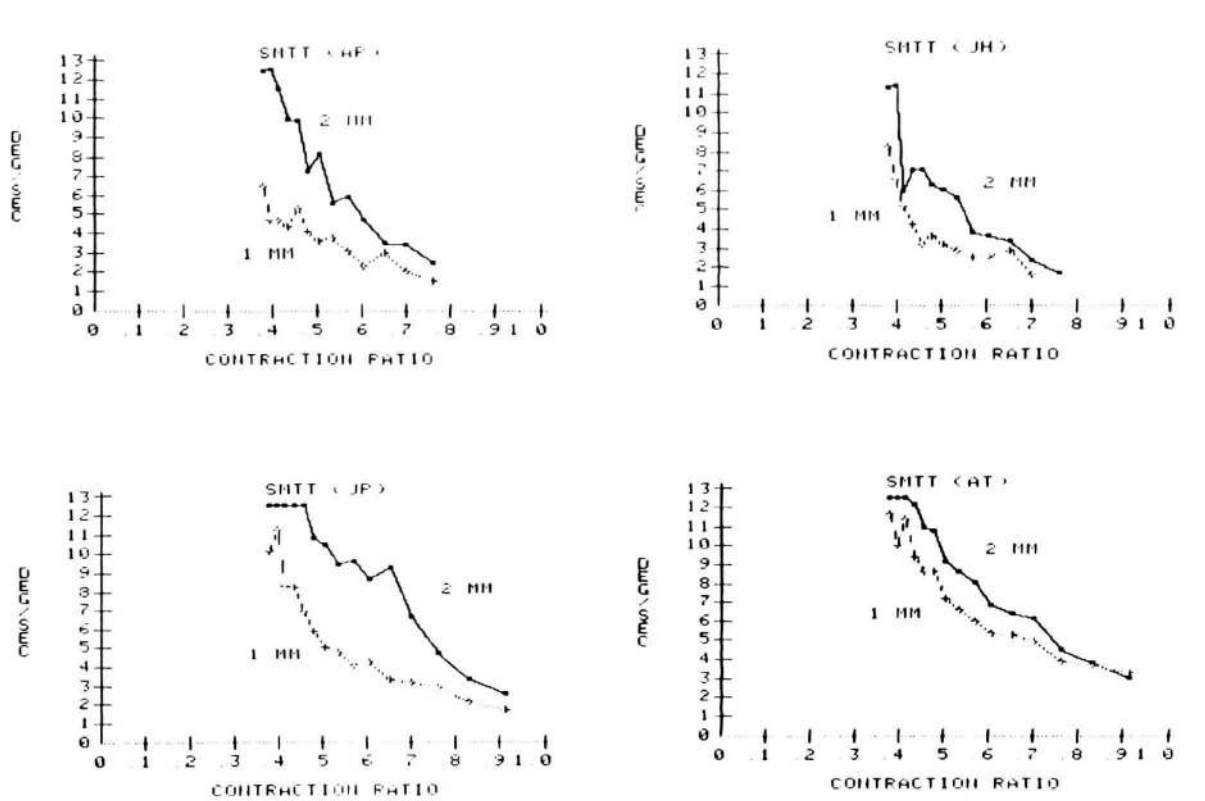


Figure 3. Experiment 1. Plot of adjusted stimulus rates against contraction ratios needed to achieve a perceived width-to-height match for each of four subjects.

Procedure

The apparatus, instructions, and procedure were the same as in Experiment 1, except for the following differences. The width of the aperture in the mask was fixed at 0.2 degrees and data were collected under two different intensities of incident illumination. Two subjects were tested with 8.0 foot-candles of incident illumination followed by 2.0 foot-candles, and for the other two the order of the illumination conditions were reversed.

Results

In this experiment, the independent variables were: (1) the height/width ratios of the stimulus patterns; (2) the levels of ambient illumination. The dependent variable was the stimulus velocity set by the subject to give the perception of height/width equivalence. Figure 4 shows the results plotted separately for each of the four subjects. Using a binomial test, three subjects showed no significant effect of illumination level, while one subject (AT) showed a highly significant effect

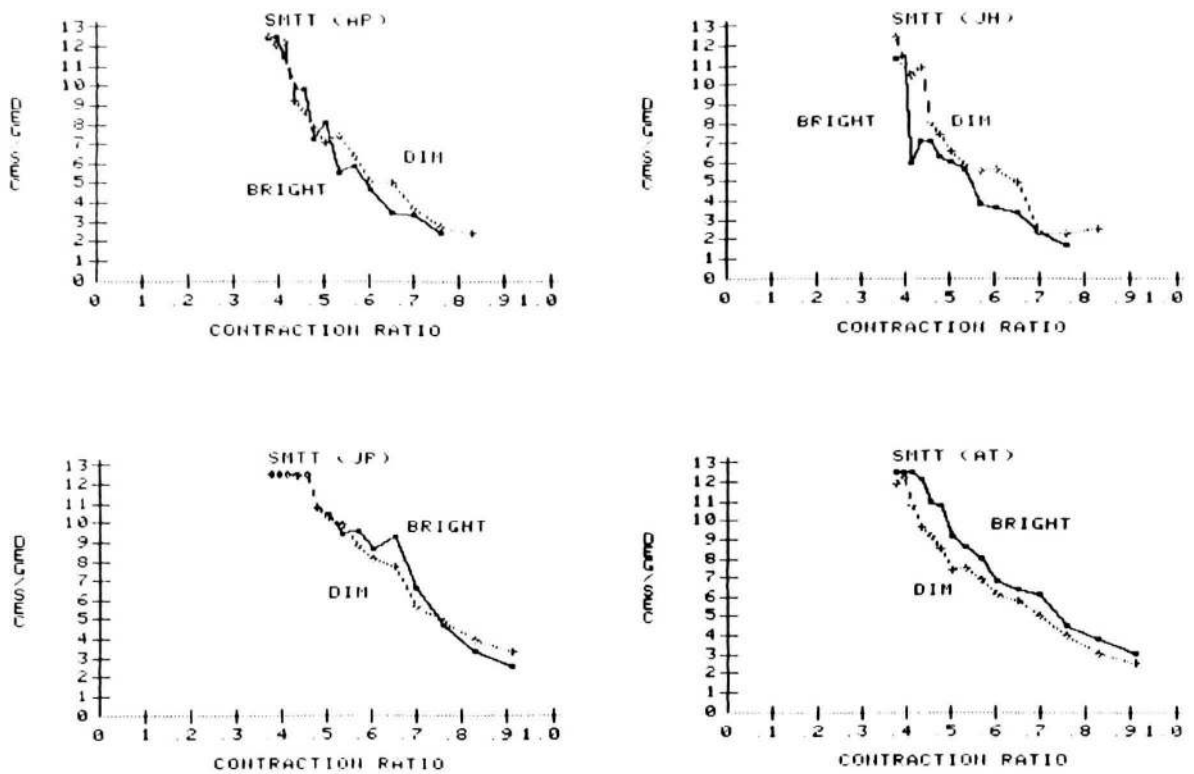


Figure 4. Experiment 2. Plot of adjusted stimulus rates against contraction ratios needed to achieve a perceived width-to-height match for each of four subjects.

in accord with Hypothesis III ($p<.005$). On the basis of these results, we are unable to confirm Hypothesis III.

EXPERIMENT 3

Hypothesis IV

In the SMTT paradigm with stimulus parameters held constant, the absolute number of stimulus sweeps required to squelch the SMTT effect will increase as ambient illumination increases.

This follows from the assumption that the small interneurons $\underline{i}$ will fatigue under prolonged driving. Thus $\underline{F}_i$ will decrease to the point that it no longer adds sufficient excitation to the target autaptic cells ($\underline{\mu}$) for them to reach discharge threshold. When this occurs, pattern translation over $\underline{\mu}$ ceases and SMTT will be squelched. Since the theoretical model assumes that ambient illumination (λ) also adds excitation to $\underline{\mu}$, it follows that SMTT squelching should occur later (more stimulus sweeps will be required) under higher general illumination. Note that

this prediction is opposite what might be expected given differences in control, since one would expect the lower background illumination (higher contrast) figure to survive longer.

Procedure

The apparatus in Experiment 3 was the same as used in Experiments 1 and 2. The mask aperture was fixed at 0.2 degrees in visual angle. The stimulus presented on each trial was a triangle with a height of 0.7 degrees and a base of 1.7 degrees of visual angle. The triangle oscillated horizontally at a constant rate of 10.0 degrees/second in visual angle, and disappeared behind the mask on both left and right excursions. In this experiment, the subject could not vary the rate of stimulus motion. Instead, the subject was instructed to fixate the center of the aperture on each stimulus presentation, and to press the controller button as soon as the perceived pattern clearly shifted from a horizontally oscillating triangle to a short, vertically oscillating line segment above a short, stable line segment (the veridical retinal input). When the button was pressed, the number of stimulus sweeps required for the perceptual shift was automatically recorded; the screen was cleared, and after a 5-second interval, a warning buzzer sounded and was followed a second later by another trial.

Each subject was given ten consecutive trials under each of two intensities of incident illumination. Two subjects were tested first with 8.0 foot-candles, followed by 2.0 foot-candles, while the order of illumination was reversed for the other two subjects. A rest interval of five minutes was allowed between the illumination conditions.

Results

In this experiment, the independent variable was the level of ambient illumination. The height/width ratio of the stimulus pattern as well as its velocity were held constant. The dependent variable was the number of successive stimulus sweeps required to squelch the SMTT effect. Figure 5 shows the results plotted separately for each of the four subjects. Examination of the differences in response under the two illumination conditions for each subject reveals that Hypothesis IV is confirmed. Using a binomial test, the confidence levels for each of the four subjects are as follows: (AP) $p<.005$; (JH) $p<.025$; (JP) $p<.025$; (AT) $p<.005$. Thus it can be concluded that in the SMTT paradigm with stimulus parameters held constant, the number of stimulus sweeps required to squelch the SMTT effect will increase as ambient illumination increases.

GENERAL DISCUSSION

The SMTT phenomenon provides a strong argument against a static "iconic" representation in visual short-term memory, and argues for a more dynamic form of storage (McCloskey and Watkins, 1978). The retinoid mechanism offers a well-articulated and biologically plausible model of what such

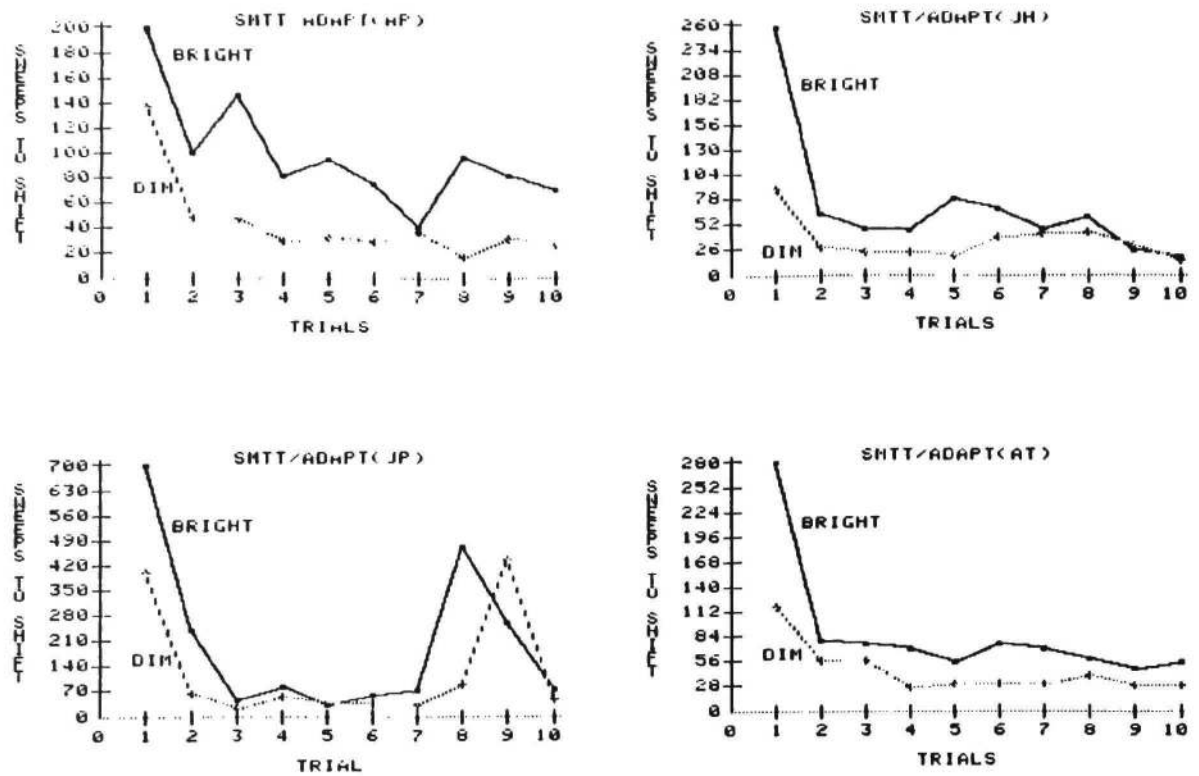


Figure 5. Experiment 3. Plot of the number of successive stimulus sweeps required to squelch the SMTT effect over ten successive trial blocks for each of four subjects.

a dynamic storage system would look like. It explains the SMTT phenomenon and makes several confirmed predictions about the effects of changing stimulus parameters.

A different explanation, depending upon a homunculus called the "executive agency", has been proposed by Rock (1983). The perceived shortening of the moving figure in the SMTT paradigm has been commented upon by several investigators (Parks, 1965; McCloskey and Watkins, 1978; Rock, 1983). Rock postulates that the motion detection system underestimates the velocity of the moving object, and the "executive agency" then uses the erroneous velocity estimation to calculate where all the points of the figure should be if they were all visible. Presumably, in Rock's theory, since missing parts of the figure should be in a particular location, they are seen in that location. In our view, this begs the central question. A major functional difference between the two models can be framed in these terms: computations and representational processes in the retinoid model are local and non-inferential, while in the Rock model they are global and inferential at the level of the unexplicated "executive agency".

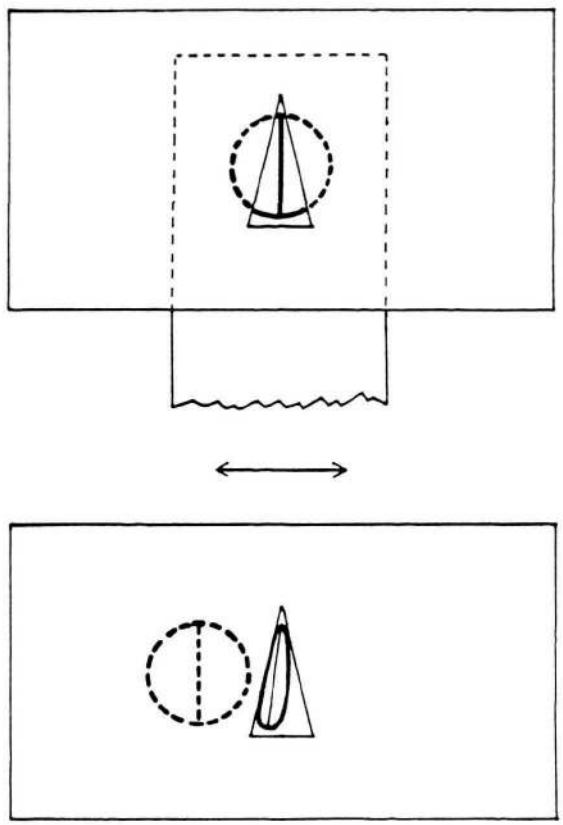


Figure 6. (Top) Screen with triangular aperture through which a part of the stimulus pattern (circle with vertical line) can be seen. (Bottom) Appearance of pendulum illusion when stimulus is moved horizontally back and forth at approximately 2 cycles/sec.

The following demonstration tests the adequacy of a local versus global explanation of SMTT and supports the superiority of the retinoid model over Rock's homunculus. If a figure were to oscillate behind a triangular aperture instead of a rectangular slit, the Rock model would predict only the usual perceived contraction along the axis of motion with no other perceptual distortion related to the shape of the aperture. The dynamics of the retinoid model, on the other hand, takes account of the change in the number of autaptic cells collinear with the direction of stimulus motion which are directly exposed to retinal stimulation as a function of the triangular shape of the aperture. This model predicts that horizontal translation rate over the retinoid surface should increase from the apex to the base of the triangular aperture. Thus, a figure which is moving back and forth should be seen to be swinging in pendular fashion as though pivoting near the apex of the triangular opening. In fact, this pendulum illusion can be clearly seen with simple materials. Using a heavy black line, draw near the top of a strip of white cardboard a circle approximately 25 mm in diameter with a vertical line through

its center. In a sheet of gray cardboard, cut out an isosceles triangle approximately 30 mm high and 15 mm wide. Holding the gray cardboard screen at normal reading distance, position the circle on the white strip behind the triangular aperture and slide it laterally back and forth at a rate of about 2 cycles/sec (taking care to move the circle beyond the aperture in each direction. If the center of the cutout is fixated, an egg-shaped figure will be perceived swinging like a pendulum within the aperture (see Fig. 6). The apparent contraction of the figure along the axis of motion occurs because the rate of translation on the retinoid is less than the veridical rate of stimulus motion. Moreover, the rate lag is greater at the apex of the window than at the base, which accounts for the egg-shaped transformation of the circle. Thus the retinoid model correctly predicts a complex illusion which should not occur according to Rock's theory.

REFERENCES

Helmholtz, H. (Originally published, 1867; 3rd edition trans., 1962). Handbook of Physiological Optics. New York: Dover.

McCloskey, M. and Watkins, M. J. (1978). The seeing-more-than-is-there phenomenon: Implications for the locus of iconic storage. Journal of Experimental Psychology: Human Perception and Performance. 4, 553-564.

Parks, T. E. (1965). Post-retinal storage. American Journal of Psychology, 78, 145-147.

Rock, I. (1983). The Logic of Perception, Cambridge, Mass: MIT Press.

Shepherd, G. M. (1974). The Synaptic Organization of the Brain. New York: Oxford University Press.

Trehub, A. (1977). Neuronal models for cognitive processes: Networks for learning, perception and imagination. Journal of Theoretical Biology, 65, 141-169.

Trehub, A. (in press). Visual-cognitive neuronal networks. In M. A. Arbib and A. R. Hanson (Eds.), Vision, Brain and Cooperative Computation. Cambridge: MIT Press, in press.

Zollner, F. (1892). Uber einer neue Art anorthoskopischer Zerrbilder. Annalen der Physik und Chemie, 27, 477-484.

AN ARCHITECTURE FOR MATHEMATICAL COGNITION

Kurt Ammon[1]
Department of Computer Science
University of Hamburg

ABSTRACT

This paper presents the architecture of the discovery system SHUNYATA which models studies and research in higher mathematics. SHUNYATA analyzes mathematical proofs and produces concepts and proof strategies which form the basis for the discovery of more difficult proofs in other mathematical theories. Its architecture avoids combinatorial explosions and does not require search strategies. The proof strategies contain two categories of predicates. A predicate of the first category selects a small set of proof steps and the predicates of the second category evaluate partial proofs and decide which predicate of the first category should be applied next. Thus, the proof strategies include feedback loops. A detailed example is given. It contains a simple proof in group theory, the analysis of this proof, and the discovery of a proof in lattice theory whose degree of difficulty represents the state-of-the-art in automated theorem proving. The most important result of this work is the discovery of a holistic logic based on the concept that cognitive structures arise from simple perceptions, evolve by reflection and finally contain their own evolution mechanisms.
Keywords: Learning, knowledge acquisition, cognitive evolution, automated theorem proving.

1 INTRODUCTION

Traditional research in machine learning assumes the existence of domain-independent and objectivizable cognitive structures and discovery mechanisms (e.g., [4]). This approach entails the following difficulties:

- The learning system cannot change its representation language and its structure (e.g., [11]).
- After a period of time, the efficiency of the system decreases drastically (e.g., [8]).

SHUNYATA automatically changes its language and its structure on the basis of experience which increases its efficiency. A crucial consequence is that it is impossible to objectivize its architecture, i.e., it cannot be analyzed completely.

The organization of this paper is as follows. Section 2 gives an overview of SHUNYATA. Section 3 describes the reflection system which forms the core of SHUNYATA. Section 4 introduces the concept of analytical spaces and Section 5 presents holistic logic. Sections 6 and 7 give a simple proof in group theory and the analysis of this proof. Section 9 describes the discovery of a difficult proof in lattice theory.

[1]Copyright © 1986 by Kurt Ammon. All rights reserved. Further, storage and utilization of the described programs on data processing installations is forbidden without prior written permission of the author.

2 OVERVIEW OF SHUNYATA

SHUNYATA has the structure of a tree in functional representation, i.e., every object is either a symbol or else has the form

$$f(a_1, ..., a_n),$$

where f is a function and a_1, ..., a_n are arguments. At the most general level SHUNYATA does not consist of interacting components but has only this functional structure which is very suitable for representing reasoning processes. SHUNYATA contains three kinds of functions:

- *Calculable functions.* These are conventional effectively calculable functions, i.e., they are objectivizable because they have a precise representation and produce precisely describable values.

- *Strange functions.* The value of a strange function is composed of the name of the function and its arguments. Finite sets, tuples, and bags can be regarded as strange functions. Example: The set $\{a_1, ..., a_n\}$ is represented by

  $$fset(a_1, ..., a_n),$$

  where the value of $fset(a_1, ..., a_n)$ is $fset(a_1, ..., a_n)$.

- *Holistic functions.* Holistic functions evolve through experience and cannot be objectivized. Their development is accompanied by division and unification processes. Existential quantification, universal quantification, and the operator that constructs sets from predicates are holistic functions. The undecidability of predicate logic [5] indicates that the decision procedure is a holistic function.

SHUNYATA contains elementary knowledge for evaluating holistic functions. In particular, it uses a basic procedure to evaluate predicates that define finite sets. The central mechanism of this procedure is the replacement of element relations by disjunctions of equality relations. In SHUNYATA holistic functions presently produce the value *not-evaluable* if the system does not contain sufficient knowledge for evaluating the function.

The most important functions of SHUNYATA are the *analyze* function and the *proof* function which perform the proof analyses and generate proofs. The *analyze* function has the form

$$analyze(C, T, P),$$

where C is a predicate calculus, T a theorem, and P a proof for the theorem. It produces a proof strategy. The *proof* function has the form

$$proof(C, T, S),$$

where C is a predicate calculus, T a theorem, and S a proof strategy. It produces a proof for the theorem. The *analyze* and *proof* functions illustrate the spiral organization of SHUNYATA: The *analyze* function constructs powerful proof strategies from proofs and the *proof* function constructs complex proofs from proof strategies.

3 THE REFLECTION SYSTEM

The reflection system is the core of SHUNYATA. It contains over one hundred simple functions and their relations which are represented in a language closely related to predicate logic. The universe of this language is the set of all symbols and all trees. The functions are meta-level concepts such as the subset relation, the union of sets, and the addition of natural numbers. Thus, the reflection system contains precise knowledge about finite trees.

An inference can be represented by

$$m = i(a_1, ..., a_n),$$

where m is a meta-level theorem, i is an rule of inference, and $a_1, ..., a_n$ are the arguments. The system uses predicates to construct small sets of inferences. These predicates are modified dynamically. The activity of the reflection system changes rapidly because it is controlled by the results of the inferences. The set of meta-level theorems generated by the reflection system is called *reflection space*. The architecture described avoids combinatorial explosions because the reflexion space can be pruned by additional predicates. This mechanism produces a combinatorial reduction of the size of the reflection space with regard to the number of predicates.

4 ANALYTICAL SPACES

Cognition permanently reduces huge amounts of information to a few concepts. Miller [10] argues that only some seven concepts are contained in short-term memory simultaneously. This reduction forces cognition to divide the world into analytical spaces which consist of a few concepts and the environment these concepts refer to. An analytical space represents a view of the world. Examples are geometry and space in mathematics, properties of programs and programs in computer science, and proof strategies and proofs in SHUNYATA. The principle of complementarity [3] states that quantum theory requires the existence of different analytical spaces. Gödel's Theorem [6] implies that there are different analytical spaces for the theory of natural numbers. Traditional epistemology regards the world and cognition as completely separable entities. From a holistic point of view the world and cognition are divided into different analytical spaces which contain essential but incomplete knowledge.

5 HOLISTIC LOGIC

The SHUNYATA system is the first step towards an implementation of holistic logic. The central hypotheses of this logic are:

- The basis or kernel of cognition is formless.
- Cognitive evolution creates new analytical spaces (*division*) and integrates existing analytical spaces (*unification*).

New information in a cognitive system generates elementary analytical spaces which cause perturbations in its structures. This process can be considered as low-level perception. In integrating the new analytical space the system must preserve its previous efficiency. It first tries to assimilate the new space into existing structures. If this is impossible, the new analytical space begins its own evolution, i.e., the system performs a division process. The evolution is controlled by the activity of the reflection system. Cognition achieves the most important advancements in its development by

integrating different analytical spaces, i.e., by unification processes. Even the high-level structures of the reflection system are permanently revised and improved by divisions and unifications on the basis of its low-level structures. Because cognitive structures arise from simple perception, their origin can be regarded as an empty or formless kernel, i.e., the kernel lacks properties and structures. Thus, the theory of holistic logic implies that it is impossibe to objectivize cognition and discovery mechanisms completely. The integration of new information can be regarded as an induction process which creates new structures. These are modified by future experience, i.e., by feedback processes. Therefore, cognitive structures evolve by induction and feedback [1].

6 A SIMPLE PROOF

This section gives a formalization of the elementary theory of groups, a theorem, and a complete formalization of a proof which includes the reasons for the theorems. This approach to formalization forms the basis for the analysis of the proof. It can be applied to predicate logic and meta-level reasoning. The symbols and functions used in this section are defined in the appendix.

1. *One binary predicate letter:* p. We write $x = y$ for $p(x, y)$.
2. *Three binary function letters:* f, g, h. We write xy for $f(x, y)$.
3. *Constant:* c.
4. *Axioms:* $(xy)z = x(yz)$, $g(x, y)x = y$, $xh(x, y) = y$.
5. *Rules of inference:*
   (a) Substitution Rule: $r = u \in E$, $s \in subs(r = u) \Longrightarrow sub(r = u, s) \in E$. Function: $sub(r = u, s)$.
   (b) Reflexivity Rule: $r \in R \Longrightarrow r = r \in E$. Function: $ref(r)$.
   (c) Symmetry Rule: $r = u \in E \Longrightarrow u = r \in E$. Function: $sym(r = u)$.
   (d) Transitivity Rule: $r = u \in E$, $u = v \in E \Longrightarrow r = v \in E$. Function: $tran(r = u, u = v)$.
   (e) Replacement Rule: $r \in R$, $d \in D$, $u = v \in E$, $d(r) = u \Longrightarrow rep(r, d, u = v) \in E$. Function: $rep(r, d, u = v)$.
   (f) Chain Rule: $r \in R$, $d \in D$, $u = v \in E$, $s \in subs(u = v)$, $d(r) = st(1, sub(u = v, s)) \Longrightarrow rep(r, d, sub(u = v, s)) \in E$. Function: $chain(r, d, u = v, s)$.
6. *Theorem:* $g(c, c)x = x$.
7. *Proof:*

Theorems	Reasons
$g(c,c)x = g(c,c)(ch(c,x))$	$chain(g(c,c)x,\ 2,\ y = xh(x,y),\ \{(x,c),(y,x)\})$
$g(c,c)(ch(c,x)) = (g(c,c)c)h(c,x)$	$chain(g(c,c)(ch(c,x)),\ (),\ x(yz) = (xy)z,$ $\{(x, g(c,c)),(y,c),(z,h(c,x))\})$
$(g(c,c)c)h(c,x) = ch(c,x)$	$chain((g(c,c)c)h(c,x),\ 1,\ g(x,y)x = y,\ \{(x,c),(y,c)\})$
$ch(c,x) = x$	$chain(ch(c,x),\ (),\ xh(x,y) = y,\ \{(x,c),(y,x)\})$

The proof steps containing the Symmetry Rule or the Transitivity Rule are omitted. The proof in ordinary scientific notation is

$$g(c,c)x = g(c,c)(ch(c,x)) = (g(c,c)c)h(c,x) = ch(c,x) = x.$$

7 THE ANALYSIS OF THE PROOF

This section describes the automatic analysis of the previous proof. The result of this analysis is a proof strategy.

The proof in the previous section contains four steps Y_1, Y_2, Y_3, and Y_4. These proof steps are tuples

$$(r = t, chain(r, d, u = v, s)),$$

where $r \in R$, $d \in D$, $u = v \in E$, and $d(r) = st(1, sub(u = v, s))$. SHUNYATA analyzes the four proof steps successively. The analysis produces two categories of predicates which represent the proof strategies. These predicates are generated from the functions of the reflection system like the terms and well-formed formulas of a predicate calculus. The predicates of the first category have the form

$$p(C, T, X, Y),$$

where C is a predicate calculus, T the theorem to be proved, X a tuple of proof steps, and Y a proof step. At the beginning of the analysis the tuple X is empty. SHUNYATA tests whether these predicates satisfy the following requirements:

- The evaluation of $p(C, T, X, Y_i)$ yields the truth value *true* where Y_i, $i \in \{1, 2, 3, 4\}$, is the proof step to be analyzed.
- The evaluation of $\{Y \mid p(C, T, X, Y)\}$ yields a small set of proof steps in a limited space of time.

The predicates of the second category have the form

$$q(C, T, X).$$

where C is a predicate calculus, T the theorem to be proved, and X a tuple of proof steps. They decide which predicate of the first category should be applied next. Therefore, the proof strategies include feedback loops. The proof steps generated by the predicates of the first category are appended to the tuple X until X contains the complete proof, i.e., the proof steps Y_1, Y_2, Y_3, and Y_4.

1. *The analysis of the first proof step.* SHUNYATA generates predicates and tests whether they satisfy the requirements described. It discovers the proof strategy:

$$\begin{array}{|l} r = st(1, T) \wedge \\ u = v \in axms(C) \cup \{b \mid (\exists a)(a \in axms(C) \wedge b = sym(a))\} \wedge \\ pr(2, s) \subseteq con(T) \cup var(T) \end{array}$$

   This proof strategy produces twelve proof steps.

2. *The analysis of the second proof step (division).* The previous strategy does not generate the second proof step. Thus, the second proof step can be considered as a perturbation with regard to the previous strategy. SHUNYATA tests further predicates but they fail to produce the two proof steps. Therefore, it has to perform a division process, i.e., it contructs a second predicate that produces the second proof step and a criterion that decides when the first and the second predicate should be applied. The division yields the strategy:

$$\begin{array}{|l} \ldots \\ \hline r \in pr((1,2), X) - pr((1,1), X) \wedge \\ u = v \in axms(C) \cup \{b \mid (\exists a)(a \in axms(C) \wedge b = sym(a))\} \wedge \\ ne(lvs(u)) = ne(lvs(v)) \end{array}$$

The dots represent the first predicate. The criterion states that the first predicate is applied only if the second predicate generates no proof steps, i.e.,

$$\{Y \mid p(C,T,X,Y)\} = \emptyset,$$

where p denotes the second predicate. Thus, the second predicate has priority. This strategy produces four additional proof steps.

3. *The analysis of the third proof step (unification).* It is very difficult to integrate the third proof step. The division process first fails because it yields a strategy that generates an infinite number of proof steps. Then, it produces a complex strategy with three predicates:

...
...
$r \in pr((1,2),X) - pr((1,1),X) \wedge$ $u = v \in axms(C) \cup \{b \mid (\exists a)(a \in axms(C) \wedge b = sym(a))\} \wedge$ $ne(lvs(u)) > ne(lvs(v)) \wedge$ $not(t \in pr((1,1),X) \cup pr((1,2),X))$

The dots represent the first and the second predicate. The third predicate has priority. Thus, the third proof step causes a strong perturbation in the previous strategy. Finally, SHUNYATA unifies the second and the third predicate and discovers a simple and efficient strategy:

...
$r \in pr((1,2),X) - pr((1,1),X) \wedge$ $u = v \in axms(C) \cup \{b \mid (\exists a)(a \in axms(C) \wedge b = sym(a))\} \wedge$ $ne(lvs(u)) \geq ne(lvs(v)) \wedge$ $not(t \in pr((1,1),X) \cup pr((1,2),X))$

The dots represent the first predicate. This strategy produces the third proof step.

4. *The analysis of the fourth proof step.* The previous strategy also produces the fourth proof step.

Thus, the proof strategy produces eighteen proof steps, i.e., eighteen theorems:

$g(c,c)x = g(g(c,c)c,c)x$			
$g(c,c)x = g(ch(c,c),c)x$			
$g(c,c)x = g(c,g(c,c)c)x$			
$g(c,c)x = g(c,ch(c,c))x$			
$g(c,c)x = g(g(x,c)x,c)x$			
$g(c,c)x = g(xh(x,c),c)x$			
$g(c,c)x = g(c,g(x,c)x)x$			
$g(c,c)x = g(c,xh(x,c))x$			
$g(c,c)x = g(c,c)(g(c,x)c)$	$\ldots = (g(c,c)g(c,x))c$		
$\underline{g(c,c)x = g(c,c)(ch(c,x))}$	$\underline{\ldots = (g(c,c)c)h(c,x)}$	$\underline{\ldots = ch(c,x)}$	$\underline{\ldots = x}$
$g(c,c)x = g(c,c)(g(x,x)x)$	$\ldots = (g(c,c)g(x,x))x$		
$g(c,c)x = g(c,c)(xh(x,x))$	$\ldots = (g(c,c)x)h(x,x)$		

The dots represent the terms on the right side of the preceding column. The proof is underlined. The first column is generated by the first predicate of the proof strategy and the other columns are generated by the second predicate.

The proof steps can be considered as new information for the system which can cause perturbations in its structures. The analysis of the proof steps produces predicates and proof strategies, i.e., new concepts. These predicates are generated like the terms and well-formed formulas of a predicate calculus and they are selected by experimentation. The predicates and proof strategies that were produced by analyzing the previous proof steps form the basis for the analysis of the next proof step. Thus, SHUNYATA changes its language and its structure on the basis of experience. The next section shows how the strategy for this simple proof in group theory can be used for discovering a difficult proof in lattice theory.

8 THE DISCOVERY OF A DIFFICULT PROOF

This section describes the discovery of a proof for SAM's Lemma which was an open problem until 1969. It was solved by Guard et al. [7] and subsequently by McCharen et al. [9]. The degree of difficulty for discovering a proof for SAM's Lemma represents the state-of-the-art in automated theorem proving [2]. The size of a proof in ordinary scientific notation is approximately one and a half pages [12].

1. *One binary predicate letter:* p. We write $x = y$ for $p(x, y)$.

2. *Two binary function letters:* f, g. We write xy for $f(x, y)$ and $x + y$ for $g(x, y)$. In order to omit parentheses, we assume that the first function has priority.

3. *Constants:* $0, 1, a, b, c, d$.

4. *Axioms:*

$(xy)z = x(yz)$	$(x + y) + z = x + (y + z)$	$xy = yx$	$x + y = y + x$
$xx = x$	$x + x = x$	$x(x + y) = x$	$x + xy = x$
$0x = 0$	$0 + x = x$	$1x = x$	$1 + x = 1$
$(a + b)c = 0$	$(a + b) + c = 1$	$(ab)d = 0$	$ab + d = 1$

$x + z = z \Rightarrow (x + y)z = x(y + z)$

5. *Theorem:* $(c + da)(c + db) = c$.

6. *Proof Strategy:* The proof strategy for the theorem in group theory is applied but the axiom that contains the implication requires special treatment: The consequence is treated like the other equations if the condition can be proved by the repeated application of the second predicate of the strategy. This strategy generates seventy-four theorems. The application of the first predicate may be regarded as an extension step and the repeated application of the second predicate as a simplification step. The extension step produces seventy-two theorems and the simplification step two theorems.

Extension	Simplification
$(c + da)(c + db) = (cc + da)(c + db)$	
...	
$(c + da)(c + db) = (c + da(a + b))(c + db)$	$(c + da(a + b))(c + db) = c$
...	
$(c + da)(c + db) = (c + da)(c + db(b + a))$	$(c + da)(c + db(b + a)) = c$
...	
$(c + da)(c + db) = (c + da)(c + d1b)$	

Some parentheses are omitted because of the associativity of the first function. The first proof is underlined.

7. *Proof*: The strategy produces the proof

$$(c + da)(c + db) = (c + da(a + b))(c + db) = c.$$

9 RELATED WORK

The first project concerned with automated mathematics research was AM [8, p. 137]. AM focuses on research in elementary mathematics, SHUNYATA on research in higher mathematics. AM discovers mathematical concepts, SHUNYATA additionally develops powerful cognitive structures, e.g., new ideas. AM uses rules to represent heuristic knowledge whereas SHUNYATA uses a modified predicate calculus, in particular predicates that generate finite sets of trees. AM contains sophisticated heuristic rules from the beginning and does not learn from experience. In contrast, SHUNYATA analyzes new information on the basis of a simple language closely related to predicate logic and develops concepts and discovery mechanisms on the basis of experience. AM does not generate proofs.

10 CONCLUSIONS

In this paper, I have described an idealized trace of cognitive evolution from the very beginning to a level that is suitable for research in higher mathematics. The origin of cognitive structures is a formless kernel. This formulation is a short paraphrase of the hypothesis that cognitive structures arise from simple perceptions, evolve by reflection and finally contain their own evolution mechanisms. Their development is accompanied by division and unification processes and creates an increasing objectivization of the environment. The discovery system SHUNYATA models these high-level cognitive processes and constructs more advanced theories from weaker ones. It is an evolving tree in functional representation. Its core is a reflection system which contains a language that is potentially universal because this language is permanently revised and improved. Cognitive structures are holistic in the sense that they cannot be reduced to simpler structures and that they cannot be completely separated from the objects they refer to, i.e., they are domain specific because they change on the basis of experience. The experiments suggest that artificial cognitive systems must have at least the computational capacity of human cognition and that they must use the same cognitive structures and the same organization.

SHUNYATA is written in ZETALISP and is presently running on a Symbolics 3600. The system has constructed many proofs from proof strategies, for example the previous proof for SAM's Lemma without special procedures for associativity and commutativity. If a representation similar to the representation humans use for associativity and commutativity is integrated into SHUNYATA, its efficieny increases by a factor of over two hundred. In generating finite sets of proof steps from predicates, the system produces more than one hundred simple LISP programs per minute and evaluates them. SHUNYATA has analyzed several proofs and has acquired concepts and strategies for these proofs, for example the strategy described in this paper. The proof analysis requires multiprocessing. The strategies are LISP programs which contain holistic functions and which are written and compiled by SHUNYATA. They can be regarded as heuristic knowledge which is modified on the basis of experience. Thus, the proof analysis involves the automatic and evolutionary development of programs by division and unification. The automatic revision and extension of the reflection system has not yet been implemented.

APPENDIX: NOTATIONS

$axms(C)$	The axioms of a predicate calculus.
$bag(a_1, ...a_n)$	The bag containing the elements $a_1, ..., a_n$. A bag is an unordered group of elements. Example: $bag(x, y, x)$.
$con(u = v)$	The constants of an equation $u = v$. Example: $con(g(c, c)x = x) = \{c\}$.
D	The set of subtree descriptors. Subtree descriptors are tuples of natural numbers that denote subtrees of trees. Examples: The subtree descriptor () denotes the subtree $f(a, b)$ of the tree $f(a, b)$, the subtree descriptor (2) the subtree $g(b, c)$ of the tree $f(a, g(b, c))$, and the subtree descriptor $(2, 1)$ the subtree b of the tree $f(a, g(b, c))$. We write 1 for the subtree descriptor (1) and 2 for the subtree descriptor (2).
E	The theorems of a predicate calculus.
$lvs(t)$	The bag of leaves of a tree t. Example: $lvs(f(g(x, y), x)) = bag(x, y, x)$.
$ne(b)$	The number of elements in a bag b. Example: $ne(bag(x, y, x)) = 3$.
$pr(d, t)$	The projection of a set of trees t described by the subtree descriptor d. Examples: $pr(1, \{(a, b), (c, d)\}) = \{a, c\}$, $pr((1, 2), \{(f(a, b), c), (f(d, e), c)\}) = \{b, e\}$.
R	The terms of a predicate calculus.
$rep(r, d, u = v)$	The replacement of the subterm of a term r that is selected by the subtree descriptor d by the term v of the equation $u = v$. Example: $rep(g(c, c)x, 2, x = ch(c, x)) = g(c, c)(ch(c, x))$.
$st(d, t)$	The subtree of the tree t that is selected by the subtree descriptor d. Examples: $st((), f(a, b)) = f(a, b)$, $st(2, f(a, b)) = b$, $st((1, 2), f(g(a, b), c)) = b$.
$sub(u = v, s)$	The application of the substitution s for the variables in the equation $u = v$ to the equation $u = v$. Example: $sub(y = xh(x, y), \{(x, c), (y, x)\}) = x = ch(c, x)$.
$subs(u = v)$	The substitutions for the variables of an equation $u = v$. Example: $\{(x, c), (y, x)\}$ is a substitution for the variables of the equation $y = xh(x, y)$.
$var(u = v)$	The variables of an equation $u = v$. Example: $var(g(c, c)x = x) = \{x\}$.

ACKNOWLEDGEMENTS

I wish to thank Wilfried Brauer, Bernd Neumann, and Hans-Joachim Novak for providing constructive comments on an earlier draft of this paper. I also wish to thank Ernie Cohen and John McCarthy for helpful conversations. Special thanks to Russell Block for improving my English.

REFERENCES

1. Ammon, K. The Automatic Discovery of Concepts by Induction and Feedback: Towards a Theory of Intelligence. In H. Trost and J. Retti (Eds.) *Österreichische Artificial Intelligence-Tagung*. Springer-Verlag, Berlin and Heidelberg, 1985.

2. Antoniou, G., and Ohlbach, H. J. TERMINATOR. *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, Karlsruhe, West Germany, 1983.

3. Bohr, N. The Quantum Postulate and the Recent Development of Atomic Theory. *Nature* 121 (1928), pp. 580-590.

4. Carbonell, J. G., Michalski, R. S., and Mitchell, T. M. An Overview of Machine Learning. In R. S. Michalski, J. G. Carbonell, and T. M. Mitchell (Eds.) *Machine Learning: An Artificial Intelligence Approach*. Tioga, Palo Alto, 1983.

5. Church, A. A Note on the Entscheidungsproblem. *The Journal of Symbolic Logic* 1 (1936), pp. 40-41. Correction, ibid., pp. 101-102.

6. Gödel, K. Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatshefte für Mathematik und Physik* 38 (1931), pp. 173-198.

7. Guard, J. R., Oglesby, F. C., Bennett, J. H., and Settle, L. G. Semi-Automated Mathematics. *Journal of the ACM* 16 (1969), pp. 49-62.

8. Lenat, D. B. AM: Discovery in Mathematics as Heuristic Search. In R. Davis, and D. B. Lenat (Eds.) *Knowledge-Based Systems in Artificial Intelligence*. McGraw-Hill, New York, 1982.

9. McCharen, J. D., Overbeek, R. A., and Wos, L. A. Problems and Experiments for and with Automated Theorem-Proving Programs. *IEEE Transactions on Computers* C-25 (1976), pp. 773-782.

10. Miller, G. A. The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Infomation. *Psychological Review* 63 (1956), pp. 81-97.

11. Mitchell, T. M., Utgoff, P. E., and Banerji, R. Learning by Experimentation: Acquiring and Refining Problem-Solving Heuristics. In R. S. Michalski, J. G. Carbonell, and T. M. Mitchell (Eds.) *Machine Learning: An Artificial Intelligence Approach*. Tioga, Palo Alto, 1983.

12. Ohlbach, H. J. The Markgraf Carl Refutation Procedure: The Logic Engine. Interner Bericht 24/82, University of Karlsruhe, 1982.

CLUSTER: An Approach to Modeling Context*

Yigal Arens

Computer Science Department
University of Southern California

1. Introduction: CLUSTER and the Context of an Utterance

This paper outlines a theory of modeling the context of an utterance. It is part of the CLUSTER** theory of language understanding in context (Arens, 1986). It has been implemented in a version of UC (the UNIX Consultant) (Wilensky, Arens, & Chin, 1984).

CLUSTER builds on the following observations concerning the process of language understanding:

- Language understanding is not sentence-based.
- New input is understood in reference to the context, not necessarily the previous text.
- The context is updated and changed in the course of processing input. Insertion of a fact into the context may cause the insertion of a group of associated facts.
- Elements of the context have a degree of salience associated with them.
- This salience is a status that decays over time and may eventually be lost.
- Information in the context should be recorded periodically in the more permanent form.

CLUSTER theory posits several levels of memory structures. At the lowest level are *entries,* 'atomic' representations of elements of the context. Entries are grouped together in *clusters.* Such a grouping indicates that the entries tend to co-occur in the world. A collection of clusters is stored in an understanding system's *Long Term Memory.* The mention of an entry in an input sentence, say, will result in the recall of clusters containing it. The recalled clusters will then be placed in *the Context Model,* where they will be available to any processes that require knowledge of the context.

1.1. Entries

There are three general classes of entries. Each contains representations of different aspects of the world within which the understanding system operates.

Objects – Entries in this class stand for physical objects, events, and actions. An entry of this type must be present in the Context Model before a reference to such an object may be understood.

Assertions – Entries in this class describe states of the world, state-changes, or actions that have taken place. An entry of this type must be present before the effects of the thing it represents may be accessed.

Intentions – These entries describe intentions the understanding system has of performing certain actions. The actions the system can perform include output to the user (i. e., saying something), adding or deleting an existing entry in the Context Model, and more.

* Some of the work reported here was performed when the author was at the University of California at Berkeley, where it was sponsored by the Office of Naval Research under contract N00014-80-C-0732, and DARPA (DOD), ARPA order No. 3041, monitored by the Naval Electronic Systems Command under contract N00039-82-C-0235. The work performed at the University of Southern California was supported by the USC Faculty Research and Innovation Fund, and by TRW Defense Systems under Subcontract No. D71810AN5S.

** Contextual Language Understanding with Salience Tracking and Environment Recall.

The same real-world object or event may be represented by an entry of more than one type. For example, the system's intention to say something to the user may also be represented in an entry of type *object*. This would permit the system to understand references to this intention in the user's input, a situation that would not be possible otherwise.

1.2. Clusters

Entries do not, as a rule, occur in isolation. A collection of entries is arranged in a cluster if those entries tend to be found together in the real world. The choice of *associatedness* as the basis for cluster structure is supported by psychological evidence of its importance. There is considerable psychological evidence for the existence of an automatic process that makes information available on the basis of associative relatedness (McKoon & Ratcliff, 1979) (Neely, 1977) (Warren, 1977) (and see Anderson (1983) for a review). The co-occurrence of entries signified by their forming a cluster is usually due to the fact that they represent one of the following:

- the various parts of a complex object;
- the several steps of a complex action;
- the actors and instruments of a particular activity;
- intended or associated uses of a particular object;
- standard associations between certain stimuli and responses, causes and effects; or
- actions related by convention (e. g., an inquiry into whether one can pass a salt shaker and a request that the salt shaker actually be passed).

For example, a cluster from the UNIX domain relating the making of a file executable to the command that must be issued in order to do so will include the following entries*:

?F is a file
?U issues command 'chmod' with argument '755'
?F changes its 'executability' state to 'on'

The existence of this cluster in long term memory asserts, among other things, that the last two entries are related, i. e., issuing the command 'chmod 755 <filename>' is associated with making a file executable, and vice versa. Mentioning any one of these two facts in a conversation causes the second to become present in the Context Model, hence in context.

A more complex cluster, containing a variety of entries is displayed below:

?A is an action
?U has asked how to perform action ?A
?U wishes to know how to perform ?A
?U cannot perform action ?A
System intends to find plan for ?A
?U expects a response

This cluster reflects the association between the user asking how to perform some action, the user's wish to know how to perform it, and the system's intention of finding such a plan and informing the user of it.

* Entries are presented in a simplified notation. ?F and ?U are variables, allowing this cluster to be used regardless of who the particular user (?U) is, and which file (?F) is involved.

1.3. Activation

Entries in the Context Model are marked with a measure of their salience in the context, indicated by a level of activation. In CLUSTER, the level of activation is viewed as a point on a scale, thus distinguishing it both from Quillian's (1969) and Charniak's (1983) marker passing models. The processing of input is accompanied by the activation of entries representing concepts mentioned in it and a flow of activation from those entries to others in their cluster(s), and so forth. Through the accumulation of activation flow from several sources, the Context Modeler has the ability to identify entries that are salient in the current context, although not closely associated with any single entry.

Activation flow is *cluster-directed*. When an entry's activation level is increased, all clusters to which it belongs are identified. The entry's increase is divided by the number of clusters and that amount flows on to each cluster. Within each of the clusters, however, every entry receives the same increase, regardless of the size of the cluster. In this manner, CLUSTER distinguishes between entries associated with many different situations (e. g., a file and its numerous uses) and entries associated with only a few (e. g., the notion of execution). In the latter case we would like to infer that the current context is probably one of those few clusters/situations, regardless of the number of entries in each. Thus, when 'execution' is mentioned we can realize that there is some program being discussed regardless of the number of facts we know about the execution of programs.

The level of activation of an entry in the Context Model influences the choices made by processes that need to know about the context. As a rule, a process in search of a particular type of entry will choose the one of that type with the highest level of activation. See section 2. for examples.

Unless reinforced, as described above, the activation of all entries in the Context Model decays gradually with the passage of time. When the activation of an entry decreases below a certain threshold it is dropped from the Context Model.

Entries of the class *intention* are triggered when their activation grows high enough. Their intended action is then carried out.

1.4. Processing Outline

New entries are passed on to the Context Modeler by outside processes. In the current implementation, the only such process is the language analyzer. It occasionally identifies a fragment of the input language that is meaningful to it and passes its meaning to the Context Modeler. The latter forms an entry representing the meaning of the language fragment. The following process then takes place:

1. The Context Modeler attempts to unify the new entry with those already existing in the Context Model. Either a match is found or else a new entry is created in the Context Model.

   If an existing entry is found:

2a. All clusters of which the existing entry is a member are reinforced, with activation flow continuing on from those.

   Otherwise:

2b. All clusters associated with the new entry are retrieved from long term memory and their members are inserted recursively.

3. This process continues until stability is reached.

4. The Context Model is inspected and highly activated *intentions* are performed.

At this point the Context Modeler awaits new input from the language analyzer.

2. Usefulness of the Context Model to Language Analysis

In a conversational context the language analyzer cannot treat sentences as independent entities. Rather, it must relate their content and import to information gathered during the processing of previous utterances. In particular, objects, actions, and any other concepts mentioned in, or whose existence is implied by, an utterance must be disambiguated. Their interpretations must be identified from among those concepts present in the Context Model. For this to take place the language analyzer must be capable of identifying meaningful fragments in the input and passing them on to the Context Modeler as soon as possible.

A language analyzer with these characteristics is used in the implementation of CLUSTER. This is PHRAN, the PHRasal ANalyzer (Wilensky and Arens, 1980a, 1980b). Upon recognizing a phrasal construct in the input, PHRAN constructs its paired concept and passes it to the Context Modeler. The concept becomes part of the entry representing the meaning of the recognized language fragment.

2.1. Ambiguity Resolution

When PHRAN processes input it often encounters ambiguous language fragments. In such instances entries representing the meaning of *all* interpretations are inserted in the Context Model. In the absence of any selectional restrictions or other linguistic clues, or if such restrictions are insufficient to uniquely determine the meaning, PHRAN chooses the most highly activated entry of the appropriate type as representing the meaning of the ambiguous fragment. Naturally, the presence of related clusters in the Context Model will cause an entry's activation to be higher.

2.2. Definite References

By itself, a language analyzer is incapable of determining precisely which real world entities definite references refer to. This is because the understanding of most definite references inherently involves the use of information not present in the sentence being processed.

The Context Model provides a pool of candidates for the interpretation of referring expressions. As a rule, the language analyzer will choose as the referent the most highly activated entry consistent with the restrictions provided by the referring expression. When no such entry can be found, the Context Modeler realizes that a new entity is being discussed. It then creates an appropriate new entry and inserts it in the Context Model.

2.3. Pronouns and Demonstratives

As is the case with definite references, the language anlyzer searches the Context Model for a highly activated entry consistent with the restrictions derivable from the pronoun or demonstrative used. These will often include gender information.

3. Processing Example

This section contains an example of the operation of the UNIX Consultant (UC). I will sketch the processing involved in engaging in the dialogue that follows. I will pay special attention to the determination of the referent of the word 'it' in line [5].

[1] User: How do I print the file fetch.l?
[2] UC: To print the file fetch.l type 'lpr fetch.l'.

. (intervening commands and questions)

[3] User: Has the file fetch.l been printed yet?
[4] UC: The file fetch.l is in the line printer queue.
[5] User: How can I cancel it?
[6] UC: To remove the file fetch.l from the line printer queue type 'lprm arens'.

The fragments recognized by PHRAN in the user's first question, [1], are, in the order in which they are identified:

I
the file fetch.l
How do I print the file fetch.l ?

The analysis of each fragment results in a structure representing its meaning. This structure is inserted in the Context Model. The third fragment recognized is a request concerning printing. Its processing results in a complete cluster being retrieved from long term memory and inserted in the Context Model. The cluster contains structures describing the command to print, its effect, the object printed and the fact that a printer is involved. This information is used to determine the answer provided by the system in [2].

After UC's reply, the Context Model contains representations of the following objects and actions:

(1) The user
(2) The file fetch.l
(3) The line printer
(4) The request to print fetch.l
(5) The command to print the file ('lpr fetch.l')
(6) A printing event
(7) The user has requested to print the file
(8) UC has told the user how to print the file

The contents of the Context Model are now preserved in long term memory in the form of a new cluster. This is done following an explicit request by the user, as this process is not currently automatic. The new cluster will be recalled when the representation of any of the entries (2), (4), (5), or (7), is encountered again.

The conversation now shifts to other topics and the structures previously in the Context Model are gradually deleted due to decay of their activation levels. By the time the user asks question [3], there is no trace left in the current context of the structures used in the earlier exchange. For the sake of this example, I will assume that when line [3] is input by the user the Context Model is empty.

When the user's second question is processed by PHRAN, the following fragments are recognized:

the file fetch.l
Has the file fetch.l been printed yet?

When an entry describing the first one is created in the Context Model, it serves as a key for retrieving the new cluster described above. The concept associated with the second fragment recognized causes the

retrieval of other clusters. These clusters enable the system to respond with [4], but are otherwise irrelevant to this example.

When question [5] is asked by the user, it is analyzed by PHRAN and the following input fragments are identified:

I
it
How can I cancel it?

The structure created to describe the first fragment is identified with the structure describing the user, which is already present in the Context Model.

When a structure is created to represent the meaning of the second fragment, *it*, no determination of its precise meaning by UC is possible. However, when the last fragment is recognized, PHRAN notes that the referent of *it* must be a command. The only command found in the Context Model is the one to print the file fetch.l. This was entry (5) in the cluster formed after processing question [1], which was stored in long term memory and retrieved during the processing of question [3] above. The identification of this entry as the referent enables UC to correctly interpret the user's question and eventually to provide the answer [6].

4. References

Anderson, J. R., 1983. A Spreading Activation Theory of Memory. *Journal of Verbal Learning and Verbal Behavior,* 22, pp. 261-295.

Arens, Y., 1986. CLUSTER: An Approach to Contextual Language Understanding. Ph. D. Thesis. Report No. UCB/CSD 86/293, Computer Science Division (EECS), UC Berkeley. April 1986.

Charniak, E., 1983. Passing Markers: A Theory of Contextual Influence in Language Comprehension. *Cognitive Science* 7, pp. 171-190.

McKoon, G., and Ratcliff, R., 1979. Priming in Episodic and Semantic Memory. *Journal of Verbal Learning and Verbal Behavior,* 18, pp. 463-480.

Neely, J. H., 1977. Semantic Priming and Retrieval from Lexical Memory: Roles of Inhibitionless Spreading Activation and Limited-capacity Attention. *Journal of Experimental Psychology: General,* 106, pp. 226-254.

Quillian, M. R., 1969. The Teachable Language Comprehender: A Simulation Program and Theory of Language. In *Communications of the ACM,* 12(8), August 1969, pp. 459-476.

Warren, R. E., 1977. Time and Spread of Activation in Memory. *Journal of Experimental Psychology: Learning and Memory,* 3, pp. 458-466.

Wilensky, R., and Arens, Y., 1980a. PHRAN: A Knowledge-Based Approach to Natural Language Analysis. Memo UCB/ERL M80/34, Electronic Research Laboratory, UC Berkeley, August 1980.

Wilensky, R., and Arens, Y., 1980b. PHRAN: A Knowledge-Based Natural Language Understander. In *Proceedings of the 18th Annual Meeting of the Association for Computational Linguistics,* Philadelphia, PA. June 1980.

Wilensky, R., Arens, Y., and Chin, D., 1984. Talking to UNIX in English: An overview of UC. In *Communications of the ACM.* June, 1984.

EFFECTS OF FOCAL BRAIN DAMAGE ON CATEGORIZATION OF VISUAL AND HAPTIC FEATURES

Bruce Bolster
University of Winnipeg
Jane Honeyman
University of Winnipeg
Valerie Gillis
University of Winnipeg
Martha Wilson
University of Connecticut

ABSTRACT

Previous research has shown that stimulus values on a sensory continuum are perceived in a categorical manner by human subjects and by rhesus monkeys (Wilson, 1972; Streitfeld & Wilson, in press). That is, stimuli which are judged to belong to different perceptual categories are discriminated more accurately than are stimuli which are perceived as belonging to the same perceptual category. In these experiments, the category boundary was defined as the adaptation level (AL) established by the stimulus series presented to the subject. Wilson and DeBauche (1981) showed that resection of visual "association cortex" in monkey abolished categorical perception of visual features and they hypothesized that modality-specific neural substrates that preserve the effects of stimulation provide an internal referent which determines the manner in which given stimulus values are identified and discriminated. In the study described here, the effects of focal brain damage on categorical perception of three stimulus continua was examined in neurological patients. Processing of visual features as members of perceptual categories was doubly dissociated from processing of haptic stimuli; posterior lesions in the right hemisphere selectively impaired categorization of visual stimuli differing in length and orientation while anterior lesions in the left hemisphere selectively impaired categorical perception of weight. Implications for the neural dynamics of categorization are discussed in the context of AL theory and principles of neural organization.

BACKGROUND

Adaptation-level theory (Helson, 1948, 1964) postulates that perceptual judgments in a given sensory domain are based on an internal standard, the adaptation level (AL), the value of which is determined by a weighted, logarithmic mean of focal, background and residual stimulation. The AL represents a neutral point of functioning, i.e., the level of stimulation to which the organism is adapted or habituated. It follows that stimuli whose values coincide

with the physical value representing the AL will elicit a null response while stimulus values that differ from the AL will elicit perceptual responses which depend upon the direction and magnitude of their discrepancy from the AL. In this sense, the AL can be equated with a category boundary, structuring a stimulus continuum into two complementary perceptual classes. For a series of stimuli differing in length, for example, stimuli whose physical values lie above the AL would be perceived as "long" while stimuli whose values lie below the AL would be perceived as "short". In studies with rhesus monkeys, Wilson (1972) showed that the discriminability of stimuli, as well as the way in which stimuli are identified (i.e., assigned to perceptual classes), is a function of their relationship to the AL. Two stimuli that lay on different sides of the AL (a "long" vs. a "short" stimulus) were discriminated more accurately than two stimuli that lay on one or the other side of the AL (two "long" or two "short" stimuli) even though the physical difference between the stimuli in the pairs was the same.

Subsequently, Wilson & DeBauche (1981) showed that lesions of visual association cortex, but not primary visual cortex, in monkeys interfered with categorical perception of three visual continua. It was hypothesized that portions of the extrastriate visual system, including inferotemporal cortex and the pulvinar nucleus of the thalamus, serve to maintain reverberatory circuits which sustain the effects of visual experience over time. Such a system would yield an experience-sensitive "set-point" or frame of reference against which visual input is compared.

The experiments reported here were designed to further test the hypothesis that categorical perception in humans, as in monkeys, is vulnerable to forebrain damage which involves cortical areas that lie outside primary sensory cortex (Wilson, in press). To this end, we examined visual and tactual-kinaesthetic discriminative abilities in patients with verified damage to restricted brain areas. While our data do not allow us at this time to contrast the effects of damage to primary sensory cortex vs. intrinsic or "association" cortex, we can address the issue of whether modality-specific impairments in categorization follow from damage to specific extrasensory brain areas.

METHOD

Subjects.

Neurosurgical patients who had sustained focal forebrain damage due to vascular accident, trauma, or removal of localized tumor, served as subjects on a voluntary basis. Not all patients participated in all tests. Subjects were recruited at one of two centers, either Hartford, CT or Winnipeg, Manitoba and were identified by medical staff as having sustained damage appropriate to the

investigation. Location and extent of brain damage was confirmed by reference to radiological data and/or surgical reports.

Procedure.

Four sensory continua were investigated for evidence of presence or absence of categorical perception as a function of locus of brain damage. Stimuli on two visual continua, line length and orientation (angular disparity from the horizontal), and two tactual-kinaesthetic continua, line length and weight, were presented. Data for tactual judgments of line length are not presented here as only a small number of patients have been tested on this task to date. For judgments of visual length, black bars 2 mm thick which ranged in length from 33-38 mm in 1 mm increments were mounted on 10 cm^2 white cardboard. For the orientation task, bars of equal length were varied in orientation from 38-48 deg in 2 deg increments. The stimuli for judgments of weight were small vials of identical size and appearance which were filled with lead shot and cotton packing, and which ranged from 100 to 150 gr in 10 gr increments.

For all tasks, the six stimuli on the continuum were presented in pairs representing all possible combinations. The subjects were instructed to indicate by pointing to one of the stimuli in a given pair which was judged to be longer, more horizontal, or heavier. Visual stimuli were exposed for 2 sec. For weight judgments, subjects lifted each stimulus once prior to making a response. The left-right position of each stimulus and the position of the correct response were balanced over trials. No information was provided about the correctness of response. The 15 possible combinations of the six stimuli on each continuum were each presented five times, making a total of 75 trials for each stimulus condition.

Data Analysis.

The category boundary was calculated for each subject individually. For the visual length continuum, for example, the percentage of choices of each stimulus as "longer" was computed, and the stimulus value that was chosen 50% of the time was identified as the AL (the category boundary between stimulus values that were perceived as "long" and those that were perceived as "short"). This value fell between the two middle stimulus values in the set for most subjects, consistent with AL theory which predicts that the pooled effect of the physical values in the series presented determines the neutral point in the absence of background or residual factors. After the location of the boundary was established, percentage correct response was calculated for pairs of stimuli in which both stimuli lay on one side of the boundary (Within-category discriminations), and for pairs in which the stimuli straddled the category boundary (Between-category discriminations). This was done for one-step and two-step differences between the stimulus values on each sensory continuum.

RESULTS AND DISCUSSION

Mean percentage correct discrimination for Between-category pairs vs. Within-category pairs is shown in Table 1 for the two visual discrimination tasks and for weight discrimination. It can be seen that for visual length discrimination, neither left nor right-hemisphere frontal damage affected categorical perception while performance was severely compromised by damage to right temporal cortex. Similarly, damage to right temporal and right parietal-temporal cortex selectively affected categorical perception of visual orientation, in contrast to lesions in other areas which did not interfere with categorization. On the weight discrimination task, both left frontal and left parietal-temporal damage appeared to abolish categorical perception but the results for judgments based on use of the hand ipsilateral to the lesion suggest that left frontal cortex is the critical area for categorization of this dimension. That is, the data from the ipsilateral hand are more compelling in terms of localizing the neural substrate for categorization of a weight continuum since such judgments are less likely to be degraded by direct effects on the sensory projection system. It should be noted that Within-category discrimination accuracy is not impaired in those groups which fail to show the category-boundary effect.

These data provide evidence on several points. First, Between-category judgments are more accurate than Within-category judgments for most patients on most of the tasks, indicating significant categorical perception of nonverbal stimuli in spite of forebrain damage. The effects are similar in degree to those exhibited by neurologically-intact subjects tested by Streitfeld & Wilson (in press). Thus, categorical perception appears to be a relatively robust phenomenon from a neuropsycholgical perspective in the sense that it is not vulnerable to all brain injury. On the other hand, deficits in categorization in specific sensory domains resulted from damage to localized areas of association cortex. Perception was not categorical for either visual length nor visual orientation following damage to the right temporal cortex, in contrast to categorical perception of weight which was affected by damage to the left frontal areas of the brain. Patients with damage to parietal-temporal cortex in the right hemisphere were impaired in categorical perception of visual orientation but not visual length. These results suggest that categorization of stimulus features is a modality-specific process that occurs independently within portions of the thalamocortical system devoted to each sensory channel. Moreover, the results are consistent with the notion that categorical perception depends upon stored representations of experience in a particular sensory domain since the AL defined the category boundary for each of the sensory continua examined.

Taking previous findings into account, the data suggest further that damage to intrinsic or "association" areas in the frontal,

parietal, and temporal lobes is sufficient to produce impairments in categorical perception, independent of primary sensory deficit.

REFERENCES

Helson, H. (1948). Adaptation level as a basis for a quantitative theory of frames of reference. Psychological Review, 55, 297-313.

Helson, H. (1964). Adaptation-level theory: An experimental and systematic approach to behavior. New York: Harper & Row.

Streitfeld, B. & Wilson, M. (in press). The ABC's of categorical perception. Cognitive Psychology.

Wilson, M. (1972). Assimilation and contrast effects in visual discrimination by rhesus monkeys. Journal of Experimental Psychology, 93, 279-282.

Wilson, M. (in press). Brain mechanisms in categorical perception. In S. Harnad (Ed.), Categorical perception. Cambridge: Cambridge University Press.

Wilson, M. & DeBauche, B. A. (1981). Inferotemporal cortex and categorical perception of visual stimuli by rhesus monkeys. Neuropsychologia, 19, 29-41.

AUTHOR NOTES

The research reported herein was supported by grants MH 36582 from the National Institute of Mental Health and A2560 from the Natural Sciences and Engineering Council of Canada. We thank the staff members of Hartford Hospital, and the Health Sciences Center and the St. Boniface General Hospital, Winnipeg, for cooperating in the study.

Address correspondence to: Dr. Bruce Bolster, Department of Psychology, University of Winnipeg, 515 Portage Avenue, Winnipeg, Manitoba, Canada, R3B 2E9.

TABLE 1

MEAN PERCENTAGE CORRECT DISCRIMINATION FOR COMBINED ONE-STEP AND TWO-STEP DIFFERENCES BETWEEN STIMULI ON THREE SENSORY CONTINUA.

Lesion Group	N	Between-category	Within-category	Between-Within difference
		Visual Length		
LF	4	98	78	20
LPT	3	85	77	8
LT	2	93	70	23
RF	8	95	72	23
RPT	7	82	69	13
RT	5	75	73	2
		Visual Orientation		
LF	2	83	74	9
LPT	2	85	70	15
LT	1	100	80	20
RF	3	83	64	19
RPT	4	77	77	0
RT	4	76	77	-1
		Weight (contralateral hand)		
LF	2	75	71	4
LPT	3	62	76	-14
LT	4	80	69	11
RF	7	80	68	12
RPT	8	77	66	11
RT	4	75	63	12
		Weight (ipsilateral hand)		
LF	3	75	77	-2
LPT	2	89	77	12
LT	3	97	75	22
RF	6	84	68	16
RPT	5	81	71	10
RT	4	88	73	15

NOTE: Lesion groups are LF (Left Frontal); LPT (Left Parietal-Temporal); LT (Left Temporal); RF (Right Frontal); RPT (Right Parietal-Temporal); RT (Right Temporal).

BELIEF MAINTENANCE WITH UNCERTAINTY

A. Julian Craddock & Roger A. Browse

Department of Computing and Information Science
Department of Psychology
Queen's University at Kingston
Ontario, Canada

ABSTRACT

A framework for representing and reasoning with uncertain information is described. A network knowledge structure is used which makes the reasons for believing or not believing a proposition explicit. These reasons, or endorsements, are quantified by a measure of belief and certainty. Heuristics are integrated with the knowledge structure to collect, and evaluate the endorsements.

1. INTRODUCTION

The research reported in this paper pursues the problem of developing representational and inference mechanisms which are capable of dealing with incomplete, inaccurate, and uncertain information. The direction taken is based on the assumption that methods which deal effectively with uncertainty must play an integral role in both models of human reasoning, and flexible computational reasoning systems.

Most formal reasoning systems combine both the extent of belief and certainty of belief into a single truth value, whether binary or multi-valued (McCarthy 1980; McDermott & Doyle 1980; Reiter 1978a; Zadeh 1983). In many cases, this compression is justified, but consider the proposition RICK LIKES MATH. The extent of belief in this proposition may be high whether it is quite certain (Rick has taken, and enjoyed a wide range of math courses) or quite uncertain (Rick has only taken a single math course).

Recently Cohen (1983) has formulated a model of reasoning which maintains that reasons for believing or disbelieving propositions can be collected, providing a more comprehensive description of belief. Our approach is to employ a knowledge structure such that these reasons, or endorsements, are made explicit. The endorsements for propositions can be quantified by a measure of belief and certainty. In addition, a network of endorsements among propositions may be used to: (1) determine how supportive a body of evidence for a particular hypothesis is and (2) represent evidential relationships such as conflicts between decisions (Craddock 1986).

The algorithms which compute the belief and certainty of a proposition may be formulated to operate uniformly on all supporting knowledge, or the algorithms may be subject to heuristics which emphasize the importance of selected portions of the supporting knowledge. In the development of heuristic methods we have been guided by the approach taken by Kahneman and Tversky (1982a, b) Their model indicates that humans employ a set of basic heuristics which aid in making decisions in conditions of uncertainty. These heuristics enable humans to constrain problem domains such that the uncertainty becomes manageable but still useful. In addition humans can also employ heuristics to determine complex evidential relationships between different sources of evidence.

2. BELIEF AND CERTAINTY

The development of our reasoning system requires the definition of a network containing nodes which are propositions with associated belief and certainties. Interconnections in the network represent the support that one proposition offers another. First, let P be a set of cognitive units $P = \{n_1 \ldots n_m\}$. Each of these cognitive units may represent a proposition such as I LIKE MATH or relationships among objects or concepts. Each n_i has associated with it a belief strength b_i, which is a measure of the extent to which the cognitive unit is believable. The believability of n_i is a measure of the strength of the supporting evidence for n_i, not a measure of its incidence of occurrence or its possibility of occurrence. The belief strength b_i can be defined as follows: A cognitive unit n_i is believable if there is an endorsement for n_i or if the endorsements supporting n_i are stronger than those against it. As $-1 \leq b_i \leq 1$, we may view the cognitive units as statements in fuzzy propositional logic in which a belief of -1 indicates n_i is false and a belief of +1 indicates n_i is true.

In addition, each n_i has associated with it a certainty, c_i of the assignment of the value b_i, where $0 \leq c_i \leq 1$. The certainty of a belief value is defined as a measure of the reliability or trustworthiness of the evidence which was used to calculate a particular belief (Hamburger, 1985). Thus each cognitive unit represents two distinct aspects of the *Rational* $R_i = (b_i, c_i)$ of n_i.

Any cognitive unit may endorse another. For example, a cognitive unit representing I LIKE COMPUTING may be endorsed by I LIKE ALGEBRA, I LIKE PHYSICS, and the negative endorsement I HAVE TROUBLE WITH TECHNICAL MANUALS. Each endorsement has associated with it a numeric value corresponding to the extent of the support between the units. If n_i endorses n_j then the support node s_{ij} is the *support for the endorsement* where $-1 \leq s_{ij} \leq 1$. If $-1 \leq s_{ij} < 0$ then the endorsement n_i for n_j is said to be inhibitory and if $0 \leq s_{ij} \leq 1$ then it is said to be excitatory. A diagrammatic version is supplied in figure 1. The support nodes for the endorsement, s_{ij} may be endorsed by other cognitive units. For example: I LIKE PSYCHOLOGY may endorse I LIKE COMPUTING, but the support for the endorsement may be contingent on COMPUTATION MAY MODEL COGNITION. (see figure 2). If the belief in COMPUTATION MAY MODEL COGNITION is false then the support for the endorsement will decrease.

If we consider $P = \{n_1 \ldots n_m\}$ as the set of propositional nodes of the network, then we can define $S' = \{s_{ij} \mid n_i, n_j \in P, s_{ij} \neq 0\}$ as a subset of support nodes such that n_i endorses n_j with support s_{ij}, and $T = \{t_{s_{ij}} \mid s_{ij} \in S', n_k \in P\}$ as the other subset of support nodes such that n_k endorses s_{ij} with support $t_{s_{ij}}$. We can then define the network $N = <P, S>$ where $S = S' \cup T$ is a finit set of support nodes representing the arcs and P is a finit set of propositions.

3. COMPUTING BELIEF AND CERTAINTY VALUES

We wish to develop ways of computing the values of *Rational* R_j for a proposition n_j on the basis of the endorsements available for that node. The first important observation is that the strength of endorsement between two nodes n_i and n_j is not only dependent on s_{ij}. This strength must be computed with consideration of b_i. We can compute this endorsement strength as $e_{ij} = b_i s_{ij}$. The second observation is that the belief strength of a node may be computed from the beliefs and certainties of its endorsements. Nodes which do not have endorsements, and in fact any node in the system, can be provided with an *Intuition* represented as $I_i = (b'_i, c'_i)$. This structure appears much the same as the Rational structure except that its values are never computed, but they remain available to take part in the computation of other beliefs. Intuitive values correspond to the usual direct assignment of belief and certainty to a proposition from which other beliefs and certainties are to be determined.

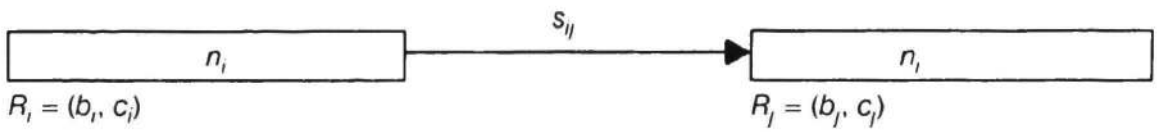


Figure 1. : Nodes of a network representing cognitive units with belief, certainty, and endorsement structure.

The computation of a belief value for a node is largely dependent on the manner in which its endorsements interact. For example, the final belief in a node is a function of the summary of the evidence and arguments. In such instances, belief depends on: (1) measuring the varying contributions of the individual endorsements and (2) measuring the effects of interaction among the different endorsements. This interaction among a set of endorsements $\{n_1 \ldots n_m\}$ for n_j depends on the relative importance of each endorsement defined as

$$r_{ij} = \frac{s_{ij}}{\sum_{k=1}^{m} |s_{kj}|} \qquad [1]$$

and the relative certainty defined as

$$rc_i = \frac{c_i}{c^*} \qquad [2]$$

where c^* is max $\{c_i \ldots c_m\}$. In this manner we can define a measure of belief using a formula such as

$$b_j = \sum_{k=1}^{m} rc_k \times r_{kj} \times b_k \qquad [3]$$

The endorsements with the greatest relative importance and greatest relative certainty have the most impact on the final belief. The interaction of endorsements is analogous to a tug-of-war where the different endorsements tug and pull against one another until an equilibrium is reached.

Since the support of an endorsement can be endorsed we can also deal with situations in which only one of several endorsements is adequate to allow a cognitive unit to be believed. For example, the statement I CAN TAKE A GRADUATE COURSE IN COMPUTING is endorsed by I AM A COMPUTING GRADUATE, OR I AM AN ELECTRICAL ENGINEERING GRADUATE, OR I AM A COMPUTING UNDERGRADUATE WITH GOOD MARKS The three endorsements are mutually exclusive; only one of them need be true. In this example a cognitive unit n_j endorsed by $\{n_1 \ldots n_m\}$ will be given an endorsement of $b^* s_{kj}$ where $b^* = \max \{b_1 \ldots b_m\}$. The node n_k will inhibit all the other endorsements by giving an inhibitory endorsement to their respective supports for the endorsement. A more complete description is available in Craddock (1986).

The certainty of a belief is calculated as a function of the agreement of the individual endorsement strengths with the final belief value calculated from them. Thus, belief must be calculated before certainty. The importance of the agreement is once again measured as a function of the relative support of the individual endorsements and their relative certainties. As these values increase so does the uncertainty associated with disagreement. Where $\{n_1 \ldots n_m\}$ are the endorsing nodes for n_j, this effect can be modelled in formulas such as:

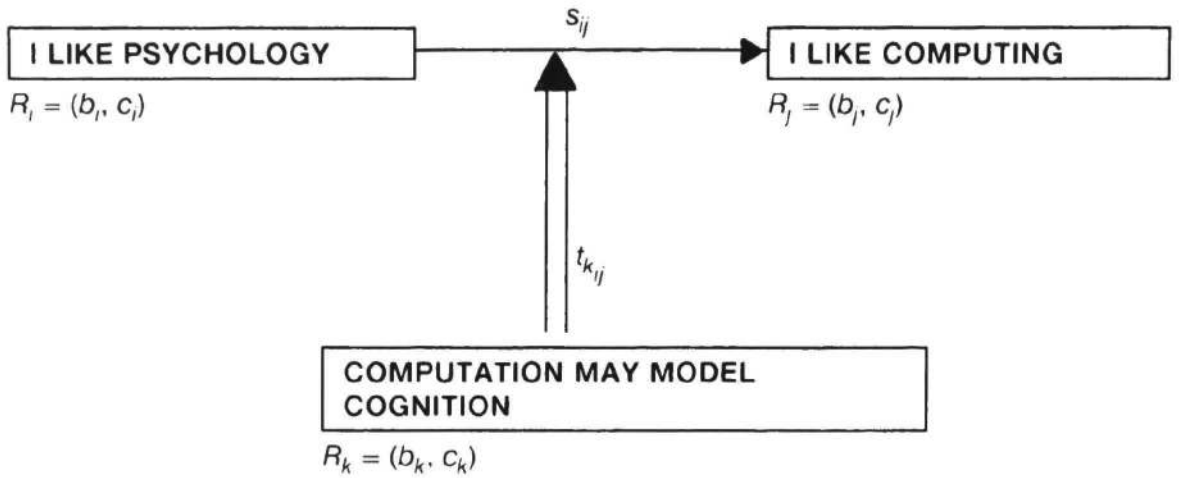


Figure 2. : An example of an endorsement which influences the strength of endorsement between two other nodes.

$$c_j = 1 - \left[\sum_{k \neq l,\, k=0}^{l} |(b_k - b_l)| \times r_{kl} \times rc_k \right] \qquad [4]$$

4. CONTRADICTIONS

Rational contradictions among endorsements are defined as follows: If A is compelling evidence against n_i but B is equally compelling evidence for n_i then the endorsements for n_i are inconsistent. In addition to a rational contraction an *intuitive contradiction* can also be defined: If the intuitive belief, b'_i is not equal to the rational belief, b_i then the two beliefs are inconsistent. If we define a threshold of intuitive contradiction, T_I then (1) if $|b'_i - b_i| > T_I$ then assign b_i to I_i and recalculate the rational beliefs of all the nodes endorsed by n_i such that $s_{ij} \neq 0$ or (2) if $|I_i - R_i| < T_I$ then assume a state of equilibrium has been reached and do not recalculate the rational beliefs of any of the nodes endorsed by n_i.

Intuitive contradictions are useful for recognizing changes in belief through a knowledge base when endorsements are added and removed. In addition they can be used to control cycles which may force more global interpretations on input propositions. When cycles exist within a network

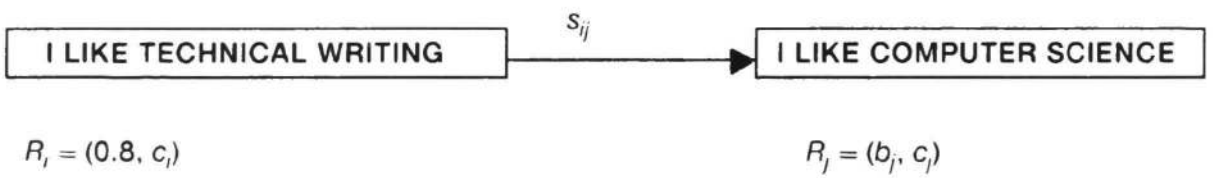


Figure 3. : An example of an endorsement which may have net positive or net negative support.

$N = <P, S>$, belief and certainty values will only be calculated for nodes in a partial network $N' = P', S'$, where $P' \subseteq P$, and $S' \subseteq S \cap (P X P')$, where there exists a node $n_i \in P - P'$ such that there is an elementary path between n_i to P' and $|I_i - R_i| > T_i$.

5. CONCLUSIONS

The model discussed in this paper seeks to develop representational and inference mechanisms capable of dealing with incomplete, inaccurate, and uncertain information. To this end, a connectionist model is proposed and heuristics are developed to collect and evaluate the endorsements for propositions in the network of beliefs. At the same time it is intended that the model represent at least some of the processes used in human reasoning.

As the model is intended to represent belief maintenance with uncertainty it differs from existing connectionist models (Anderson 1982; Rumelhart and McClelland 1982; McClelland and Rumelhart 1983) in several important respects. First, the uncertainty of a proposition is represented numerically, as the values of R_i, and non-numerically, as the structure of endorsements. Second, once the endorsements have been collected they are subject to reasoning and natural heuristics to compute numeric values as depicted in formulae [1] to [4].

Kahneman and Tversky (1982a, b) have shown that their heuristics; availability, representativeness, and adjustment and anchoring, can help describe human decision making under conditions of uncertainty. Once these heuristics are recognized as part of human reasoning it no longer appears illogical in the sense of being erratic, but rather more pragmatic and difficult to specify in terms of the logic inference mechanisms of traditional logic. Kahneman and Tversky (ibid.) provide numerous examples in which subjects reach decisions which run counter to those reached by mathematical theories. While the heuristics proposed in this paper are by no means as exhaustive nor the formulae necessarily optimal, they do illustrate how heuristics might be incorporated into the decision making model in a straight forward fashion (Craddock 1986).

In contrast, most connectionist models ignore, or do not explicitly deal with the non-numeric representation of uncertainty, depending on numeric values alone which provide no evidence as to how they were calculated, what they actually represent, or how reliable they are. Of major issue is the belief that numerical values, blindly tallied, are an inadequate representation of reasoning. Symbolic structures of support are necessary to specify how and why numeric beliefs are calculated. The availability of an endorsement structure allows the model to not only provide numerical information but also a description of its own reasoning process. The advantages of having a model whose reasoning can be readily understood are numerous and imperative if the justifications for a decision are to be made clear.

A further difference is that the proposed model can represent varying degrees of interaction between sources of evidence while models such as MYCIN (Shortcliffe, 1975) must make the assumption that all evidence is conditionally independent and that hypotheses are mutually exclusive. For example, the dynamic strengths of endorsement (see figure 2) allow us to represent evidence which is disjunctive, that is, strong belief may be propagated on the basis of only one of many supports.

ACKNOWLEDGMENT

This research was carried out with the support of Natural Sciences and Engineering Research Council of Canada grant number A2427.

REFERENCES

Anderson, R., **The architecture of cognition.** Harvard University Press, 1982.

Cohen, R., "The use of heuristic knowledge in decision theory," Diss. Stanford University, 1983.

Craddock, A. J., "Modelling uncertainty in a knowledge base," MSc. Thesis, Queeen's University at Kingston, 1986.

Edwards, W., "Conservatism in human information processing," in **Judgment under uncertainty: Heuristics and Biases**, Cambridge University Press, 1982.

Hamburger, H., **Combining uncertain estimates**. Uncertainty and Probability in Artificial Intelligence. August 14-16, 1985. UCLA: Los Angeles, California, 1985.

Kahneman D., and Tversky A., "The psychology of preferences." **Scientific American**, 246: 1982a, pp. 160 173.

Kahneman D., and Tversky A., **Judgment under uncertainty: Heuristics and Biases**. Cambridge University Press, 1982b.

McCarthy, J., "Circumscription A form of non-monotonic reasoning." **Artificial Intelligence**, 13: 1980, pp. 27 39.

McClelland, J. L., and Rumelhart, D. E., "An interactive activation model of context effects in letter perception: Part 1. An account of basic findings," **Psychological Review**, 88: 1981, 375-407.

McDermott, D., and Doyle, J., "Non-monotonic Logic I," **Artificial Intelligence**, 13: 1980, pp. 41 72.

Reiter R., "A Logic for Default Reasoning," U.B.C Technical Report, 1978a.

Reiter, R., "A logic for default reasoning," **Artificial Intelligence**, 13: 1980, pp. 81 132.

Rumelhart, D. E., and McClelland, J. L., "An interactive activation model of context effects in letter perception: Part 2. The contextual enhancement effect and some tests and extensions of the model," **Psychological Review**, 89: 1982, 60-94.

Rumelhart, D. E. and Zipser D., "Feature Discovery by Competitive Learning," **Cognitive Science**, 9: 1985, pp. 75 112.

Shortcliffe, **Computer-based medical consultations: MYCIN**. New York: American Elsevier, 1975.

Zadeh L.A., "Commonsense knowledge representation based on Fuzzy Logic," **Computer**, 16: 10, 1983, pp.61-66.

SOME FACTORS INFLUENCING THE COMPREHENSION OF PRONOUNS IN TEXTS

Rosalind A. Crawley

Department of Psychology
University of Durham
England

The interpretation of pronouns depends on a number of factors ranging from the purely linguistic through to general knowledge of the world. Thus, the study of pronoun comprehension can provide important insights into the more general question of comprehension during reading. Some of the factors influencing pronoun comprehension operate at the level of the single sentence (the local level) and others operate at the level of the text as a whole (the global level). Although there has been quite a lot of work on the factors affecting pronoun comprehension at these two levels separately (e.g. Ehrlich, 1980; Springston, 1975), they are rarely considered together. Thus, a major aim of this study was to examine the relationship between the factors affecting pronoun comprehension at the local and global levels.

Two local factors and one global factor were examined. The two local factors were the linguistic constraint of lexical agreement (specifically, gender agreement) and the heuristic strategy of subject assignment. The global factor examined was topicalisation at the text level. There is evidence to suggest that all three of these factors might be important for resolving pronominal reference. For example, the presence of a gender cue facilitates the comprehension of pronouns in isolated sentences (e.g. Ehrlich, 1980). Similarly, the use of a subject assignment strategy is suggested by the observation that in written texts the subject of the sentence is frequently the antecedent for a pronoun (e.g. Grober, Beardsley & Caramazza, 1978; Hobbs, 1976). In addition, at the global level several people have suggested that the global topic, or main character in a text is a likely antecedent for a pronoun (e.g. Clancy, 1980; Sanford & Garrod, 1981).

It seems likely that surface features of a text may signal a character as the global topic. These features include the title of a text (Kieras, 1979; Kozminsky, 1977), initial mention in a text (Kieras, 1978; Sanford & Garrod, 1981) and frequency of mention (Perfetti & Goldman, 1974). Thus, a second aim of this study was to determine whether variation in the number of features signalling the topic causes any variation in its influence on pronoun comprehension.

In the three experiments reported here, Subjects were asked to read short passages of text. Each passage contained one target sentence in which there were two pronouns and two potential antecedents. In some of the target sentences, the pronouns' antecedents could not be determined by gender cues (ambiguous pronouns), for example, "Shaun led Ben along the path and he

called to him to be careful". In others, the pronouns' antecedents could be uniquely identified using gender cues (unambiguous pronouns), for example, "Clare led Ben along the path and she called to him to be careful". This is the gender cue manipulation. In addition, since one of the potential antecedents occurred in subject position, the influence of the subject assignment strategy could be determined. Finally, one of the potential antecedents was the topic of the passage so the effect of topicality could also be manipulated.

EXPERIMENT 1

Twelve experimental passages were presented one sentence at a time on the screen of a microcomputer. Subjects were asked to press a key as soon as they had understood each sentence. The key press caused the next sentence to appear, following on from the previous one as in normal text. Once a sentence had appeared on the screen, it remained there until the Subject had read the whole passage.

Each passage mentioned two characters; the topic and the nontopic. The topic was signalled by using the topic's name as the title of the passage, by mentioning the topic first in the passage and by mentioning the topic more frequently than any other character in the passage. The passages were six sentences long and the target sentence appeared as the fifth sentence. There were six versions of each target sentence; two containing ambiguous pronouns and four containing unambiguous pronouns. Figure 1 shows an example of the materials.

In the first clause of each target sentence, the topic and nontopic were mentioned by name and in the second, they were referred to again using pronouns. There were two versions of the ambiguous target sentences: Either the topic or the nontopic appeared as the subject of the sentence (see sentences 1 and 2, Figure 1). The second, pronominal clause was the same in the two conditions. A question at the end of each passage enabled the assignment of these pronouns to be determined. Any preference for assigning these linguistically ambiguous pronouns to the subject of the sentence or the topic of the passage would be revealed in these assignments.

The four unambiguous conditions can best be described in terms of who the subject pronoun referred to. It referred to the topic in subject position (sentence 3), the topic in object position (sentence 4), the nontopic in subject position (sentence 5) and the nontopic in object position (sentence 6). Reading rates for these unambiguous conditions were examined. If gender cues alone are used for understanding pronouns, there should be no difference in reading rates for the four conditions. Any preference for assigning the subject pronoun to the subject of the sentence or the topic of the passage should be accompanied by faster reading rates in the conditions where the linguistically constrained assignment is in accordance with these preferences.

Topic = Shaun / Clare Nontopic = Ben

a. First four sentences

SHAUN

Shaun started to get worried as it became darker and the mist grew thicker. He was the leader of this walking expedition in the Lake District and he felt responsible for the others following him. He hadn't realised it would take them so long to walk back. They came to a place where the path narrowed over a steep drop and Shaun decided to go ahead with his friend Ben to make sure it was safe before the others followed.

(Clare was substituted for Shaun in the unambiguous passages.)

b. Six versions of the target sentence

Ambiguous

1.	T=S	Shaun led Ben along the path and he called to him to be careful.
2.	NT=S	Ben led Shaun along the path and he called to him to be careful.

Unambiguous

	Pronoun refers to:	
3.	T,S	Clare led Ben along the path and she called to him to be careful.
4.	T,O	Ben led Clare along the path and she called to him to be careful.
5.	NT,S	Ben led Clare along the path and he called to her to be careful.
6.	NT,O	Clare led Ben along the path and he called to her to be careful.

c. Final sentence

They got safely over to the proper path and shouted to the others that it was all right and eventually they all made their way down to their minibus at the bottom.

FIGURE 1 EXAMPLE PASSAGE FROM EXPERIMENT 1
(T = Topic, NT = Nontopic, S = Subject, O = Object)

The ambiguous conditions will be considered first. Table 1 shows the number of subject and object assignments in these two conditions. There were far more assignments to the subject than to the object (Min F' = 52.21, df = 1, 24, $p < .01$). Thus, there was a strong preference for making assignments in accordance with the subject assignment strategy. There was also an interaction between condition and assignment to the subject or object ($F1 = 8.31$, df = 1, 119, $p < .01$; $F2 = 4.17$, df = 1, 11, $p = .06$). This indicates a stronger tendency to assign the subject pronoun to the subject of the sentence when it was also topic of the passage. Thus, the topic of the text also influenced

TABLE 1 ASSIGNMENTS - EXPERIMENT 1, AMBIGUOUS PASSAGES

	Assignment to SUBJECT	OBJECT
T = S	194	46
NT = S	167	73
Means	181	60

the assignment of the linguistically ambiguous pronouns.

Reading rates for the two conditions also showed an effect of topic (see Table 2). Sentences in which the topic was subject were read faster than sentences in which the nontopic was subject (Min F' = 4.86, df = 1, 19, p <.05).

The unambiguous conditions will now be considered. Table 3 shows the reading rates for the unambiguous conditions. (Here, the discussion will be confined to the assignment of the subject pronoun in each target sentence.) The clearest effect in these unambiguous conditions is an effect of topic. Reading rates were faster when the pronoun referred to the topic rather than the nontopic (Min F' = 5.39, df = 1, 31, p <.05). There was also some suggestion of an effect of the subject. Sentences in which the pronoun referred to the subject were read faster than sentences in which the pronoun referred to the object. However, this effect only held up in the F1 analysis (F1 = 4.23, df = 1, 119, p <.05; F2 = 1.46, df = 1, 11, p =.25). There was no interaction.

The results of Experiment 1 show that both local and global factors act together to influence pronoun comprehension. While the local subject of the sentence was clearly preferred as the antecedent for linguistically ambiguous pronouns, this effect was modified by an influence of the global topic. The global topic also influenced the ease of reading the ambiguous target sentences. Similarly, even when there was a clear gender cue present in the unambiguous conditions, both the local subject and the global topic influenced the ease of assignment, although in these sentences, the effect of the topic was stronger than the effect of the local subject.

TABLE 2 MEAN READING RATES (WORDS PER SECOND) - EXPERIMENT 1, AMBIGUOUS PASSAGES

T = S	NT = S
4.25	3.65

TABLE 3 MEAN READING RATES (WORDS PER SECOND) - EXPERIMENT 1, UNAMBIGUOUS PASSAGES

Pronoun referent	TOPIC	NONTOPIC	Means
SUBJECT	4.37	4.05	4.21
OBJECT	4.14	3.87	4.01
Means	4.26	3.96	

However, it is not clear exactly which features of the topic produced these effects (title, initial mention or frequency of mention). In the next two experiments, the number of features signalling the topic was reduced to two (title and initial mention) to see whether or not frequency of mention was critical for the topic's influence on pronoun comprehension. In the next two experiments, ambiguous and unambiguous pronouns were studied separately; ambiguous pronouns in Experiment 2 and unambiguous pronouns in Experiment 3.

EXPERIMENT 2

The twelve experimental passages used in this experiment were similar to the ambiguous passages used in Experiment 1 except that the frequency with which the topic and nontopic were mentioned was equalised. Thus, the topic was no longer distinguished from the nontopic in terms of how often it was mentioned. There were also a few other minor changes to the materials. For example, most of the target sentences were shortened so that any information superfluous to the assignment of the pronouns was discarded, and the recency with which the topic and nontopic were mentioned before the target sentence was counterbalanced. In every other way, the materials were the same as the

TABLE 4 ASSIGNMENTS - EXPERIMENT 2

	Assignment to SUBJECT	OBJECT
T = S	57	14
NT = S	55	17
Means	56	16

TABLE 5 MEAN READING RATES (WORDS PER SECOND) - EXPERIMENT 2

T = S	NT = S
4.03	3.63

ambiguous passages used in Experiment 1. As before, there were two versions of each target sentence: The topic or nontopic was subject of the sentence. The task was the same as in Experiment 1.

Table 4 shows the number of assignments to the subject and object in each condition. There were more assignments to the subject than to the object (Min F' = 32.05, df = 1, 17, p <.01). But there was no difference in the pattern of assignments in the two conditions and no interaction. Thus, while there was still a strong effect of the local subject, there was no effect of topic on assignments in this experiment. However, there was an effect of topic on reading rates. Table 5 shows the mean reading rates in each condition.

Sentences in which the topic was subject were read faster than those in which the nontopic was subject (F1 = 6.78, df = 1, 23, p <.05; F2 = 3.06, df = 1, 11, p >.1). (This effect was also significant by passages (F2) when only sentences in which subject assignments had been made were considered (Min F' = 4.24, df = 1, 29, p <.05) - see Crawley, 1985.)

EXPERIMENT 3

The passages used in this experiment were identical to those used in Experiment 2 except that the topic and nontopic were different genders so that the pronouns in the target sentences could be disambiguated by gender. As in Experiment 1, there were four versions of the unambiguous target sentences (see Figure 1).

The mean reading rates in each condition are shown in Table 6.

TABLE 6 MEAN READING RATES (WORDS PER SECOND) - EXPERIMENT 3

Pronoun referent	TOPIC	NONTOPIC	Means
SUBJECT	3.86	3.63	3.75
OBJECT	3.54	3.69	3.62
Means	3.70	3.66	

Analysis of these reading rates showed no reliable differences. An examination of the effect of recency of mention of the topic and the nontopic revealed no significant effects either here or in Experiment 2.

Thus, when unambiguous pronouns were presented in the modified passages of Experiment 3, neither the local subject nor the global topic affected the ease of pronoun comprehension. It seems that readers relied on gender cues alone in this experiment.

Overall, the results of these three experiments suggest that both local and global factors act together during pronoun comprehension. At the local level, both gender cues and the local subject affected pronoun comprehension. (Strictly speaking, the subject assignment strategy could in fact be a parallel function strategy; Sheldon, 1974. These two strategies are not distinguished here.) The subject appeared to have a stronger influence on the comprehension of ambiguous pronouns than on the comprehension of unambiguous pronouns. In the absence of linguistic constraints, the subject was the preferred antecedent in both Experiments 1 and 2. When there were linguistic cues available, however, the subject had no influence except for a weak effect in the unambiguous passages of Experiment 1. Thus, heuristics like those involving the subject may only operate in the absence of other strong cues to assignment.

The global topic also influenced pronoun comprehension (Experiments 1 and 2). The topic's influence appeared to depend on the number of factors signalling it. In Experiment 1, where the topic was signalled by the title, initial mention and frequency of mention, it had an effect on both the assignment of ambiguous pronouns and the ease of reading both the ambiguous and unambiguous sentences. In Experiment 2, however, where frequency no longer distinguished the topic from the nontopic, the topic had no effect on the assignment of ambiguous pronouns, although it still had an effect on the ease of reading the target sentences. And in Experiment 3, where frequency did not signal the topic and where gender cues determined assignment unambiguously, there was no effect of topic. This contrasts with the results from the unambiguous sentences of Experiment 1. In the presence of gender cues, the topic clearly has to be very salient before it influences the ease of pronoun comprehension. A tentative explanation for these results is that the topic effect is graded and that topicality is a continuum rather than an all-or-none feature. A more systematic investigation of this proposition is currently in progress.

ACKNOWLEDGEMENTS

This research was supported by a grant from the Medical Research Council, England.

REFERENCES

Clancy, P.M. (1980) Referential choice in English and Japanese narrative discourse. in W.L. Chafe (Ed.), The Pear Stories, Vol. III: Advances in Discourse Processes. Norwood, New Jersey: Ablex Publishing Company.

Crawley, R.A. (1985) 'The effects of local and global factors on the comprehension of pronouns.' Unpublished PhD thesis, Durham University.

Ehrlich, K. (1980) Comprehension of pronouns. Quarterly Journal of Experimental Psychology, 32, 247-255.

Grober, E.H., Beardsley, W. & Caramazza, A. (1978) Parallel function strategy in pronoun assignment. Cognition, 6, 117-133.

Hobbs, J.R. (1976) Pronoun resolution. Research report 76-1, Department of Computer Sciences, City College, City University of New York.

Kieras, D.E. (1978) Good and bad structure in simple paragraphs: Effects on apparent theme, reading time and recall. Journal of Verbal Learning and Verbal Behaviour, 17, 13-28.

Kieras, D.E. (1979) The relation of topics and themes in naturally occuring technical paragraphs. Technical report number 1, University of Arizona.

Kozminsky, E. (1977) Altering comprehension: The effect of biasing titles on text comprehension. Memory and Cognition, 5, 482-490.

Perfetti, C.A. & Goldman, S.R. (1974) Thematization and sentence retrieval. Journal of Verbal Learning and Verbal Behaviour, 13, 70-79.

Sanford, A.J. & Garrod, S.C. (1981) Understanding Written Language. Chichester: J. Wiley and Sons Ltd.

Sheldon, A. (1974) The role of parallel function in the acquisition of relative clauses in English. Journal of Verbal Learning and Verbal Behaviour, 13, 272-281.

Springston, F.J. (1975) 'Some cognitive aspects of presupposed coreferential anaphora.' Unpublished PhD thesis, Stanford University.

A PSYCHOLOGICAL INVESTIGATION INTO THE DEICTIC CENTER

Joyce H. Daniels
Department of Psychology
SUNY Buffalo, Buffalo, N.Y. 14226

The Deictic Center.

When we read a narrative much of what we understand is not explicitly stated in the text. Not only do we usually not have a detailed description of the environment within the narrative, but the information that we need to understand movement through narrative space and time is not always explicitly provided for us. Nevertheless, comprehenders of narrative seem effortlessly to understand who characters are, where they are at any given moment in narrative time, and proper sequences of time intervals for events that take place. We know when a character leaves the current scene, and we know how much time has elapsed, accurately enough to make proper temporal judgments about the flow of the narrative. Since we also have available our general knowledge of how events are usually structured, space is generally laid out, and time progresses, a combination of linguistic and non-linguistic information can be used to construct a mental model of the scenes and events in a narrative. This model and the input from the narrative enable us to construct our comprehension of the narrative.

Members of the Graduate Group in Cognitive Science at SUNY Buffalo have been examining the methods by which we understand the flow of time and space in narrative text. We hypothesize that information about time, space, and the focal-character form a single data structure. They are part of what we call the deictic center, referred to from now on as the DC. The DC is composed of a WHO-point, a WHERE-point, and a WHEN-point, corresponding to the character, space, and time elements of the current place in the narrative. The concepts of deictic time, deictic place and person deixis are discussed by Fillmore (1975). We propose that the reader's mental model of the current narrative contains the DC. The DC tracks the movement of the narrative in time, space, and focal-character, cued by particular linguistic devices in the text.

Current investigations of the DC are being carried out by members of our group. A computer model of temporal movement was reported in Almeida and Shapiro, (1983) and is being continued in Almeida, (in progress). Rapaport and Shapiro, (1984) and Rapaport and Wiebe, (forthcoming) examine the concept of the WHO-point in relation to the DC. Bruder, Engl, and Schultz, (1985), have reported sentence

reading time research on the psychological validity of the role of the DC. The research reported here concerns the psychological validity of the concept of the deictic center.

Lexical Controllers of the Deictic Center.

We have identified some of the specific lexical items that act as controllers of deictic movement. The deictic verb 'go' or the presence of preposed adverbials, such as 'outside the house' at the beginning of a sentence, tend to indicate a shift from one DC to another. The nature of the linguistic devices demands the movement of the comprehender spatially, (i.e. the DC shifts to a new location) so that the new WHERE-point must actually be mentioned. In contrast, the deictic verb 'come' seems to indicate that the current WHERE-point is not moving and the reader should assume that the following textual information is coming into the current deictic center. The absence of a preposed adverbial, or lack of information about a new WHERE-point indicates that the reader should assume that the current DC is being maintained.

Thus, the comprehender learns to expect that shifts of the scene of action or event, the WHERE-point, will be cued by certain markers ('go' or a preposed adverbial) along with a naming or brief description of the new DC, and the comprehender expects no shift of the DC due to their absence (e.g., lack of a preposed adverbial) or due to the use of a maintenance marker (e.g.,'come'). If the reader finds a movement marker, the new DC will be established at that point, with the new information describing it. Minimal description is necessary if the reader is moving within the model already built, but a more detailed description is necessary if the new location has not been previously mentioned. Based on this information, the old DC will be at some greater mental distance, from where the reader currently is in the narrative. If the reader does not find a movement marker, or finds one that indicates maintenance, then he will remain at the current DC. If a shift in DC occurs, but there is no movement marker for the reader indicating the shift, then the reader should become confused by a description of a WHERE-point inconsistent with her current position established with her mental model. In addition, the reader should assume, in the absence of this marker, that he is in the same Deictic Center if no new description is provided.

Preliminary Experimental Evidence for the Deictic Center.

A preliminary examination of this hypothesis, an untimed reading comprehension experiment with naturally occurring text was conducted and reported by Segal, Bruder,

and Daniels, (1984). The results were not significant in all cases, but, generally, the direction of the results were in line with these predictions: (1) sentences with a maintenance of the DC should be answered with greater certainty of validity, and answered more uniformly among subjects. (2) if a movement marker should have been provided and was not, then subjects should have more difficulty answering statements about that event than if they had been provided with the original linguistic input.

The Experiment.

The research currently being reported is a reaction time study. The stimuli are artificially constructed stimuli and reflect situations in the real world, acknowledging that general information has an influence on responses. General information should be used in construction of the mental model by the reader, not in construction of the response.

This experiment specifically examines the effect of the deictic verbs 'come' and 'go', and of preposed adverbials on conceptual movement of the DC in narrative text. It is assumed that the verb 'come' will maintain the DC where it is currently and that 'go' will move the center to a new location. If a subject must respond to a statement about a DC that s/he has moved away from, s/he should take longer to do so, since s/he must leave where s/he currently is in memory and go back to a previous DC to respond. We assume that if a preposed adverbial is present in the text it will move the subject to a new DC causing the subject to take longer to respond than if the adverbial is absent. This would be due to the subject returning mentally to a previous DC to find the relevant information to determine the answer to the statement. However, the absence of a preposed adverbial may cause confusion if the comprehender cannot determine where s/he is in the mental model, because an important cue is missing. This will leave the comprehender in the previous DC, trying to incorporate current linguistic input into an incomplete model.

METHOD. Subjects. Subjects were 48 SUNY Buffalo undergraduate students, who participated to fulfill a requirement for their introductory psychology course.

Materials. Each subject was presented with fifty narratives of four sentences each. Ten narratives investigated the role of preposed adverbials; five compared sentences with 'come' versus 'go'; and five compared a sentence with 'go' versus a non-movement marker such as 'remain', or 'stay', or the verb 'to be'.

In each case, the narrative was used as its own comparison set. For example, the starred sentences were the

only difference in each presentation set. Each set had only one of the starred sentences.

Sample experimental set:

John and Mary were eating dinner when there was a knock at the door.
*John got up and went to answer the door.
*John looked up to see his partner come in.
Kevin greeted John with a bottle of champagne and a big hug.
They had just won a large advertising account.
****
Mary is in the dining room.

The target, or statement requiring a true-false response, was the same in all versions. It always referred back to the original DC. Subjects also saw twenty-eight fillers, each also composed of four sentences. Fillers generally described a situation, or scene, without any movement. Fillers were the same for all versions. The narrative presentation order was randomized. After the first four sentences had been presented, the subject saw a row of asterisks flash on the screen to signal that s/he should respond 'TRUE' or 'FALSE' to the statement that would immediately follow, based on the information provided in the preceding sentences.

Procedure: Narratives were presented on an Apple IIe computer with a Thunder Clock Card, controlling the timing to the nearest millisecond. Subjects were given printed instructions to read. They were then given oral directions, with general printed directions repeated again on the computer screen. This was followed by ten practice narratives. After the practice session, general directions were again displayed. Subjects pressed the space bar to have the present sentence disappear, and the next sentence or row of asterisks appear. The true/false test statement was presented automatically after the row of asterisks disappeared. Subjects pressed the space bar if the statement was true, and any other key if it was false.

Results.

	MOVEMENT		NON-MOVEMENT	
GO	3.17 sec	COME	2.64 sec.	
PREPOSITION	2.77 sec	NO PREP	2.49 sec	

Mean reading/reaction times to the same final statement presented in compared test sets.

The results are significant for both the come/go and the preposition/no preposition distinctions using the paired sample t-test. The come/go distinction shows significance with $t(8) = 4.74$ $p < .01$. The preposition/no preposition distinction shows significance with $t(11) = 3.11$ $p < .05$. There is also a main effect of movement versus non-movement $t(19) = 4.07$ $p < .01$. Two narrative sets examining the come/go distinction were deleted from analysis due to excessive errors in subject responses to the final statement. Due to item by item examination, it was determined that these sets should be eliminated since the overall error rate, excluding these two sets was less than 1%.

Discussion. Movement through narrative in terms of space and time are dependent not only upon general world knowledge about how events and space are laid out in the real world and specific types of linguistic markers, but also the previous DC. The WHERE-, WHEN- and WHO-points that are involved at any place in a narrative text are all currently activated in a focusing mechanism, the DC. When the DC shifts in time or space to a new WHO-, WHEN-, or WHERE-point, we, as comprehenders, update our current knowledge accordingly. The psychological validity of the DC is evident from the results investigating the movement of the WHERE-point in this experiment. If a reader remains in a DC and new information does not update this DC to a new DC, then when asked about DC1 the subject is "there" at DC1 in his/her mental model and s/he can respond quickly. If the DC has shifted to DC2 and the subject is asked about DC1, then s/he takes longer to respond, because s/he must leave the current center, DC2, and interrogate information about DC1 that is now at a mental distance.

Future work.

The next issues to be dealt with concern the determination of other specific linguistic markers that update the DC's WHERE-, WHEN-, and WHO-points. A continuing investigation of the placement of the adverbial is also warranted. Bruder et. al., 1985, has investigated sentence reading times and found a significant effect for post-posed adverbials on the sentence following the one with the adverbial, but no effect on the reading times for the sentence containing the adverbial. This suggests that the current experimental paradigm might be extended to include an examination of the effect of post-posed adverbials. We also might ask if the spatial, temporal and focal-character focus of the DC is only one component of a larger concept in which all activated parts of a narrative are dependent upon what is currently active in the DC.

Acknowledgments.

Many thanks go to William Rapaport and Gail Bruder for their generous help and comments on prior versions of this paper. This experiment was possible due to the patience and programming skill, time and logical insights of William Daniels. A special thanks goes to the Graduate Group in Cognitive Science and my advisor, Erwin Segal, for providing the opportunity to explore the deictic center concept and for their insights about deictic center operation.

BIBLIOGRAPHY

Almeida, M. J. [dissertation in progress]. REASONING ABOUT THE TEMPORAL STRUCTURE OF NARRATIVES. Computer Science Department, SUNY Buffalo.

Almeida, M. and Shapiro, S. [1983]. Reasoning about the Temporal Structure of Narrative Texts. PROCEEDINGS OF THE FIFTH ANNUAL CONFERENCE OF THE COGNITIVE SCIENCE SOCIETY, Rochester, N. Y.

Bruder, G., Engl, L., and Schultz, J. [1985]. PREPOSED ADVERBIALS SIGNAL CHANGE IN THE NARRATIVE DEICTIC CENTER. Paper presented at the Psychonomic Society Meeting, Boston.

Fillmore, C. [1975]. SANTA CRUZ LECTURES ON DEIXIS. Bloomington, Indiana: Indiana University Linguistics Club Press.

Rapaport W. and Shapiro, S. [1984]. Quasi-Indexical Reference in Propositional Semantic Networks. PROCEEDINGS 10th INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS (COLING-84).

Segal, E., Bruder, G. and Daniels, J. [1984]. DEICTIC CENTERS IN NARRATIVE COMPREHENSION. Paper presented at Psychonomic Society Meeting, San Antonio.

Wiebe, J. M. and Rapaport, W. J. [forthcoming]. Understanding de re and de dicto Belief Reports in Discourse and Narrative. IEEE Special Issue on Knowledge Representation. Geneva, Switzerland.

Towards A Comparative Psychology of Cognitive Content: Exploring Tree Preference Asymmetries In Humans, Pigeons, and Monkeys

Michael R. Dawson, Dwight S. Mazmanian, and William A. Roberts
Department of Psychology
University of Western Ontario
London, Ontario, Canada

Abstract

Conceptual structure in humans, pigeons, and monkeys was investigated using a multidimensional scaling procedure. Pigeons and monkeys were initially trained to discriminate between stimultaneously presented tree and nontree pictorial stimuli. Preference data was collected by inserting probe trials in which the animals were forced to choose between two tree stimuli. Analogous data for human subjects was collected by having subjects rate their preferences for the same stimuli. Tree preference relationships in the different datasets were obtained using the DEDICOM procedure. These analyses revealed striking interspecies differences in conceptual structure. The analysis of human tree preferences revealed a 'whole vs. part' pattern in which stimulus preference was a function of stimulus completeness. Pigeon tree preferences were qualitatively different from human tree preferences, and also appeared to be less elaborate. In general, 'branchy' stimuli were preferred over 'leafy' stimuli, and a 'whole vs. part' pattern did not emerge. The data for monkeys also illustrated a preference for 'branchy' structures over 'leafy' structures. Individual differences between monkey preferences were also revealed, and were found to be related to performance on the initial discrimination task. Monkeys that had a well-defined tree preference pattern learned this task faster than did monkeys with a less defined structure. The results of all of the analyses demonstrated interspecies differences in tree concepts, and

suggested the possibility that these may be related to different functional experiences or requirements.

A key hypothesis in cognitive science is that intelligence is the product of symbolic processing. Under this hypothesis, intelligent behaviour may best be explained by providing an account of the representational system that mediates the behaviour (e.g., Pylyshyn, 1984). This "Cognitivist" hypothesis is also beginning to be applied in the study of animal behaviour (e.g., Griffin, 1978; Roitblat, 1982). Indeed, many animal behaviour researchers now assume that nonhuman species use some form of mental representation, and many recent experiments have been concerned with determining the functional characteristics of animal representational systems (e.g., Mazmanian & Roberts, 1982; Premack, 1983; Roberts, Mazmanian, & Kraemer, 1986; Roitblat, Bever, & Terrace, 1984; Sands, Lincoln, & Wright, 1982).

Roitblat (1982) makes a distinction between the *domain* and the *content* of a representation. The domain is the class of situations in the world to which the representation applies. The content is the set of features about the represented world that can be derived from the representation. It is quite likely that different species use different contents in representing identical domains. For example, one species of animal may encode a tree as a place of shelter, while another may encode the same tree as a source of food. Thus, a major issue in a comparative psychology of conceptual structure is whether the contents of representations of various species are different, and if so, in what way.

Some recent studies have shown that multidimensional scaling (MDS) procedures may be very useful in examining the contents of animal concepts (Blough, 1985; Sands, Lincoln, & Wright, 1982). The results we report come from an experiment that extends this previous methodology in two ways. First, a paradigm was used in which the preference of one stimulus over another in terms of its being "tree-like" was measured,

using a set of photographs of trees in natural settings. This preference data was obtained for human, pigeon, and monkey subjects. Thus we were able to compare tree-preference relationships among three different animal species for the same set of stimuli. Second, preference data is intrinsically asymmetric (whan A is preferred over B, B is *not* preferred over A), and therefore traditional MDS analysis procedures are not appropriate. Instead, we used the DEDICOM MDS procedure, which represents the asymmetric structure of a matrix in terms of directional geometric patterns that can easily be interpreted (e.g., Harshman, Green, Wind & Lundy, 1982). When used to analyse preference data, DEDICOM indicates which stimuli were preferred over others, as well as the extent of this preference. In interpreting such a pattern, an examination of stimulus characteristics is undertaken in an attempt to understand why particular preference relationships hold. Our aim was to examine the patterns of tree preferences obtained from the three species in order to see if there were any systematic differences in the content of the concept "tree".

Data Collection

Sixteen photographic slides of trees in natural settings were used in an initial study in which human tree preferences were measured. These photographs were of a wide variety of trees, ranging from tree parts to full, leafy trees. A variety of settings were also depicted in the stimuli. Tree preference ratings for all possible permutations of pairs of these stimuli were obtained from four human subjects by having them indicate which member of each pair was the better tree. Subjects also indicated the extent of their preference on a seven point scale, as well as the perceived similarity between members of each stimulus pair.

An initial MDS analysis of the (symmetric) similarity ratings data was used to select a nonredundant subset of twelve stimuli to use for animal testing. Four ? pigeons and four squirrel monkeys were run in the experiment. The animals were initially trained using a forced choice discrimination procedure to choose a tree from a nontree by pressing the screen on which the tree picture was projected. Animals remained in this phase of the experiment until 85% accuracy was achieved over a block of five

sessions. In the next phase, novel tree-nontree pairs were inserted among the training slides, and both pigeons and monkeys were highly accurate at choosing the novel tree picture in these pairs. In the final phase of the experiment, six probe trials were inserted among the training trials on each session. These probe trials consisted of pairs of tree stimuli selected from the twelve stimuli we wished to examine. Thus, the animal was forced to make a tree-preference choice between members of pairs of these critical stimuli. This procedure continued until all possible permutations of pairs of stimuli had been presented twice.

Analysis of the Human Data

The human tree preference data was averaged across the four subjects, and DEDICOM was applied to this average matrix. Two preference patterns that accounted for 69% of the variance in the data were recovered, and are presented in Figure 1. In this figure, the open circle represents the origin of the preference pattern, and stimulus objects falling near this origin are *not* strongly involved in the pattern. The arrows indicate the direction of preference in the pattern; a stimulus object at the tail of an arrow is preferred over stimuli pointed to by the head of the arrow.

The first human tree preference pattern indicated that leafy tree wholes were seen as more tree-like than leafy tree parts. The most preferred stimuli in the pattern were two full, leafy trees, while the least preferred stimuli were leafless tree parts. The whole vs. part pattern is apparent if one follows the axis of the pattern (solid line) in the direction of preference: the full leafy trees are encountered first, followed by flowering trees and a coloured fall tree, followed by pictures of tree parts in which some leaves are present. The second pattern indicated a similar structure in which the presence of leaves was not as important. In this second pattern, full tree structures with leaves, flowers or needles were preferred over less typical tree structures (a lone leafless tree and a stunted evergreen), which in turn were preferred over the stump and root.

Analysis of the Pigeon Data

A DEDICOM analysis of the averaged pigeon data revealed two striking differences from the results obtained in analysing the human data. First, only one preference pattern was recovered. This pattern, which accounted for ??% of the data variance, is presented in Figure 2. The second difference is noted in the interpretation of this pattern, as it was qualitatively different from either pattern in Figure 1. For instance, the fall deciduous tree, one of the more preferred stimuli for humans, was the least preferred stimulus for pigeons. Similarly, the stump that was one of the least preferred stimuli for humans was one of the more preferred stimuli for pigeons. In general, the entire preference pattern for pigeons suggested that stimuli that had a very well-defined "branchy" structure (i.e., long narrow projections not obstructed by dense masses of leaves, and distinct from the background) were selected by the pigeons as being more tree-like than stimuli that had a poorly defined branchy structure.

Analysis of the Monkey Data

The DEDICOM analyses of the monkey data were performed on the individual subjects' datasets, as preliminary investigations indicated a variety of individual differences. One of the animals demonstrated a single preference pattern quite similar to the structure observed in the pigeon data. Two of the monkeys demonstrated two preference patterns apiece. One of these patterns indicated a preference of well-defined "branchy" structures over poorly-defined "branchy" structures. The other of these patterns indicated a preference of stimuli representing solitary branches over stimuli in which several branches were evident. The final monkey demonstrated three separate preference patterns, accounting for ??% of the data variance, which are illustrated in Figure 3. The first of these patterns revealed the preference for well-defined "branchy" structures observed in the other animals. Note the outlying position of the evergreen in this pattern, suggesting that the animal was aware of some difference between this stimulus and the other stimuli that were used, which were predominately deciduous trees. The second pattern involved preferences among the subset of stimuli that were primarily leafless, indicating that the animal was sensitive to the presence or absence of

leaves in the photographs. In this pattern, full solitary branching structures were preferred over more complicated branching structures. The final pattern revealed preference relations among (roughly) the subset of stimuli that possessed leaves. In this pattern, stimuli that had more leaves were preferred over stimuli that had few leaves.

In performing a DEDICOM analysis, use is made of a fit-to-dimensionality curve that plots the goodness of fit of a solution as a function of the number of directional patterns in the solution. This curve is used to choose the most appropriate solution for a dataset. This curve can be interpreted as providing an index of how well defined a directional structure is (c.f., Cattell, 1978). In general, if this curve is very steep and then flattens sharply, the structure is well defined (i.e., free of noise), while if the curve tends to have a more gentle initial slope, and does not sharply flatten, the structure is not well defined. Figure 4 illusrates the fit-to-dimensionality curve obtained for each monkey along with a curve inidicating the performance of the animal when learning the initial tree/nontree discrimination task. It is evident from this figure that the more well-defined the preference structure was, the faster did the animal learn the initial task. This suggests that performance on the initial learning task was related to the animal's ability to use a well-defined representational structure when making the discrimination.

Discussion

The major result of this study was that there were noticeable differences between species in terms of the preference patterns observed for the set of tree stimuli that were examined. While human subjects appeared to base their judgement on how complete stimuli appeared to be, both monkeys and pigeons were more sensitive to the "branchiness" of the stimuli. Individual differences between monkey tree preference patterns were also noted, suggesting that different animals used representations that were sensitive to different stimulus attributes. The monkey data also indicated that how well defined a preference pattern was (as indicated by fit-to-dimensionality curves) was related to performance on the discrimination task.

Although clear differences between species were noted, the reason why these

differences were found is not as evident. It could be that in the tree/nontree discrimination task, the most reliable or salient feature of a tree is the presence of some branching structure. Thus, when confronted with two different tree photographs, the animal selects the stimulus that has the most branches. However, this is apparently not the only process involved. For example, the patterns illustrated in Figure 3 show that one monkey was also sensitive to the presence and absence of leaves, as well as the number of branching structures in a stimulus.

A more speculative account of the observed differences is in terms of the functional nature of different cognitive contents. It is possible that the preference for "branchiness" in the stimuli is related to the fact that pigeons view trees primarily as places to perch, and that monkeys view trees in terms of places to climb. Structures with well-defined branches are presumably structures that can be perched upon or climbed upon quite readily, and may therefore be encoded as "good" trees. Unfortunately, the complexity of the stimuli that we used prevents firm conclusions of this type to be drawn. However, we feel that these results indicate a fruitful approach to studying the conceptual structure in animals. The current results show that a particular variable (i.e., "branchiness") is very important in tree-preference patterns in both pigeons and monkeys. A similar study using a more controlled set of stimuli, in which "branchiness" was systematically varied, might begin to provide a more precise characterization of the encoding of the concept "tree" in these animals.

References

Griffin, D.R. (1978). Prospects for a cognitive ethology. *The Behavioural and Brain Sciences, 4*, 527-538.

Harshman, R., Green, P., Wind, Y., & Lundy, M.E. (1982). A model for the analysis of asymmetric data in marketing research. *Marketing Science, 1*, 205-242.

Mazmanian, D.S., & Roberts, W.A. (1983). Spatial memory in rats under restricted viewing conditions. *Learning & Motivation, 12*, 261-281.

Premack, D. (1983). The codes of man and beasts. *The Behavioural and Brain Sciences, 6*,

125-168.

Pylyshyn, Z.W. (1984). *Cognition And Computation.* Cambridge, Mass.: MIT Press, a Bradford book.

Roberts, W.A., Mazmanian, D.S., & Kraemer, P.J. (1986). Memory for picture fragements in monkeys and humans. *Canadian Journal of Psychology,* submitted.

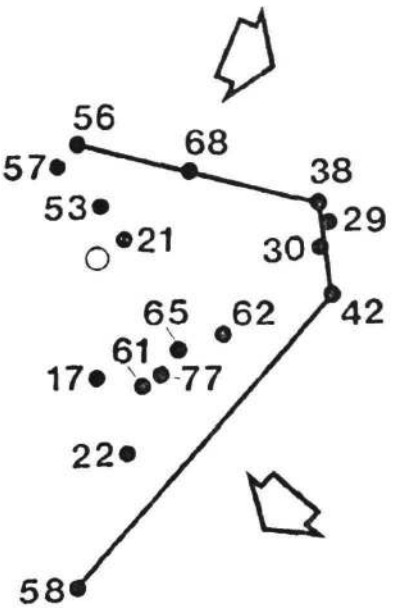
Roitblat, H.L. (1982). The meaning of representation in animal memory. *The Behavioural and Brain Sciences, 5,* 353-406.

Roitblat, H.L., Bever, T.G., & Terrace, H.S. (1984). *Animal Cognition.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

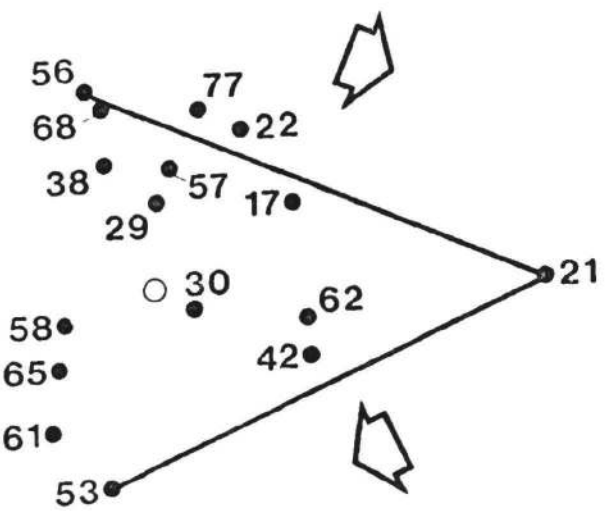
Sands, S.F., Lincoln, C.E., & Wright, A.A. (1982). Pictorial similarity judgements and the organization of visual memory in the Rhesus monkey. *Journal of Experimental Psychology: General, 111,* 369-389.

Figure 1

DEDICOM solutions for human tree preference data.



A



B

17 Branch with some leaves

21 Leafless deciduous tree

29 Full fall deciduous tree

30 Flowering tree

38 Flowering magnolia

42 Poor evergreen

53 Stump

56 Full deciduous tree

57 Full deciduous tree

58 Leafless branch

61 Root

62 Trees by water; not very leafy

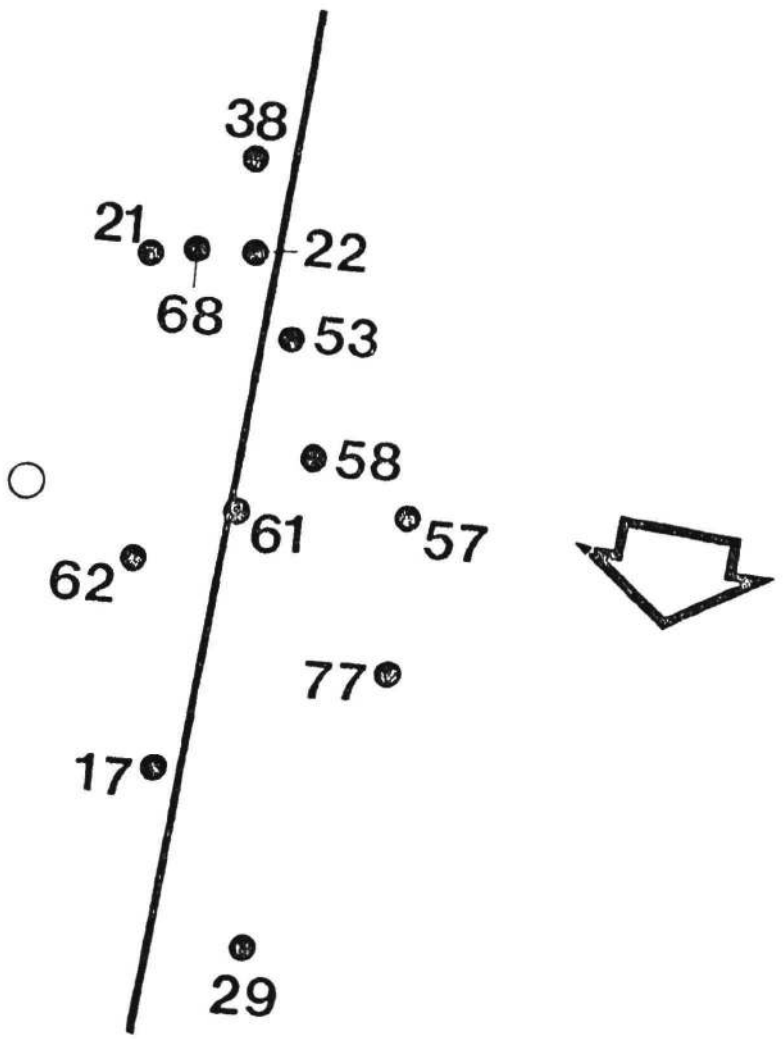
65 Tree trunk

68 Full evergreen

77 Leafless tree by building

Figure 2

The single bimension removed from the pigeon tree preference data.



17 Branch with some leaves

21 Leafless deciduous tree

29 Full fall deciduous tree

38 Flowering magnolia

53 Stump

57 Full deciduous tree

58 Leafless branch

61 Root

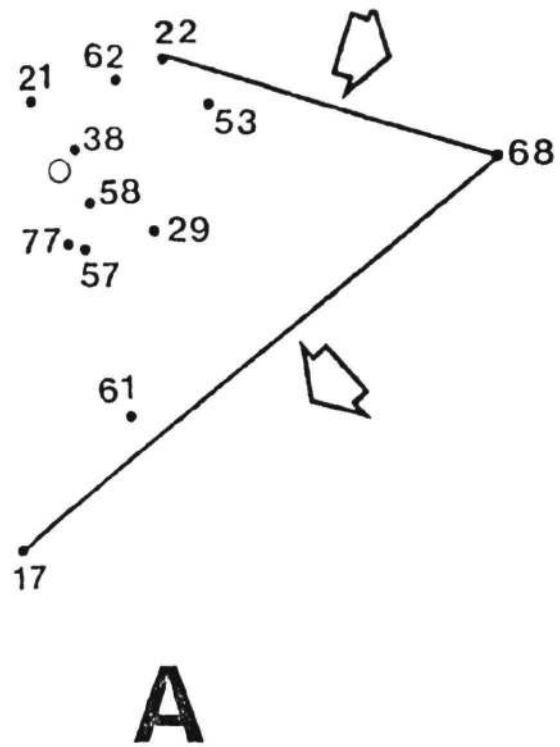
62 Trees by water; not very leafy

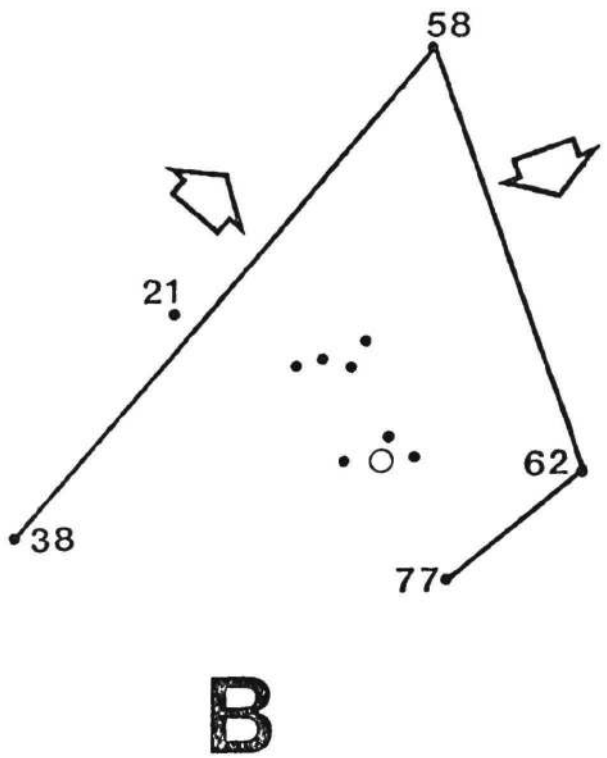
68 Full evergreen

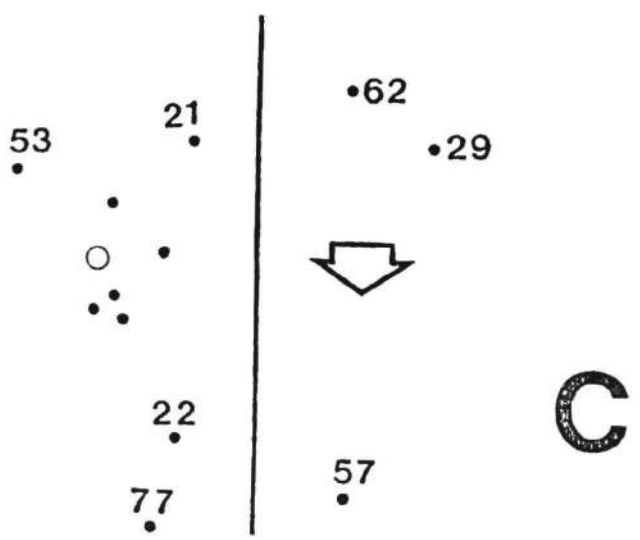
77 Leafless tree by building

Figure 3

Three bimensions recovered from a single monkey's tree preference data.







17 Branch with some leaves

21 Leafless deciduous tree

29 Full fall deciduous tree

38 Flowering magnolia

53 Stump

57 Full deciduous tree

58 Leafless branch

61 Root

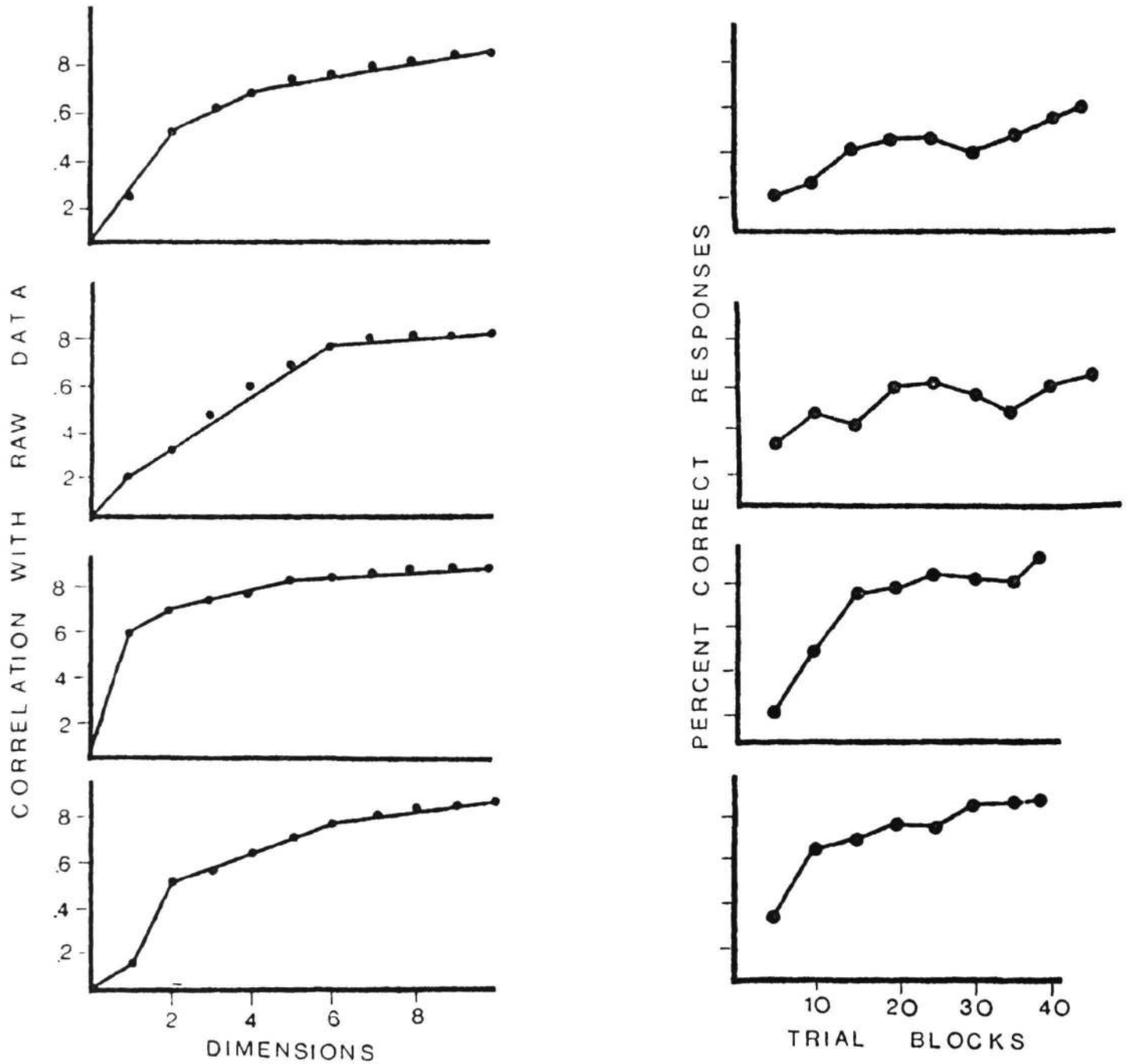
62 Trees by water; not very leafy

68 Full evergreen

77 Leafless tree by building

Figure 4

The fit-to-dimensionality curves and the learning curves for the individual monkeys.



Is Distributed Connectionism Compatible with the Physical Symbol System Hypothesis?

Mark Derthick and David C. Plaut

Computer Science Department
Carnegie-Mellon University

1. Introduction

All existing intelligent systems share a similar biological and evolutionary heritage. Based on the conviction that cognition is computation, artificial intelligence researchers are investigating computational models as a means of discovering properties shared by all intelligent systems.

One property that has been proposed as central to intelligence is the ability to construct and manipulate symbol structures. If intelligence may be described completely in terms of symbol processing, then cognitive science need not be concerned with the particular physical implementation details of either artificial or biological examples; neuroscience would no longer be part of cognitive science. On the other hand, if important aspects of intelligence evade symbolic explanation, it may prove necessary to consider phenomena below the symbol level. The connectionist approach to artificial intelligence is founded on the conviction that the structure of the brain critically constrains the nature of the computations it performs. However, if the symbolic position is correct and neural networks only implement symbol systems, then connectionism contributes little to cognitive science.

The notion of intelligence as symbol processing was made explicit by Newell and Simon with the Physical Symbol System Hypothesis (PSSH) (Newell & Simon, 1976, Newell, 1980) and the Knowledge Level Hypothesis (KLH) (Newell, 1982). Taken together, these hypotheses have significant implications for the nature of any system capable of general intelligence. We examine a number of connectionist systems in light of the hypotheses and distinguish three kinds: (1) *rule-based* systems, which are symbol systems; (2) *rule-following* systems, which are symbol systems only under a weakened version of the PSSH; and (3) systems which are not rule-following, and thus are not symbol systems even in a weak sense.

According to the PSSH, non-symbolic connectionist systems must be incapable of general intelligence. There are strong arguments both for and against this conclusion. On the one hand, such connectionist systems may provide more parsimonious accounts of certain cognitive phenomena than do symbolic approaches. On the other hand, these connectionist systems have significant limitations, relating to universality, not shared by symbol systems. We conclude that a comprehensive theory of intelligence may require a hybrid model that combines the strengths of both approaches.

2. Physical Symbol Systems

The PSSH states that a physical symbol system, defined as "a machine that produces through time an evolving collection of symbol structures" (Newell & Simon, 1976, p. 116), has the necessary and sufficient means for general intelligent action. Newell and Simon explain,

> By "necessary" we mean that any system that exhibits general intelligence will prove upon analysis to be a physical symbol system. By "sufficient" we mean that any physical symbol system of sufficient size can be organized further to exhibit general intelligence. (Newell & Simon, 1976, p. 116)

Of course, since symbol systems are universal they are sufficient for carrying out *any* behavior. To accept the sufficiency condition on this basis, however, would be to fall prey to the "Turing tarpit" (Newell, 1980), in which significant structural differences are blurred under the notion that all universal systems are equivalent. In order for the sufficiency claim to be substantive, the further organization required to exhibit general intelligence must not resort to simulating a non-symbolic system.

Physical symbol systems[1] are composed of

1. a collection of *symbols*, each a discrete, identifiable physical pattern in a machine;
2. *symbol structures*, or expressions, composed of symbols related in some physical way;
3. *processes* operating on expressions to produce other expressions.

A symbol structure *designates* another symbol structure or process if having the first structure allows behavior that affects or depends on the second. The system can *interpret* an expression if the expression designates a process and if, given the expression, the system can carry out the process. This formulation has a number of consequences for the nature of any physical symbol system (Newell & Simon, 1976), the most important for our purposes being:

1. A symbol may be used to designate any expression whatsoever.
2. There exist expressions that designate every process of which the machine is capable.
3. The number of expressions that the system can hold is essentially unbounded.

We will argue in the next section that these characteristics exclude certain connectionist systems from the class of physical symbol systems.

The semantics of a physical symbol system is developed in the context of the relation between the symbol level and a higher, *knowledge level*, in which the behavior of the system is described in terms of knowledge, goals, and actions (Newell, 1982). The Knowledge Level Hypothesis states that the knowledge level is implemented directly by the symbol level. Knowledge level entities are represented by particular symbol level structures, and each symbol structure has a coherent interpretation at the knowledge level. In other words, symbols and symbol structures are the formal entities of a physical symbol system that are given a semantic interpretation.

With these characteristics of physical symbol systems in mind, we turn to an analysis of the relationship between symbol systems and connectionist systems.

3. Connectionist Systems

We take the essential characteristic of a connectionist system to be the existence of a physical level description in terms of the operation of a large number of very simple computing devices (units) locally interacting across very low bandwidth channels (connections). Such an architecture is quite different from that underlying standard physical symbol systems. However, the fact that connectionist systems are built out of units and connections does not bear on the question of whether they are symbol systems. The critical issue here is the relation between formal operations of the system and what they represent. In this regard, certain types of "localist" connectionist systems (Feldman & Ballard, 1982, Cottrell, 1984, Shastri & Feldman, 1984), in which the activity of individual units may be given a coherent knowledge level semantics, meet the requirement of physical symbol systems that the formal level map directly to the semantic level.

On the other hand, "distributed" connectionist systems (Hinton & Anderson, 1981, McClelland & Rumelhart, 1986), in which a knowledge level semantics is ascribed only to *patterns* of activity of a large number of units, present difficulties for any attempt to assign to the system the type of formal semantics required of physical symbol systems. In particular, the formal level of the system (i.e. the interaction of

[1]Our analysis is based solely on the formulation of symbol systems developed by Newell and Simon, and does not exclude the possibility of alternative formulations which might encompass the connectionist systems we discuss.

units via connections) and the semantic level (i.e. the interactions of patterns of activity) do not correspond, nor do they operate according to the same principles. While units obey formal input/output rules (specified in terms of unit activities and connection strengths), the interaction of patterns of activity *as patterns* need not be formal. There are various ways for the interaction of patterns in such systems to generate knowledge level behavior. In the following sections we make three distinctions among distributed connectionist systems: *rule-based* systems, with explicitly encoded rules; *rule-following* systems, with implicitly encoded rules; and systems whose behavior is not strictly rule-following.

3.1. Explicit Rules

We first examine a distributed connectionist system in which the interaction of the patterns of activity *is* governed by explicit rules and is therefore formal. Touretzky and Hinton (1985) have developed a connectionist implementation of a production system. Production memory consists of sets of units, each dedicated to a particular production. The units in each set are wired to units which are active in the representation of working memory elements that are matched, added, or deleted by the production. Although the behavior of the system can be explained in terms of interactions between individual units, a higher level explanation making reference to the production rules can be used. Furthermore, this explanation is not just a way of speaking; it corresponds directly to physical structure. This is just what Newell and Simon expect will be the case for any intelligent system: the symbol level may be implemented in various technologies, but it is a real, necessary system level. The "Touretzky tarpit" (in contrast to the Turing tarpit) traps those who gratuitously distinguish this system from symbolic ones by attributing psychological importance to its underlying connectionist basis. It *is* symbolic, and any theory based on it could as well be non-connectionist.

3.2. Implicit Rules

In contrast, a connectionist system developed by Rumelhart and McClelland (1986) exhibits rule-following behavior without containing explicit representations of the rules. The task is to form a phonological representation of the past tense of English verbs from a phonological representation of the present tense form. Linguists typically model this task with a large number of rules, which form a hierarchy of exceptions, exceptions to exceptions, and so on. The rule for regular verbs consists of adding /ed/. Irregular verbs may be grouped according to other rules, such as changing /ing/ to /ang/ (*sing*/*sang*), changing /d/ to /t/ (*build*/*built*), etc.

Instead of using explicit rules, Rumelhart and McClelland's system captures the rule following nature of forming past tenses in terms of regularities between the substructures of the phonological codes of the present and past tense forms. This substructure is represented in terms of the activity of a set of units, each representing a context-sensitive triple of phoneme features (called "Wickelfeatures" in the spirit of Wickelphones (Wickelgren, 1969)). The important characteristics of this code are that it can sufficiently discriminate between any two English verbs, and that it provides a natural basis for generalizations to emerge about what aspects of a present tense form correspond to what aspects of the past tense form (Rumelhart & McClelland, 1986).

In the model, a fixed encoding network converts the actual phonological representation of the present tense form into a slightly blurred pattern of activity over a large set of *input* units, each representing a particular Wickelfeature. Each input unit is connected to each member in a similar set of *output* units for representing the phonological substructure of the past tense form. The activity of the output units is then decoded by a second fixed network into its corresponding phonological representation. The goal of the network is to produce in the output units the pattern of activity representing the past tense form given the pattern of activity over the input units for the present tense form. The network is presented with the codes for a large number of present/past tense pairs, and a learning algorithm modifies the strengths of the connections between the input and output units to reduce for each pair the difference between the correct phonological representation and the one

produced by the network. As it learns the task, the performance of the network passes through three important stages (which we describe below), eventually producing the appropriate rule-following and exception behavior, as demonstrated by proper generalization to verb pairs not seen in the training.

The way in which Rumelhart and McClelland's system produces rule-following behavior violates an important constraint on the structure of a physical symbol system; processes in the system do not have symbol structures which designate them. Regularities between the substructure of present and past tense forms are encoded (in connection strengths) in terms of the interaction of "microfeatures" (in this case, Wickelfeatures). The presence of each microfeature in the input representation provides support for some microfeatures in the output representation while inhibiting others. These "microinferences" allow the substructure of the input to be combined in very complex and subtle ways to produce the appropriate substructure for the output. The actual semantic rules (which the input/output activity patterns can be described as following) are nowhere stated explicitly, but emerge from complex interactions among the microinferences (Hinton, 1981).

The lack of explicit rules excludes this system from the class of physical symbol systems. However, rule-following connectionist systems are compatible with a weaker version of the PSSH. It could still be maintained that intelligence can be explained with a rule-based system, regardless of the fact that it is also possible to explain intelligence in terms of a system which is only rule-following. This corresponds to interpreting the symbol level primarily as a means of *explaining* knowledge level behavior, and not necessarily as the means of *implementing* it.

3.3. Not Rule-following

A major contribution of Rumelhart and McClelland's system is modeling the stages that children pass through in acquiring the ability to form past tenses: (1) an initial stage in which all past tenses are learned as separate words; (2) an intermediate stage in which an insufficient number of rules are used to form all past tenses, resulting in overregularization; and (3) a final stage in which more and more rules are learned, so all known verbs are handled correctly and novel verbs generalize appropriately. While the stages are relatively well defined, the transition from one stage to another is gradual, so that at times a child may use several past tense forms of the same verb in the same conversation. Such behavior is difficult to account for using rules, but is explained quite elegantly (in terms of competing microinferences) in the connectionist model. While the system can be described as following rules once the ability to form past tenses has been fully learned, the system viewed as evolving over time cannot be given an adequate formal symbol level *explanation*, and hence is not a symbol system even in the weaker sense described above.

Of course, a rule-based approach with a very large set of highly interacting, fine-grained rules which fire in parallel might succeed in reproducing such graded behavior.[2] In general, rule-based systems embody symbol systems, and so by virtue of their universality they are, in principle, capable of reproducing *any* behavior (the Turing tarpit argument). Yet as the complexity of such systems increases, it becomes more and more difficult to give a clear account, in terms of the task or environment, of what a single rule is doing. More and more of the structure of a rule depends on the entire set of rules with which it interacts. The system becomes a less and less parsimonious *rule-based* explanation, and more and more similar to a microinference-based *connectionist* explanation.

4. Limitations of Connectionist Systems

The PSSH states that only symbol systems are capable of general intelligence. As we have shown, there are connectionist systems that are not symbol systems even in a weak sense. Therefore, either the PSSH is false, or connectionist systems of this type cannot be extended

[2] In fact, the evolution of large expert systems, such as R1 (McDermott, 1982, Bachant & McDermott, 1984), seems to lead in this direction.

from limited domains to general intelligence. We contend that non-symbolic connectionist systems *are* limited with respect to symbol systems: they are not universal, and hence lack the flexibility to potentially carry out any possible behavior.

Any computational system is implemented by what Pylyshyn (1980) calls the *functional architecture*; "the fixed functional capacities provided by the biological substrate...out of which cognitive processes are composed." Pylyshyn (1984) argues that the essential difference between connectionist systems and symbol systems hinges on significant differences in their functional architectures. In a symbol system, an unbounded number of symbols may be manipulated by a finitely specified control. In a connectionist system, *all* of the knowledge used to carry out a process is contained within a finite structure; control and data are not separate. The unbounded storage capacity that underlies the universality of symbol systems is, in principle, not available in the specification of connectionist systems.

It could be argued that if an unbounded pool of uncommitted units were available, and if learning could take place during the execution of an operation, then it might be possible to store a potentially unlimited amount of knowledge. However, allowing learning to be a primitive operation of the functional architecture would be allowing universality in by the back door. By definition, the capabilities of the functional architecture are fixed. In a connectionist system learning is a method for *modifying* the functional architecture and cannot be controlled at the semantic level.

Yet Touretzky and Hinton's system *can* add a potentially unbounded amount of knowledge without the use of learning. Productions match clauses represented in *clause spaces*, which are groups of units that can represent one clause at a time. The clauses which correspond to stable patterns are not determined by connection strengths, but by the states of units in working memory space. Connectionist systems require learning to modify weights, but they modify states during ordinary operation. Thus it is possible in this system to store an arbitrary amount of knowledge in working memory and have it appropriately affect clause retrieval. Changing the states of working memory units is done by the production memory units, so universality appears exactly when the division between control and data does.

5. A Hybrid System

We have seen how generalization arises naturally in non-symbolic connectionist system, but that such systems are limited with respect to the universality of symbol systems. In order to determine the relative significance of these two approaches, it is necessary to characterize the processes that must be accounted for by a comprehensive theory of intelligence.

Pylyshyn (1980) argues that the class of processes that psychology is committed to explain are those that are "cognitively penetrable"; processes whose function depends on the agent's beliefs and goals in meaningful ways. Any function that is cognitively *impenetrable* represents a fixed primitive operation of the functional architecture. By definition, these primitive operations are not symbolic. Therefore, it is possible to suppose that, for example, past tense formation is primitive. In general, any process found to require a non-symbolic explanation may be relegated to the functional architecture. However, if too much of cognition is swept under the primitive rug, there is nothing left for a theory of intelligence to do: psychology becomes trivial.

Pylyshyn claims that *all* cognitively penetrable processes are symbolic. We believe that processes like past tense formation are both cognitively penetrable and best carried out non-symbolically.

Therefore, we claim that a comprehensive theory of intelligence will require both symbolic and non-symbolic processes. The relative strengths and weaknesses of connectionist and symbolic approaches suggests a natural division of labor within such a hybrid system. A large collection of task-specific connectionist modules would carry out overlearned processes under the executive control of a flexible symbol system. The fact that this division is aligned with a number of classic dichotomies used to charac-

terize various aspects of cognition (e.g. "controlled vs. automatic" (Shiffrin & Schneider, 1977), "central vs. peripheral" (Fodor, 1983)) suggests that it may reflect a real structural property of the cognitive architecture.

6. Conclusion

We have shown that there are connectionist systems which generate knowledge level behavior but fall outside Newell and Simon's characterization of physical symbol systems. The claim of the Physical Symbol System Hypothesis that symbol processing is necessary for intelligence implies that such systems cannot be extended from simple domains to general intelligence. The lack of universality of non-symbolic connectionist systems supports the claim that some symbolic processing is necessary for intelligence.

On the other hand, non-symbolic connectionist systems, like Rumelhart and McClelland's, provide a more satisfactory explanation of certain aspects of cognition than do symbol systems. This supports the claim of the connectionist approach that implementation details are important to understanding cognition, and is incompatible with the claim of the PSSH that symbol processing is *sufficient* for general intelligence. It is striking and perhaps quite significant that what is natural for each approach is quite difficult and unwieldy in the other. A theory of intelligence incorporating both connectionist and symbolic components would be more capable than either approach alone of integrating the various aspects of cognition into a comprehensive explanation of intelligent behavior.

Acknowledgements

We wish to thank the Boltzmann research group at Carnegie-Mellon University for useful discussions on the relationship between symbol systems and connectionist systems, and David Ackley, Ron Brachman, Hector Levesque, Jay McClelland, Steven Nowlan, Benjamin Pierce, and David Touretzky for providing useful comments. This research was supported by an ONR Graduate Fellowship to the first author, and an R.K. Mellon Fellowship to the second author.

References

Bachant J. and McDermott J. R1 revisited: Four years in the trenches. *AI Magazine*, 1984, *5*, 21-32.

Cottrell G.W. A model of lexical access of ambiguous words. *Proceedings of the National Conference on Artificial Intelligence*, Austin, TX, August, 1984.

Feldman J.A. and Ballard D.H. Connectionist models and their properties. *Cognitive Science*, 1982, *6*, 205-254.

Fodor J.A. *The modularity of mind.* Cambridge, MA: M.I.T. Press, 1983.

Hinton G.E. Implementing semantic networks in parallel hardware. In G.E. Hinton and J.A. Anderson (Eds.), *Parallel models of associative memory,* Hillsdale, NJ: Lawrence Erlbaum Associates, 1981.

Hinton G.E. and Anderson J.A. (Eds.). *Parallel models of associative memory.* Hillsdale, NJ: Lawrence Erlbaum Associates, 1981.

McClelland J.L. and Rumelhart D.E. (Eds.). *Parallel distributed processing: Explorations in the microstructure of cognition.* Cambridge, MA: Bradford Books, 1986.

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

Newell A. Physical symbol systems. *Cognitive Science*, 1980, *4*, 135-183.

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

Newell A. and Simon H.A. Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, 1976, *19*, 113-126.

Pylyshyn Z.W. Computation and cognition: Issues in the foundations of cognitive science. *Behavioral and Brain Sciences*, 1980, *3*, 111-169.

Pylyshyn Z.W. Why "computing" requires symbols, pages 71-73. *Proceedings, 6th Conference of the Cognitive Science Society*, Boulder, CO, 1984.

Rumelhart D.E. and McClelland J.L. On learning the past tenses of English verbs. In J.L. McClelland and D.E. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition,* Cambridge, MA: Bradford Books, 1986.

Shastri L. and Feldman J.A. *Semantic networks and neural nets.* Technical Report 131, Computer Science Department, University of Rochester, January 1984.

Shiffrin R.M. and Schneider W. Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review*, 1977, *84*, 127-190.

Touretzky D.S. and Hinton G.E. Symbols among the neurons: Details of a connectionist inference architecture. *Proceedings, 9th International Joint Conference on Artificial Intelligence*, Los Angeles, August, 1985.

Wickelgren W.A. Context-sensitive coding, associative memory and serial order in (speech) behavior. *Psychological Review*, 1969, *76*, 1-15.

TOPOLOGICAL AND DYNAMICAL ASPECTS OF A NEURAL NETWORK MODEL FOR GENERATION OF PURSUIT MOTOR PROGRAMS

R. ECKMILLER
University of Düsseldorf, FRG

A B S T R A C T

A computational model of a motor program generator (MPG) for horizontal pursuit eye movements (PEM) is proposed. The MPG model consists of two neural networks (velocity maps). Neurons are arranged in a single circular layer with lattice structure and connected only to their immediate neighbors. During PEM one of the two maps always features an activity peak (AP) which travels with constant velocity from one neuron to the next. A memory trace of the most recent portion of the trajectory is created by means of a temporary increase of the interneuronal connectivity strength between previously activated neurons. This novel MPG model may be useful for designing parallel processors for motor control of robots.

I N T R O D U C T I O N

The new areas of Cognitive Science, Artificial Intelligence, and Robotics share a common interest with Neuroscience in the various topological (structural) and functional principles of the primate central nervous system, which uses the general concept of parallel processing and a hierarchical organization of numerous neural maps. One important function of the primate brain that has concerned neuroscientists for many years is Motor Control. Assuming the predominantly accepted view is correct that movements in primates are based on internally generated motor programs rather than on reflectory responses to sensory stimuli, the fundamental and still open question is: How are motor program generators (MPG) (which certainly consist of neurons and synapses) designed in detail to control the motor activity time courses of various sets of muscles for speech movements, eye movements, or limb movements?

Neither neurophysiological nor neuroanatomical research has so far been able to offer answers or even plausible hypotheses regarding the architecture and dynamics of motor program generators, although an enormous amount of data has been compiled (Miles & Evarts, 1979; Tatton & Bruce, 1981). One possible exception is the generation of locomotion programs at the level of the spinal cord (Grillner & Wallén, 1985; Herman et al., 1976; Miller & Scott, 1977). By means of cognitive modelling several groups have recently begun to identify the problem of motor program generation and to search for neural networks with appropriate parallel processing features (Anderson, 1983; Arbib, 1981; Feldman & Ballard, 1982).

The present study leads to a proposal of a MPG for pursuit eye movements (PEM). This MPG model was developed on the basis of extensive single unit studies and accompanying lesioning and neuroanatomical studies on the pursuit control system in trained monkeys (Macaca fascicularis). Accordingly, the necessary assumptions concerning the network topology and dynamics of the MPG model are neurophysiologically plausible or even supported by direct neurophysiological evidence.

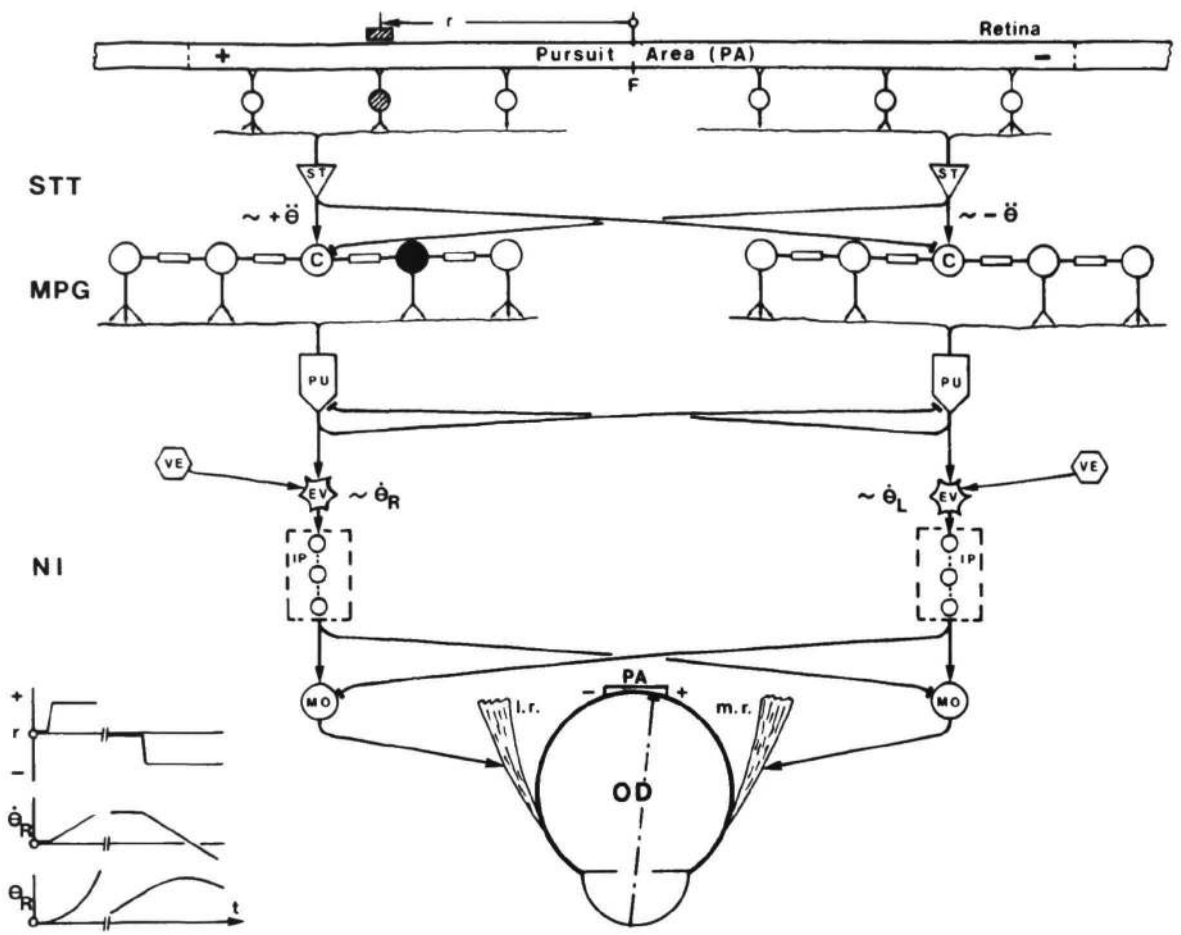


FIG. 1: Model of the oculomotor pursuit system in primates for horizontal pursuit of the right eye (OD). STT: spatio-temporal translation; MPG: motor program generation; NI: neural integration. Details, see text. Inset bottom left: relationship between position error r, eye velocity to the right $\dot{\theta}_R$, and eye position θ_R.

A) Neurophysiological, Neuroanatomical, and Behavioral PEM Data

PEM in the horizontal head plane (oblique or vertical PEM are assumed to be generated by additional signals from a vertical PEM system) in pursuit of maneuverable or predictable (e.g. sinusoidal) moving targets against a homogeneous background continuously minimize the position error r between the center of the fovea centralis on the retina and the retinal target projection. PEM approximately yields a continuous fixation of the moving target for the purposes of pattern recognition (Eckmiller, 1981; Eckmiller, 1986a). PEM can follow a continuous time course without large correctional saccadic eye movements as long as target velocity $\dot{\theta}_t$ is less than $\pm$ 50 deg/sec at target accelerations $\ddot{\theta}_t$ below $\pm$ 250 deg/sec^2. These maximal values imply that sinusoidal PEM at an amplitude of 10 deg will be possible for frequencies up to at least 0.8 Hz (Eckmiller, 1983; Eckmiller & Mackeben, 1978).

Recent neurophysiological studies in trained monkeys revealed that PEM must be controlled by two MPG for eye velocity to the left $\dot{\theta}_L$ and to the right $\dot{\theta}_R$. These two separate velocity signals are probably represented by two clusters of pursuit neurons (PU) close to the abducens nuclei in the brain stem (Eckmiller, 1983; Eckmiller & Bauswein, 1985). In fact, all oculomotor control

signals at the pre-motor level exist only as eye velocity signals and have to be transformed into eye position time courses by means of a neural integrator (Eckmiller, 1986b; Robinson, 1981).

B) Concept of the MPG model

The MPG model consists of two neural networks (velocity maps), which represent a continuum of velocity values $\dot{\theta}_R$ and $\dot{\theta}_L$ respectively, as locations on the map with zero in the center and increasing velocity (up to 50 deg/sec) at increasing radius R. The only input to each MPG network comes from ST neurons (Eckmiller, 1983) that carry the position error r to the center neuron C. Fig. 1 shows that positive values of r activate the left ST neuron. Its output is proportional to $\ddot{\theta}$ in the sense that it causes a change of $\dot{\theta}_R$ that is generated by the MPG network. If r becomes negative, then the ST neuron on the opposite side is activated and inhibits the center neuron in the left MPG network while exciting the center neuron in the right network.

C) Network Topology

Neurons are arranged in a single circular layer with lattice structure and connected only to their immediate neighbors as indicated in Fig. 2.

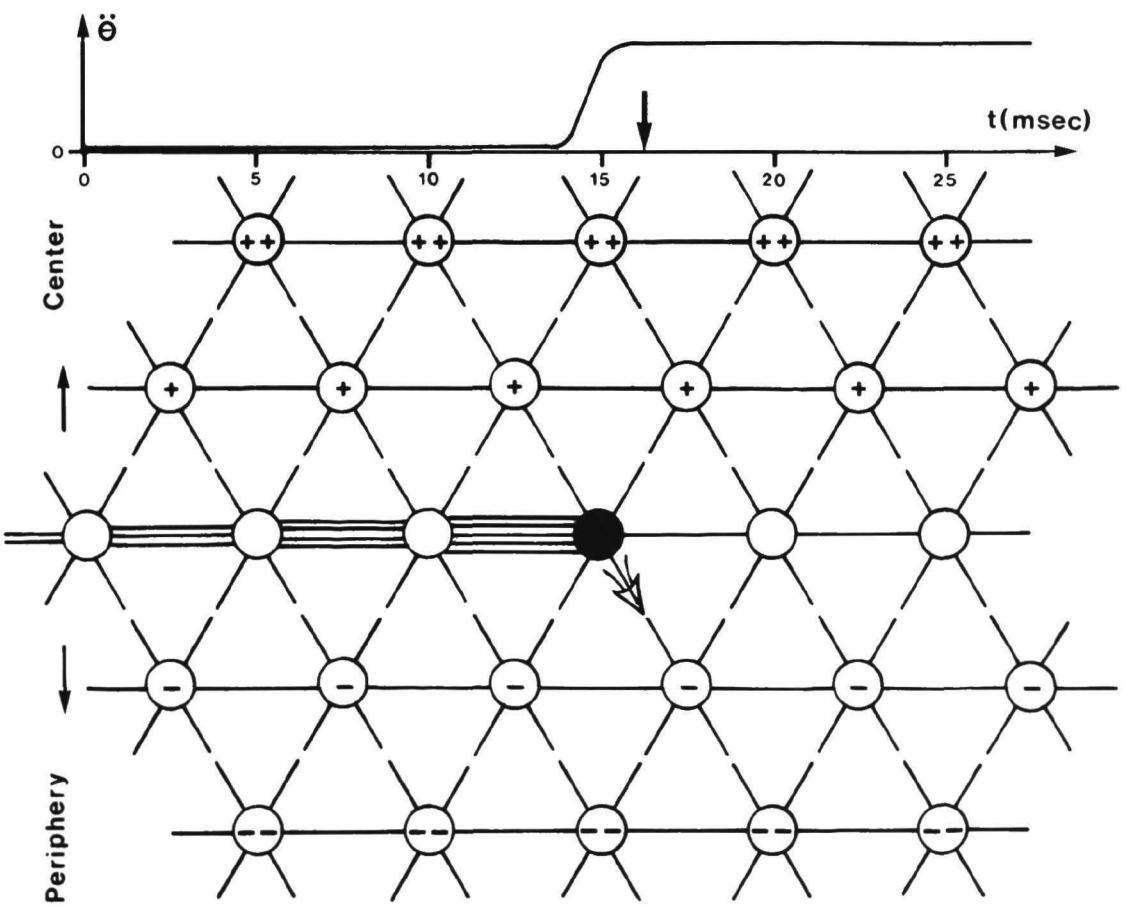


FIG. 2: Spatial distribution of neurons in the MPG network. Horizontal connections follow circles about the center. Activity peak, which is presently (see large arrow on the time scale) located at the central neuron (filled circle) of this network portion, had travelled from the left and is about to move downwards because of the sudden acceleration increase (time course at the top). The potential gradient is indicated by (++), (+), and (-), (--).

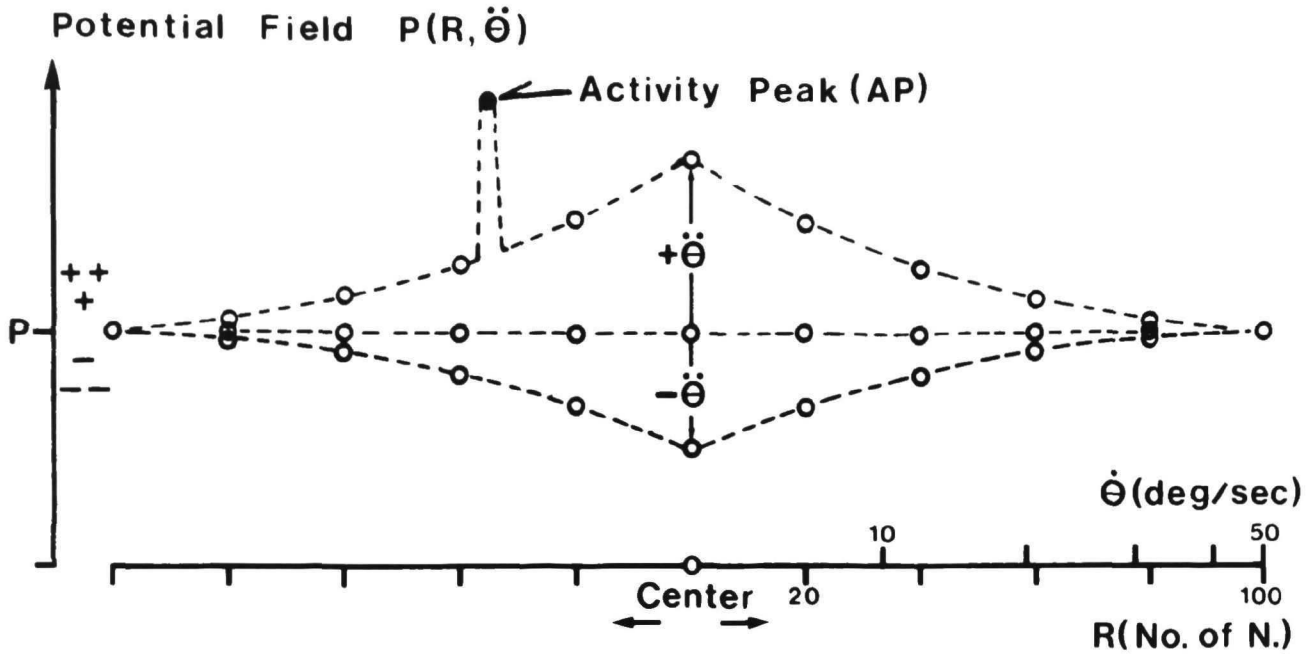


FIG. 3: Potential field $P(R,\ddot{\theta})$ along a cross section of the MPG map through the center at three different values of the input signal ($+\ddot{\theta}$, 0, $-\ddot{\theta}$). Abscissa gives the radius R as the number of neurons from the center as well as the eye velocity $\dot{\theta}$(deg/sec) with a possible non-linear scale. The activity peak AP is about to travel towards the periphery due to the potential gradient.

The connectivity strength c of tangential (horizontal in Fig. 2) connections c_t and radial connections c_r are initially constant, whereby c_t is slightly larger (indicated by a continuous line) than c_r (indicated by an interrupted line). This arrangement gives a slight preference to activity movements along a tangential trajectory. The time course at the top of Fig. 2 indicates an abrupt change in $\ddot{\theta}$ just before the movement (large arrow) at which the distribution of connectivity strength between the neurons and the activity state of each neuron is indicated. The activity state of each neuron is given by its potential P, which is initially constant throughout the network.

Fig. 3 indicates the potential field $P(R,\ddot{\theta})$ of the MPG network in a cross section through the center. The abscissa gives two possible and plausible scales: the radius R is measured by the number of neurons from the center; the eye velocity signal $\dot{\theta}$ can have a non-linear scale along the radius R as exemplified in Fig. 3. The topology can be changed by the input signal to the center neuron as indicated in Fig. 3 for several neurons at a positive ($+\ddot{\theta}$) and negative ($-\ddot{\theta}$) acceleration value. The potential field $P(R,\ddot{\theta})$ of all neurons is analogous to a flat circular membrane whose center can be pushed up or down by an amount proportional to the retinal position error r. The activity peak (AP) will be discussed later.

D) Network Dynamics

During PEM one of the two MPG networks, which are connected in a push-pull fashion (Fig. 1), always features an activity peak (AP) which travels with constant velocity v_T from one neuron to the next. The shape of the potential field defines whether AP travels in a circle ($\dot{\theta}$ = constant) in case of $\ddot{\theta}$ = 0, towards the periphery ($\dot{\theta}$ increase) or towards the center ($\dot{\theta}$ decrease). The

topology of the network assures that only one AP can exist at a time and that AP always travels with v_T = constant.
These two postulates are neurophysiologically plausible and can be simulated with little effort. The constant travel velocity of AP represents a linear passage of real time, which is an essential feature of any generator of time functions. AP is indicated in Fig. 2 as the filled circle neuron. This figure also shows that the connectivity strength of the most recently active connections is temporarily increased, as indicated by the number of connection lines. AP came along a circle (horizontal line in Fig. 2) from the left at a suggested travel velocity of v_T = 5 msec per jump between neighboring neurons

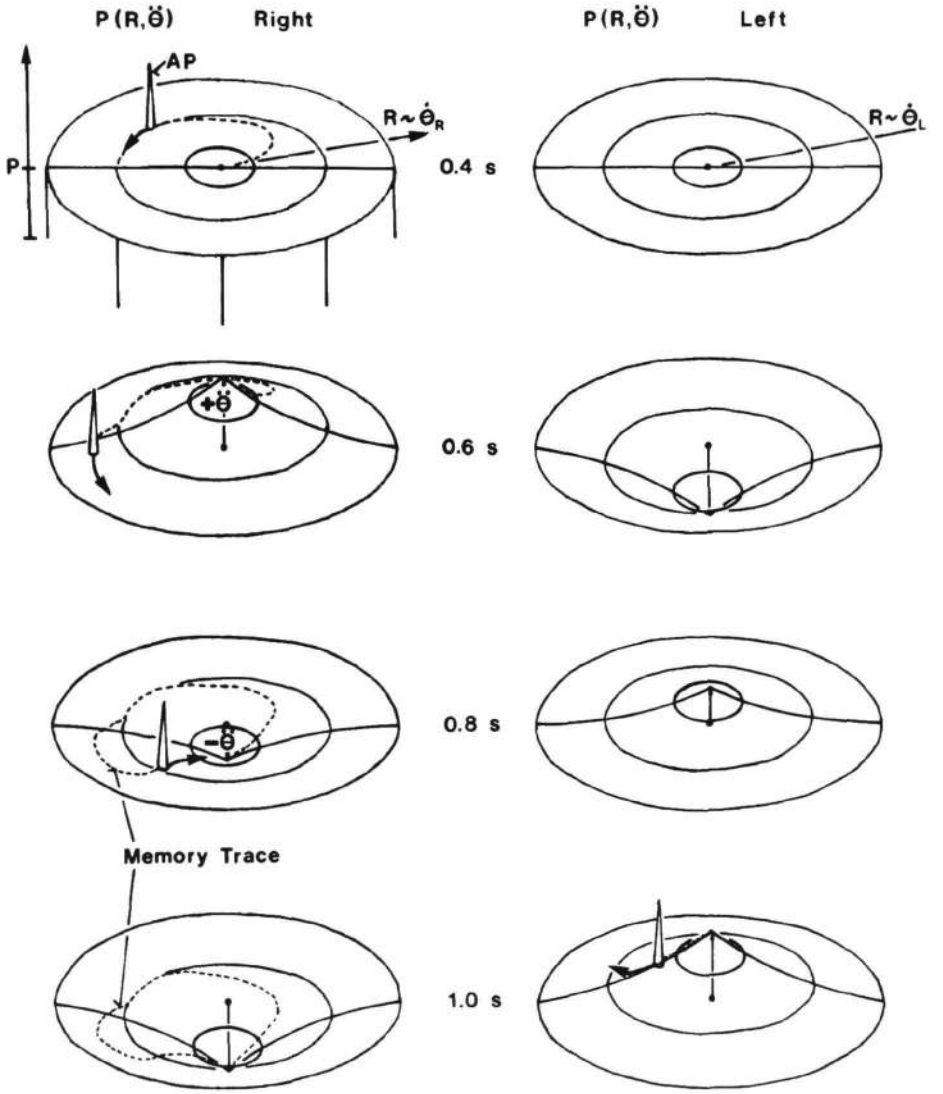


FIG. 4: Potential fields $P(R,\ddot{\theta})$ of the two MPG networks for generation of eye velocity time courses to the right ($\dot{\theta}_R$) and to the left ($\dot{\theta}_L$) at four different times. At 0.4 sec the activity peak AP is travelling in a circle since the potential field is flat. At 0.6 sec AP travels towards the periphery and at 0.8 sec towards the center. At 1.0 sec AP has already left the map for $\dot{\theta}_R$ and is travelling on the other map. Note that the potential fields of both maps are always identical except for the sign. The most recent trajectory is temporarily existent as a memory trace.

at $P(R,\ddot{\theta}) = \text{constant}$, when $\ddot{\theta}$ changed suddenly from zero to a positive value. This change led to the occurrence of a potential gradient as indicated in Fig. 2. AP will now travel towards the periphery. Depending on the size of the potential gradient, AP can travel always, or only occasionally towards more peripheral (or central) neurons, thereby generating trajectories with varying curvature on the velocity map. Alternating phases of $\dot{\theta}_R$ and $\dot{\theta}_L$ are made possible by repeated transfers of AP from the center neuron of one map to that of the other.

A memory trace of the most recent portion of the trajectory is created by means of a temporary increase of the connectivity strength as already mentioned. A small portion of such a memory trace is indicated in Fig. 2. Fig. 4 gives the events on both maps including the location of AP at four different times and the gradually increasing length of the memory trace.
Several features are noteworthy with regard to Fig. 4:

a) Trajectories on a map always start and end in the center,
b) the topology of both networks changes in opposite directions when $\ddot{\theta}$ changes from zero to a positive or negative value.
c) It is assumed that AP travels away from the center in the same angular direction under similar potential field conditions.
d) The memory trace gradually fades with a time constant of about 1 sec. Given a small potential gradient towards the center, AP can be pulled out of a weak memory trace and travel towards the center in a spiral after target disappearance.

This memory trace is used as explanation for the neural predictor mechanism (Eckmiller & Mackeben, 1978; Westheimer, 1954). Once a memory trace exists it can be used to reduce the amount of necessary updating (position error signals) during periodical pursuit movements, which are generated by repeated generation of the same two velocity trajectories on both maps. In such a case sudden and temporary target disappearance episodes can be bridged since AP simply follows the memory trace in the absence of a retinal input.

D I S C U S S I O N

The key feature of the proposed MPG model is the existence of an activity peak AP which travels with constant velocity on a neural map along a modifiable trajectory (Eckmiller, 1986c). This MPG model may also prove applicable to limb movements in addition to eye movements. A comparison of the model with the technical solutions that are presently available for generating various movement trajectories (including trajectory learning, memory, and prediction) indicates that it may be used for designing parallel processors for motor control of robots.

It is evident that the neural network as proposed for the MPG cannot be detected by means of single unit recordings. Simultaneous recordings from larger ensembles of neurons, however, may be able to monitor an AP as it travels over the map.

Desired velocity trajectories can be permanently stored by means of neurons that gradually develop connections only to those neurons in the map that are repeatedly activated by an AP during a learning phase (Eckmiller, 1986c). These pattern retrieval (PR) neurons can later be used to reactivate a strong memory trace with a single activation burst.

A C K N O W L E D G E M E N T

This research was supported by the Deutsche Forschungsgemeinschaft, SFB 200/A1.

R E F E R E N C E S

1. Anderson, J.A. (1983) Cognitive and psychological computation with neural models. IEEE Trans. SMC-13, 799-815.
2. Arbib, M.A. (1981) Perceptual structures and disturbed motor control. In: Hdb. Physiology, Section 1: The nervous system, Vol. II, Motor control, Part 2, J.M. Brookhart , V.B. Mountcastle, and V.B. Brooks (eds.) Baltimore: Williams & Wilkins, ch.33, pp. 1449-1480.
3. Eckmiller, R. (1981) A model of the neural network controlling foveal pursuit eye movements. In: Prog. Oculomotor Res., (Fuchs & Becker, eds.), Elsevier: New York, pp. 541-550.
4. Eckmiller, R. (1983) Neural control of foveal pursuit versus saccadic eye movements in primates - single unit data and models. IEEE Trans. SMC-13, 980-989.
5. Eckmiller, R. (1986a) Neural control of pursuit eye movements. Physiological Reviews (submitted).
6. Eckmiller, R.(1986b) The transition between pre-motor eye velocity signals and oculomotor eye position signals in primate brain stem neurons during pursuit. In: Adaptive processes in visual and oculomotor systems. (Keller & Zee, eds.), Pergamon Press: New York, in press.
7. Eckmiller, R. (1986c) Computational properties of a neural net with a triangular lattice structure and a travelling activity peak. Proc. 1986 IEEE Int. Conf. Systems, Man, and Cybernetics.
8. Eckmiller, R., & Bauswein, E. (1985) Smooth pursuit eye movements. In: Oculomotor and skeletomotor system - Differences and similarities (Freund et al., eds.), Prog. Brain Res. 64, 313-323.
9. Eckmiller, R., & Mackeben, M. (1978) Pursuit eye movements and their neural control. Pflügers Arch. Eur. J. Physiol. 377, 15-23.
10. Feldman, J.A., & Ballard, D.H. (1982) Connectionist models and their properties. Cognitive Sci. 6, 205-254.
11. Grillner, S., & Wallén, P. (1985) Central pattern generators for locomotion, with special reference to vertebrates. Ann. Rev. Neurosci. 8, 233-261.
12. Herman, R.M., Grillner, S., Stein, P.S.G., & Stuart, D.G. (eds.) (1976) Neural control of locomotion. Plenum: New York.
13. Miles, F.A., & Evarts, E.V. (1979) Concepts of motor organization. Ann. Rev. Psychol. 30, 327-362.
14. Miller, S., & Scott, P.D. (1979) The spinal locomotor generator. Exp. Brain Res. 30, 387-403.
15. Robinson, D.A. (1981) Control of eye movements. In: Hdb. Physiology see ref. 2, ch. 28, pp. 1275-1320.
16. Tatton W.G., & Bruce I.C. (1981) Comment: A scheme for the interactions between motor programs and sensory input. Can. J. Physiol. Pharmacol. 59, 691-699.
17. Westheimer, G. (1954) Eye movement responses to a horizontally moving visual stimulus. Arch Ophthalmol. 52, 932-941.

Three Constructive Algorithms For Network Learning

Stephen I. Gallant

College of Computer Science
Northeastern University

ABSTRACT: *Machine learning methods for connectionist models usually operate by attaching weights to a prespecified network so that a certain functionality is achieved. This is the classical credit assignment problem.*

This paper explores a constructive approach to connectionist learning where both a network and weights must be generated. It is argued that this is an easier problem to solve and is sufficient for many applications since network topology is usually not as important as functionality.

Three algorithms are presented for constructing networks from training examples. As cells are added and iterations are made, each method produces a network having optimal expected behavior (i.e. it correctly classifies the maximum number of training examples possible) with arbitrarily high probability $p < 1$.

Learning speed for these algorithms is currently being investigated.

Keywords: Learning, Connectionist Models, Pocket Algorithm

Acknowledgment: This research was partially supported by a grant from the Northeastern University Research and Scholarship Development Fund. Thanks to Mark Frydenberg, Mitch Wand, Bob Futrelle and Ken Baclawski for helpful comments.

I. Introduction

A common single cell connectionist model is the linear discriminant [Fisher 1936, Duda & Hart 1973] or perceptron [Rosenblatt 1961]. A linear discriminant consists of a set of numerical weights W_i. An input vector V with components or "features" $< V_1, ..., V_n >$ is classified into one of two categories (True or False) according to whether the weighted sum of its components is greater than some threshold W_0 (see Figures 1, 2). Linear discriminants and perceptrons have been extensively studied and critiqued [Minsky & Papert 1969].

Linear discriminants are quite popular for practical applications such as machine vision, pattern recognition, and

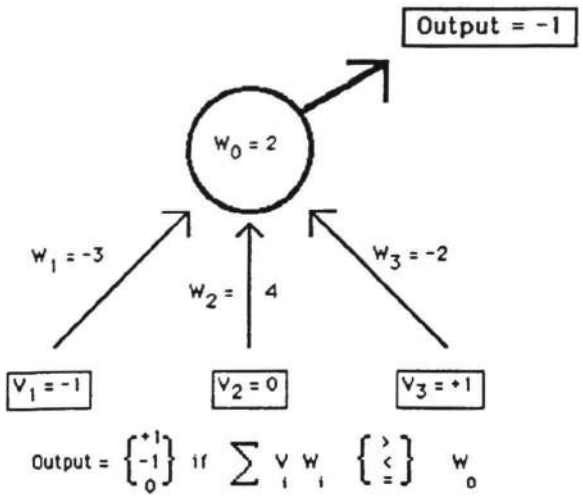


Figure 1: Linear Discriminant

Output of top cell is -1 since
$(-1)(-3) + (0)(4) + (+1)(-2) < 2$

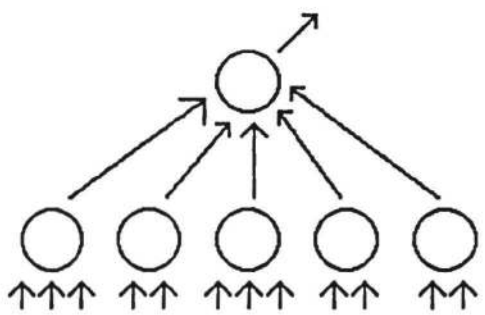


Figure 2: Linear Discriminant Network

classification problems. There are well known methods for determining good, but sub-optimal, linear discriminant weights. Recently we have developed a new method, the Pocket Algorithm, which gives optimal weights with arbitrarily high probability and appears to decrease misclassifications by roughly one-fifth over standard methods [Gallant 1985b]. The Pocket Algorithm is also important because it serves as the basis for the network growing algorithms to be discussed.

While an individual discriminant has several appealing features, it can only represent a small subset of boolean functions called separable functions (see [Gallant 1985b]). However a *network* of linear discriminants (Figure 2) can represent any boolean function on n boolean variables.

The problem of determining weights for a predetermined network of cells is much harder than for a single cell. This is the classical credit assignment problem identified by Minsky [1961]. Recent progress on this problem has been made by Barto [1981, 1985], C. Anderson [1982], Pearl [1985a, 1985b], Sutton [1984], Hinton [1984a], and Rumelhart et. al. [1985]. Of these, Rumelhart's *Back Propagation Algorithm* appears most promising with respect to speed.

II. The Constructive Approach

Rather than assuming a fixed network and trying to determine appropriate weights for it, one of our approaches has been to determine both network and weights in order to achieve the desired functionality. There are several motivations for this *constructive approach.*

First, the constructive algorithms to be studied always produce (in theory) networks and weights which correctly classify a maximum number of training examples. (If there is no pair of contradictory examples, all examples will be correctly classified.) The constructive problem seems easier to solve than the fixed problem because the necessity of realizing a solution in a fixed network is an additional constraint on the problem. Thus the requirement of defining a network is not an added burden, it is an extra degree of freedom.

Second, for most applications the topology of the final network is not vitally important. Usually it is more important to have a network and weights implementing some desired functionality than it is to have conformity with some prespecified structure. For our work with connectionist expert systems [Gallant 1985a], functionality is much more important than network topology.

Another reason to examine constructive approaches is speed of learning. The Boltzmann learning algorithm [Hinton 1984a] has largely been abandoned because of its slowness (although work with the Boltzmann model continues to be fruitful [Touretzky & Hinton 1985, Cottrell 1985]). Rumelhart's Back Propagation Algorithm appears more promising with respect to speed. Nevertheless the prospect of a fast learning method less susceptible to convergence problems gives strong impetus for exploring constructive network algorithms.

The above arguments motivate the development of constructive algorithms and comparisons with Back Propagation where appropriate.

III. Algorithms for Generating Weights for a Single Discriminant

Generating weights for a single linear discriminant so that a maximum number of training examples are correctly classified is a classical problem reviewed in [Gallant 1985b]. To quickly summarize the situation:

1. *Perceptron learning [Rosenblatt 1961, Minsky & Papert 1969] works for separable problems, but fails for nonseparable problems.*
2. *Other standard methods solve similar but different problems. Usually they produce sub-optimal weights for non-separable problems and frequently gives sub-optimal weights for separable problems as well [Duda & Hart 1973, Kaleca 1980].*
3. *The Pocket Algorithm [Gallant 1985c, 1985b] solves separable problems in finite time and gives weights for non-separable problems with an expected quality that improves as the number of iterations increases. The probability that the generated weights will be optimal approaches 1 as the number of iterations grows. Initial comparison with a standard method (Wilks' method) gave a rough estimate of a 20% decrease in misclassifications using the Pocket Algorithm.*

The algorithms for generating single discriminants are important in their own right due to their heavy use in applications such as machine vision and pattern recognition. They are also important because they are part of algorithms that generate networks.

IV. Constructive Algorithms For Network Learning

We shall briefly describe three constructive algorithms for network generation: the Tower Construction, the Inverted Pyramid Construction, and the Distributed Method. It should be noted that there are a number of additional variations to these methods which may have important effects on performance.

1. The Tower Construction

This method constructs a linear discriminant network in the form of a tower (Figure 3). Cells are added one by one

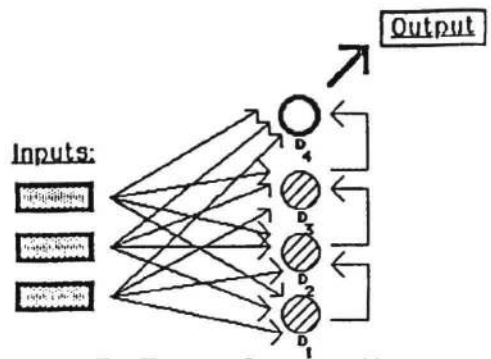


Figure 3: Tower Construction

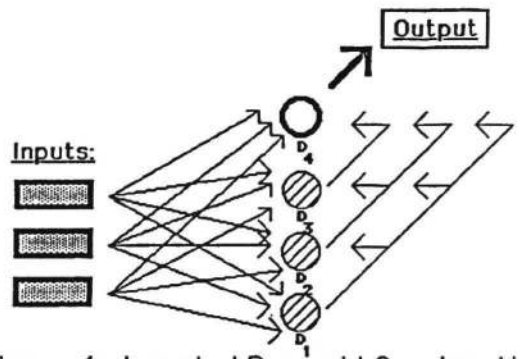
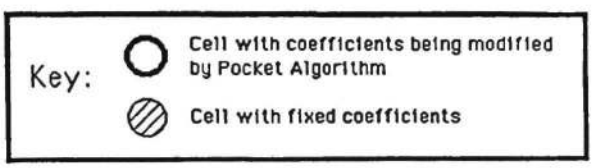


Figure 4: Inverted Pyramid Construction



and connected to the original inputs and to the previous cell as in Algorithm 1.

It can be shown that with arbitrarily high probability $p < 1$, the Tower Construction will produce a network that correctly classifies a maximum number of training examples (see [Gallant1986]). Furthermore, the output from each level will correctly classify a greater number of training examples than the output from the previous level. This guarantees theoretical convergence to a network with optimal performance for the set of training examples, as contrasted with other approaches (such as gradient descent methods) which may fail to find an optimal network.

---

1. *Construct a good (preferably optimal) set of coefficients for a single linear discriminant, D_1. If all (or the maximum number possible) of the training examples are correctly classified, we are done. Otherwise freeze the coefficients for D_1.*
2. *Construct a good (preferably optimal) set of coefficients for linear discriminant D_{n+1} having inputs from the training examples and from discriminant D_n (see Figure 3). It can be shown that D_{n+1} can correctly classify a greater number of training examples than D_n. Freeze the coefficients for D_{n+1}.*
3. *Repeat step 2 a finite number of times until all (or the maximum number possible) of training examples are correctly classified.*

Algorithm 1: *Tower Construction*

While only preliminary tests have been made with this algorithm, one of them is particularly interesting.

The Parity function is a classical non-separable function which will return true (+1) if an odd number of its inputs are +1. Otherwise the function returns false (-1). For 2 inputs the Parity function is the same as exclusive-OR. It is well known that a network of 2 cells is necessary and sufficient to represent exclusive-OR (see for example [Hinton 1984a]). We applied the Tower Construction to the Parity problem for n=3 with the expectation that a 3 level network would be

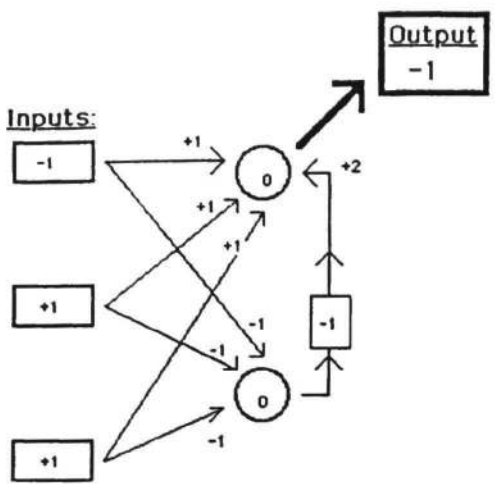


Figure 5: Tower Construction for computing parity function with three inputs.

Output is +1 if an odd number of inputs are +1. Output is -1 if an even number of inputs are +1. Note that all inputs and outputs are +1 and -1 (rather than +1 and 0).

Example shows an output of -1 when two inputs are +1. The output from the lower cell is also -1 for this example.

required to represent this function. To our surprise, the Tower Construction produced a 2 level network! The network is given in Figure 5. From this construction it became clear that a tower could compute the Parity function on n inputs by an

$$\lceil (n+1)/2 \rceil$$

level tower.

2. The Inverted Pyramid Construction

The Inverted Pyramid Construction (Figure 4) is similar to the Tower Construction, except each new discriminant sees outputs from *all* previous levels, not just the immediately preceding level. One would expect that fewer levels would be needed and this would speed up network generation. However learning speed may be adversely affected, since each level must solve a problem in a space of higher dimensions than is the case with the Tower Construction. This is currently under investigation.

3. The Distributed Method

The Distributed Method is a constructive approach which may also be viewed as a fixed network approach. The basic idea of the Distributed Method is illustrated in Figure 6. A layer of many discriminants with fixed, random coefficients is specified, after which learning takes place for the single cell on top. If enough cells have been added then any function will become separable [Minsky & Papert 1969], but families of some functions may require an exponential number of added cells for a perfect representation. On the other hand our initial testing has indicated that n training examples can usually be correctly classified by a random layer with $\frac{1}{2}n$ to $2n$ added cells. The added cells produce a distributed representation for any input [Hinton 1984b, Touretzky & Hinton 1985, Bloom 1970] since the *pattern* of cells in the added layer is more important than any single cell.

An important point here is that random linear discriminants make up the added layer rather than random boolean functions. This is intended to promote robustness of the resulting system, since an input V and a close training example V^* should both produce similar patterns in the added layer. This in turn would make it likely that the top cell would give identical outputs for V and V^*. However if random functions had been used for the added layer, V and V^* would produce to-

tally uncorrelated outputs regardless of their similarity (unless V and V^* were identical). The price to be paid for this attempt at robustness is that the added layer will require more cells to achieve separability than if strictly random boolean functions had been employed.

A major advantage of the Distributed Method is that *the same added layer can be used equally well by different top cells for computing totally different functions.* Once we invest in an added layer, it is available for every future task. Such an architecture is well suited for cells which are independent and asynchronous and it requires fewer coefficients to be learned than most other models. This structure is an old but appealing idea motivated by neural systems and the difficulty of passing large amounts of information genetically. Since we now have at least one reasonable algorithm which works with non-separable data (the Pocket Algorithm) and the ability to enhance functionality by employing the Tower or Inverted Pyramid Constructions above the added layer of fixed random cells, this idea is worth exploring again.

Space limitations preclude a more thorough examination of the Distributed Method here. The reader is referred to [Gallant 1986] for additional details.

V. Discussion and Summary

We have presented three constructive algorithms for generating connectionist networks from training examples. The Tower and Inverted Pyramid Constructions give optimal performance, after a finite number of levels have been added and after sufficient iterations, with arbitrarily high probability. Each level performs better than the previous level in terms of correct classifications.

The Distributed Method also gives optimal performance for a set of training examples with arbitrarily high probability, provided there are sufficient cells in the added layer.

All three methods are robust in the sense that an example which is sufficiently similar to a training example will be classified the same as that training example, due to the general behavior of linear discriminants.

The performance of these algorithms is just beginning to be investigated. However the Tower Construction has been incorporated into the network generation phase for the MACIE system for generating expert systems [Gallant 1985a] and has demonstrated reasonable speed and performance for several problems. More testing and practical experience is needed to fully evaluate these algorithms.

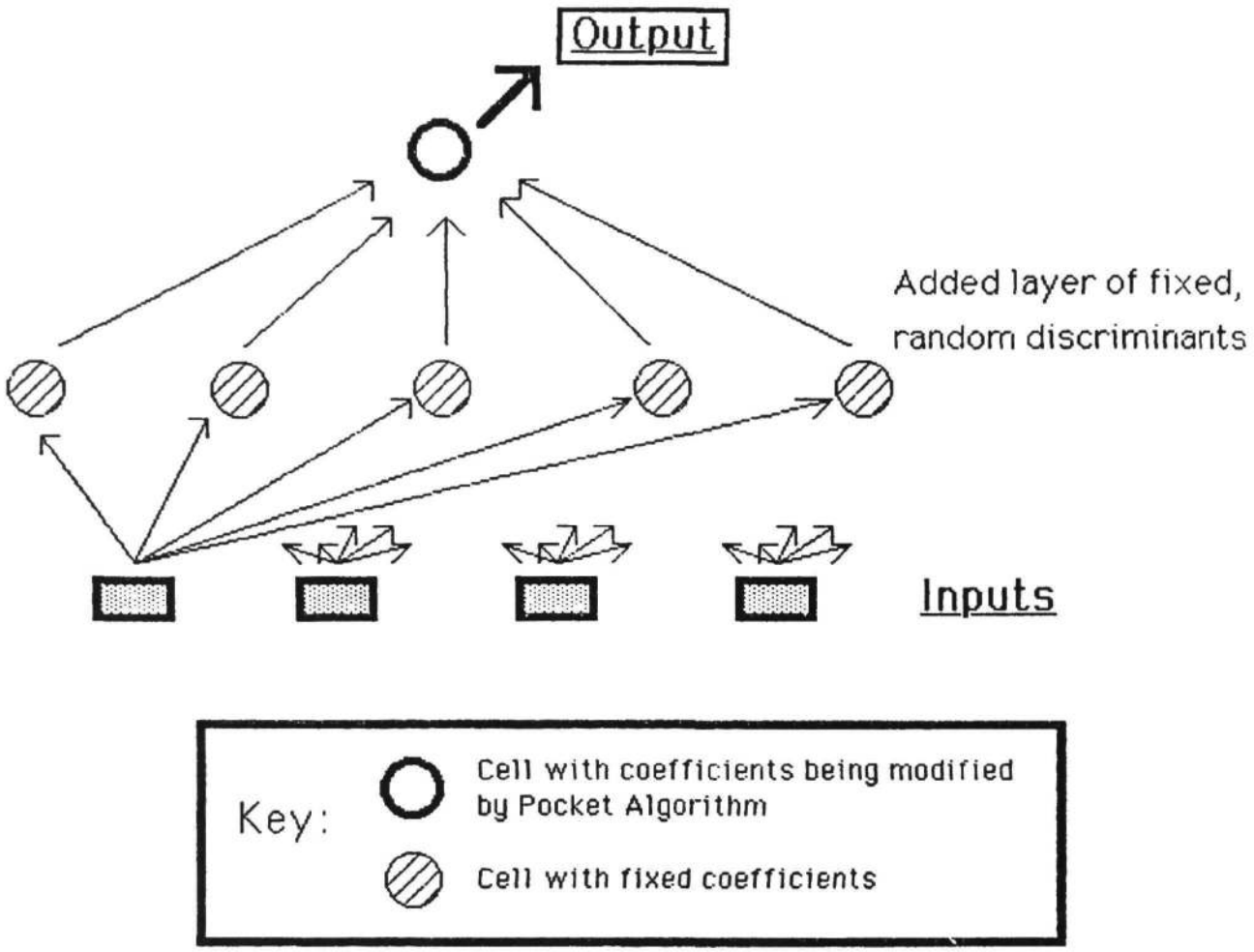


Figure 6: Distributed Method

REFERENCES

[Anderson 1982]
Anderson, C. W. Feature Generation and Selection by a Layered Network of Reinforcement Learning Elements: Some Initial Experiments. Computer and Information Science Department Technical Report 82-12, University of Massachusetts, Amherst, Massachusetts

[Barto 1981]
Barto, A. G., Sutton, R. S., & Brouwer, P. S. Associative Search Network: A Reinforcement Learning Associative Memory. Biol. Cybern. 40, 201-211 (1981)

[Barto 1985]
Barto, A. F. Learning by Statistical Cooperation of Self- Interested Neuron-like Computing Elements. Human Neurophysiology, to appear.

[Bloom 1970]
Bloom, B. H. Space/Time Trade-offs in Hash Coding with Allowable Errors. CACM 13, No. 7, July 1970, 422-426.

[Cottrell 1985]
Cottrell, G. W. Connectionist Parsing. Seventh Annual Conference of the Cognitive Science Society, Irvine, Ca. 1985.

[Duda & Hart 1973]
Duda, R. O. & Hart, P. E. Pattern Classification and Scene Analysis. (1973) John Wiley & Sons, New York.

[Fisher 1936]
Fisher, R. A. The use of multiple measurements in taxonomic problems. Ann. Eugenics, 7, Part II, 179-188. Also in *Contributions to Mathematical Statistics* (1950) John Wiley, New York.

[Gallant 1985a]
Gallant, S. I. Automatic Generation of Expert Systems From Examples. Proceedings of Second International Conference on Artificial Intelligence Applications, sponsored by IEEE Computer Society, Miami Beach, Florida, Dec. 11-13, 1985.

[Gallant 1985b]
Gallant, S. I. Optimal Linear Discriminants. Technical Report SG-85-30, Northeastern University College of Computer Science. (to appear, Eighth International Conference on Pattern Recognition, Paris, France, Oct. 28-31, 1986)

[Gallant 1985c]
Gallant, S. I. The Pocket Algorithm for Perceptron Learning. Technical Report SG-85-19, Northeastern University College of Computer Science.

[Gallant 1986]
Gallant, S. I. Three constructive Algorithms for Network Learning. Technical Report SG-86-37, Northeastern University College of Computer Science. (note: the previous working title for this paper was "Credit Assignment, Intermediate States, and the Power of Towers.")

[Hinton 1984a]
Hinton, G. E., Sejnowski, T. J., & Ackley, D. H. Boltzmann Machines: Constraint Satisfaction Networks that Learn. Technical Report CMU-CS-84-119, Carnegie-Mellon University Department of Computer Science

[Hinton 1984b]
Hinton, G. E. Distributed Representations. Technical Report CMU-CS-84-157, Carnegie-Mellon University, Department of Computer Science

[Minsky 1961]
Minsky, M. Steps Toward Artificial Intelligence. Proceedings IRE, 1961, 49, 8-30. Reprinted in Feigenbaum, E. A. & Feldman, J., Eds., *Computers and Thought* McGraw- Hill, New York, 1963.

[Minsky & Papert 1969]
Minsky, M. & Papert, S. *Perceptrons: An Introduction to Computational Geometry* (1969) MIT Press, Cambridge, Ma.

[Nilsson 1965]
Nilsson, N. J. *Learning Machines* (1965) McGraw-Hill, New York, NY.

[Pearl 1985a]
Pearl, J. Fusion, Propagation, and Structuring in Bayesian Networks. UCLA Computer Science Department Technical Report CSD-850022, R-42, June 1985.

[Pearl 1985b]
Pearl, J. How to Do with Probabilities What People Say You Can't. Proceedings of Second International Conference on Artificial Intelligence Applications, sponsored by IEEE Computer Society, Miami Beach, Florida, Dec. 11-13, 1985.

[Rosenblatt 1961]
Rosenblatt, F. *Principles or neurodynamics: Perceptrons and the theory of brain mechanisms* Spartan Press, Washington, DC.

[Rumelhart 1985]
Rumelhart, D. E., Hinton, G. E., & Williams, R. J. Learning Internal Representations by Error Propagation. ICS Report 8506, Institute for Cognitive Science, University of California, San Diego, Sept. 1985.

[Sutton 1984]
Sutton, R. S. Temporal Credit Assignment in Reinforcement Learning. Computer and Information Science Technical Report 84-02, University of Massachusetts, Amherst, Ma. (1984)

[Touretzky & Hinton 1985]
Touretzky, D. & Hinton, G. Symbols Among the Neurons: Details of a Connectionist Inference Architecture. Ninth International Joint Conference on Artificial Intelligence, Los Angeles, Ca. (IJCAI 85)

A NEURAL SIMULATION OF CLASSICAL CONDITIONING IN APLYSIA

Howard N. Henry
Information and Computer Science Department
University of California — Irvine

Psychological research into behavioral conditioning of animals is some of the most advanced work on learning available. Many biologists studying learning, on the other hand, have focused on the processes causing changes in the strength of the synaptic connections between neurons, and their research is most advanced in invertebrates such as *Aplysia.*

Relating the observable behavior and the cellular processes would help the overall understanding of learning. Recently, the cellular mechanisms for habituation, sensitization, and classical conditioning have been postulated for the defensive withdrawal reflex in *Aplysia* (Hawkins & Kandel, 1984).

It would be useful to test these mechanisms to establish the constraining limits on the learning they control. Simulating a model of neural mechanisms is ideally suited to the task, since it allows a wide range of controlled internal configurations to be tested under a wide range of controlled external conditions in a short time.

Already, a neural simulation of this reflex has identified a problem in producing "blocking", in which an established association between stimuli prevents formation of an association with a new stimulus (Gluck & Thompson, 1986). Another neural model has shown the implications of blocking and conditioned inhibition on animal processing of stimuli during classical conditioning experiments (Barto & Sutton, 1985).

This paper explores learning by a reflex in *Aplysia* in response to variations in *contingency* in classical conditioning presentations, using a neural simulation model. The main finding is that the learning is extremely sensitive to even very small changes in parameters such as the rate of decrement of synaptic strength with habituation.

THE DEFENSIVE WITHDRAWAL REFLEX IN APLYSIA

In mollusks such as *Aplysia,* a respiratory chamber *(mantle cavity)* contains the gill and is covered by a protective sheet *(mantle shelf)* that terminates in a fleshy spout *(siphon).* The gill, mantle shelf, and siphon all contract vigorously and withdraw into the mantle cavity when the siphon or mantle shelf is stimulated by touch.

Three forms of learning have been demonstrated in this reflex: habituation, sensitization, and classical conditioning (Hawkins & Kandel, 1984). Habituation involves a depression of neurotransmitter release at the synapses that the sensory neurons make on the motor neurons and interneurons. Sensitization is caused by an increase in neurotransmitter release at the same synapses, produced by a facilitator neuron that terminates on or near these synapses; stimulating the tail initiates this *presynaptic facilitation.*

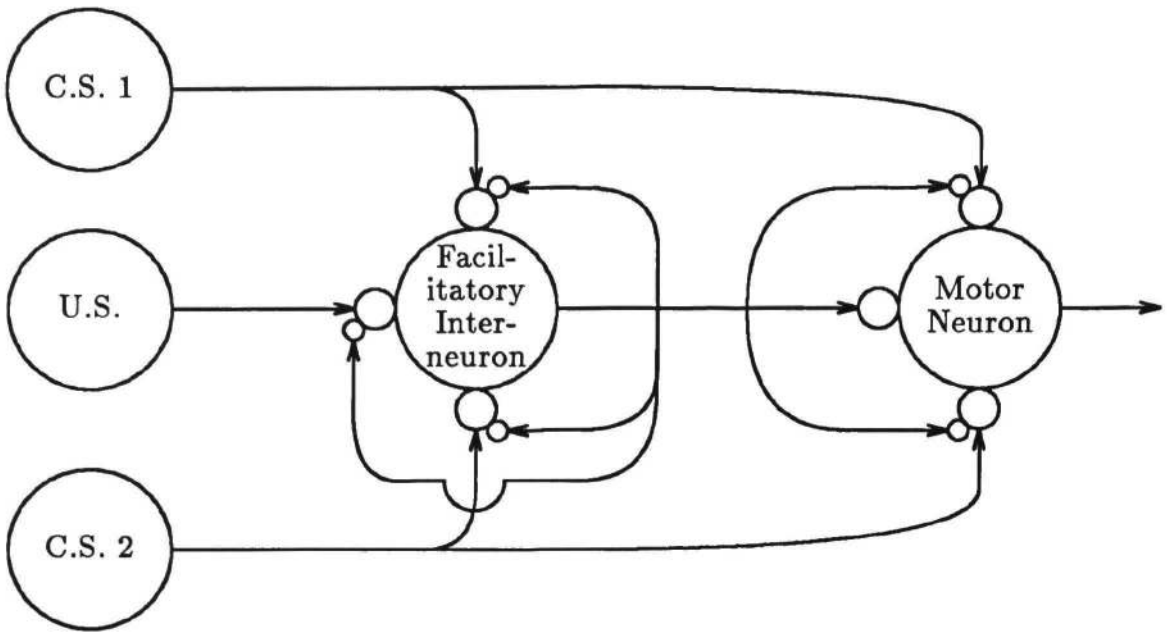


Figure 1. Simulated defensive withdrawal reflex circuit, after Hawkins & Kandel (1984).

In classical conditioning, the sensory neurons of the CS path fire just before the facilitator neuron of the US pathway becomes active. The facilitator neuron produces much more enhancement of a recently-active synapse, an effect termed an *activity-dependent amplification of presynaptic facilitation.* This enhancement is maximized when the CS precedes the onset of the US by 0.5 seconds.

The neurons, synapses, and presynapses involved in learning by this reflex are illustrated in the circuit of Figure 1 (Hawkins & Kandel, 1984). This circuit is a simplification of the actual neuroanatomy (Carew et al., 1984; Hawkins et al., 1981a), but is sufficient to demonstrate the learning in the neural simulation.

THE NEURAL SIMULATION MODEL

The input to the neural simulation model includes a description of the circuit to be simulated in terms of its neurons, synapses, and presynapses, followed by the timing of every external stimulation of a neuron for the course of the simulation. The system outputs the firing frequency of each neuron and the weight of each synapse at specified times.

Every quarter-second simulation cycle, the simulation checks for frequency changes in the input file. Then it computes the firing frequency for each neuron from the frequencies of its input neurons in the preceding cycle and the weights of the synapses between the input neurons and this neuron. Finally it updates the weight of each synapse for four possible impacts:

1. Habituation
2. Recovery from habituation
3. Sensitization and conditioning raise the weight of a synapse that has a presynapse from a rapidly-firing facilitator neuron.

4. Conditioning additionally requires recent firing of the neuron leading to the synapse. This recent activity is maintained between cycles in a *synaptic plasticity* value for each synapse.

LEARNING IN ANIMALS

A classical conditioning experiment may be described in terms of four distinct "contingency conditions" or "presentation conditions", each representing some combination of CS and US presentations (Granger & Schlimmer, 1986):

1. Perfect Pairings, containing CS - US pairs and *non-events* (having neither CS nor US);
2. Partial Reinforcement, combining CS - US pairs, spurious CS — events, and non-events;
3. Partial Warning, with CS - US pairs, spurious — US events, and non-events; and
4. Composite, having all four types of events.

All of these conditions also implicitly contain *context cues,* which are stimuli other than the CS and the US that remain constant throughout the trials.

From experiments covering various presentation conditions, Rescorla (1968) formulated a precise constraint describing those conditions that enable or prevent classical conditioning in mammals: for excitatory conditioning to occur, $P(US|CS) > P(US|\overline{CS})$. In terms of the presentation conditions, this constraint means that there should be no significant excitatory conditioning for the Composite condition, much learning for the Perfect Pairings case, and some positive conditioning for the Partial Reinforcement and Partial Warning contingencies (Granger & Schlimmer, 1986). In other words, the four presentations should be ordered as follows:

$$\textit{Perfect Pairings} > \textit{Partial Warning} \approx \textit{Partial Reinforcement} > 0 \approx \textit{Composite}.$$

Rescorla's contingency constraint is still widely held by researchers in the field of conditioning (e.g., Colwill & Rescorla, 1986).

SIMULATIONS

The simulations generally produce an increase in the strength of the synapse connecting the CS+ neuron to the motor neuron that causes the withdrawal action of the reflex. One measure of this change in strength is the *relative associative strength,* given by $(w-0.5)/0.5$, where w is the final synaptic strength and 0.5 is the initial synaptic strength.

The first simulation modeled a frequently-cited *differential conditioning* experiment in *Aplysia.* The withdrawal reflex in *Aplysia* can be differentially conditioned by pairing a tail shock reinforcing US with stimulation of either the siphon or mantle shelf, and leaving stimulation of the other unreinforced. Then stimulation of the site paired with the US (the CS+) produces greater response than the unpaired location (the CS-). In the experiment to be simulated (Carew et al., 1984; Hawkins et al., 1983; Hawkins & Kandel, 1984),

Table 1

% of Trials Spurious:	0 %	20 %	40 %	60 %	80 %	Avg. Rank	Constraint
Perfect Pairings	+0.44	—	—	—	—	1	1
Partial Warning	—	+0.22	+0.22	+0.16	+0.16	2-3 tied	2-3 tied
Partial Reinforcement	—	+0.32	+0.22	+0.14	+0.04	2-3 tied	2-3 tied
Composite	—	+0.28	+0.16	+0.04	0.00	4	4

one group of 20 *Aplysia* were given five trials of normal differential conditioning with a five-minute intertrial interval, while a second group of 20 were given the same training but with five additional US's inserted between the paired trials. The first group showed significant learning when tested, while the second group did not. Unlike the experiment but like the predictions of Rescorla's contingency constraint, the simulated second group did learn the association between CS+ and the US, though not as well as the first group without spurious US's; the relative associative strength of the first (Perfect Pairings) group was +0.44, while that of the second (20 percent Partial Warning) group was +0.22. (The presentation condition percentage refers to the proportion of spurious, unpaired trials; both figures assume a Habituation Rate equivalent to 50 percent per trial, a rate that will be tested later.)

The second set of simulations attempted to determine whether the learning by the circuit in this experimental situation conformed to Rescorla's contingency constraint. To test this required simulations over a range (0, 20, 40, 60, and 80 percent spurious trials, totaling 13 cases) of the four presentation conditions to identify the relative amount of learning under the contingencies. If learning adhered to the constraint, then performance under the four conditions would follow the prescribed relationship given earlier:

$$\textit{Perfect Pairings} > \textit{Partial Warning} \approx \textit{Partial Reinforcement} > 0 \approx \textit{Composite}.$$

The relative associative strengths for the 13 cases (at the assumed 50 percent Habituation Rate) are roughly consistent with this ordering, as shown in Table 1. Ranks of the results of the four sets of presentation conditions (1 = Highest) are also compared to the ordering of Rescorla's contingency constraint in Table 1.

The third set of simulations systematically explored the effect of variations in the Habituation Rate on the results. An analysis of the differential conditioning experiment assumed habituation at a rate equivalent to a 50 percent decrease with each US presentation (Hawkins & Kandel, 1984, p. 387). But neurophysiological measurements of habituation in this circuit's synapses average 3.9 percent per 0.1 Hz. neuron firing (Hawkins et al., 1981b), equivalent to a 25 percent decrease with each US presentation for this experiment. So in addition to the results reported at the 50 percent Habituation Rate, Figure 2 gives

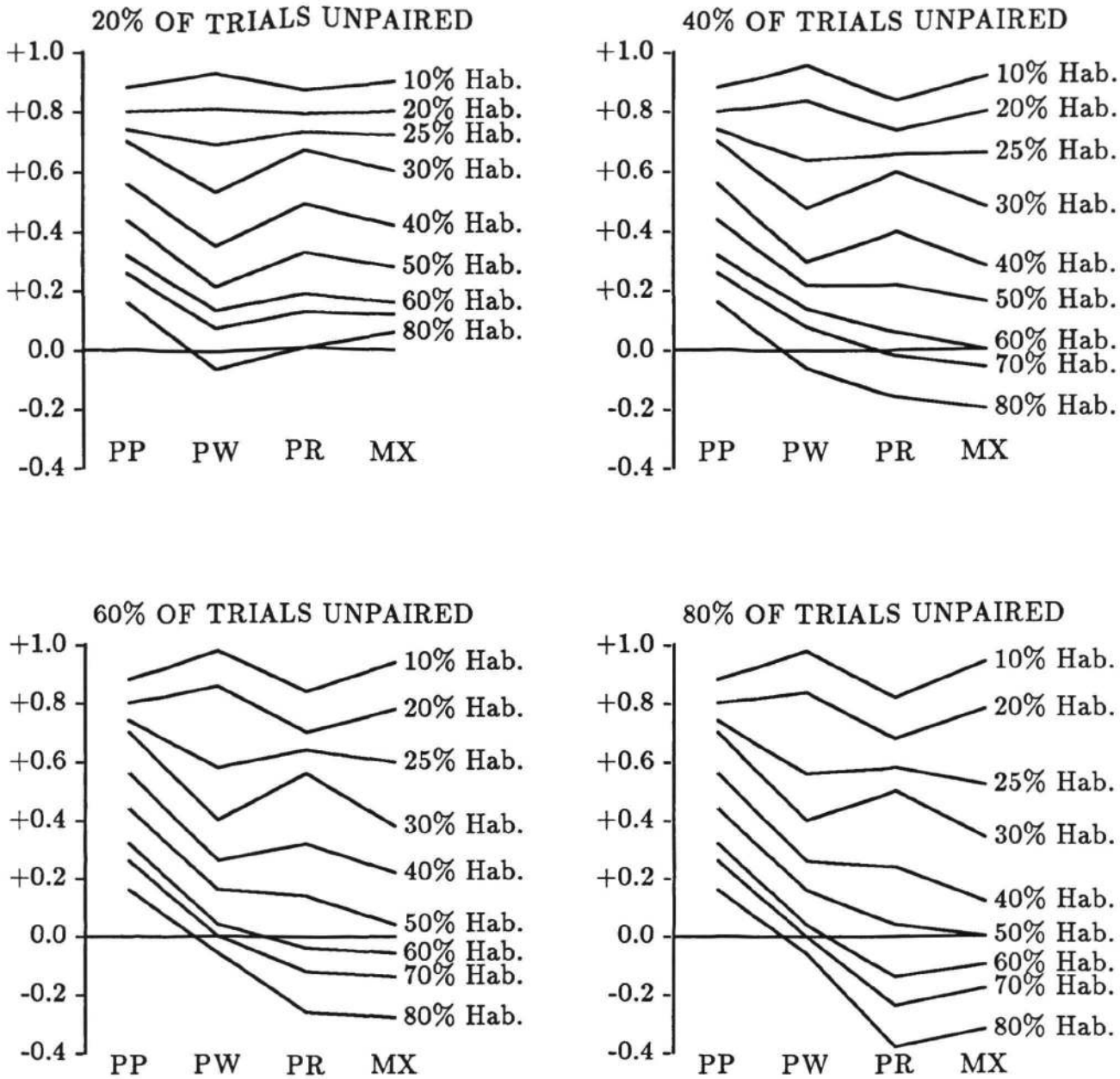


Figure 2. Each graph gives the relative associative strengths for one of the four percentages of spurious trials. Each graph contains nine curves of four points each, one curve per Habituation Rate. The four values on each curve relate to the four presentation conditions, which are situated beside each other across a graph, and are of equal levels of intensity (percentage of spurious trials). The Perfect Pairings (PP) condition is at the left, the Composite (MX) condition is at the right, and the Partial Warning (PW) and Partial Reinforcement (PR) conditions are in between. This ordering across the graph is the same as that required for Rescorla's contingency constraint, so adherence to the constraint is indicated by a curve's downward slope. (The slope of the curve between the two middle points need not fall, since these two presentation conditions should be approximately equivalent.) Conformity with the constraint also requires that the Composite value be near or below zero, and that the Partial Warning and Partial Reinforcement points be slightly positive.

Table 2

Habituation Rate	Perfect Pairings	Partial Warning	Partial Reinforcement	Composite
10	3	1	4	2
20	2	1	4	3
25	1	3-4 tied	2	3-4 tied
30	1	3-4 tied	2	3-4 tied
40	1	3	2	4
50	1	2-3 tied	2-3 tied	4
60	1	2	3-4 tied	3-4 tied
70	1	2	3-4 tied	3-4 tied
80	1	2	4	3
Constraint Prediction	1	2-3 tied	2-3 tied	4

results for the 25 percent rate and several intermediate and nearby rates so that the sensitivity of this parameter setting can be gauged. These results are summarized by Table 2's rankings of the four presentation conditions for each of the nine Habituation Rates.

DISCUSSION

These results indicate that the simulated performance of the circuit is highly sensitive to the Habituation Rate. Learning is too easy under too many conditions at low Habituation Rates, and too hard under too many conditions at high Habituation Rates. Conditioning conforms to Rescorla's contingency constraint only at a Habituation Rate approximating a 50 percent decrease per US presentation. But as indicated earlier, habituation recorded in cells of the reflex in *Aplysia* averages the equivalent of a 25 percent rate. There are four possible explanations for this discrepancy:

1. It is possible that the combination of parameter settings for the simulation is not appropriate for the experimental condition being simulated. There is likely to be an interaction among the learning rates and the timing and intensities (firing frequencies) of the stimuli in the experimental situation. While all these settings have been obtained from experimental measurements and expert assumptions discussed earlier, the various values were cast in several distinct contexts, instead of a single unified experiment. Thus each individual value may be valid for its situation, but together they may be incompatible for the experiment being simulated. Neurophysiological measurements under unifying experimental conditions would help test this possibility, as would neural simulations that systematically vary the values of each parameter to test interaction effects.

2. It is possible that the neural simulator incorporates assumptions that do not accurately reflect the processing of the neurons in the circuit.

3. It is possible that the small circuit isolated for the simulation is not sufficient for reproducing the quantitative contingency results of classical conditioning, though it is enough for demonstrating some of the more qualitative effects (e.g., Gluck & Thompson, 1986). In that case, one of the more elaborate circuits identified for the defensive withdrawal reflex in *Aplysia* (Hawkins et al., 1981a; Carew et al., 1984) may be adequate.
4. The Rescorla contingency constraint is based upon mammalian conditioning results. It is possible that learning in *Aplysia* does not adhere to the same constraints.

Further work is needed to determine which of these four possibilities best explains the discrepancy between those Habituation Rates producing learning under Rescorla's conditions and those rates that have been measured in *Aplysia*.

CONCLUSIONS

These simulations have attempted to establish whether this circuit is capable of mammalian conditioning, by testing its adherence to Rescorla's contingency constraint. The simulations have shown the sensitivity of learning to neurophysiological measures, particularly the Habituation Rate. It might be that synaptic weight change rates and neural firing frequencies will need to be completely specified before the performance of a circuit can be fully determined.

It appears that neural simulations would benefit from more extensive and detailed neurophysiological measurements for setting the values of the synaptic strength learning rates, and the *in vivo* firing frequencies resulting from stimulation by CS, US, and context stimuli. Making these various measurements under unified experimental conditions would increase the confidence in predictions made from simulating their interactions.

In addition, it would be useful to use neural simulation to systematically explore the space of joint parameter values. This would permit identification of those combinations that replicated physical experiments and those that produced results consistent with the mammalian contingency constraint. Finding the parameters most sensitive in determining performance would help prioritize the neurophysiological measurements needed.

ACKNOWLEDGMENTS

This research was supported in part by the Office of Naval Research under grants N00014-84-K-0391 and N00014-85-K-0854, by the Army Research Institute under contract MDA903-85-C-0324, and by the National Science Foundation under grants IST-81-20685 and IST-85-12419.

REFERENCES

Barto, A.G. & Sutton, R.S. (1985) Neural problem solving. In *Synaptic Modification, Neuron Selectivity, and Nervous System Organization.* Levy, W.B., Anderson, J.A., & Lehmkuhle, S. (Eds.), Lawrence Erlbaum Associates, Hillsdale, N.J.

Carew, T.J., Abrams, T.W., Hawkins, R.D., & Kandel, E.R. (1984) The use of simple invertebrate systems to explore psychological issues related to associative learning. In

Primary Neural Substrates of Learning and Behavioral Change. Alkon, D.L. & Farley, J. (Eds.), Cambridge University Press, Cambridge.

Colwill, R.M. & Rescorla, R.A. (1986) Associative structures in instrumental learning. *The Psychology of Learning and Motivation,* **20**, in press.

Gluck, M.A. & Thompson, R.F. (1986) Modeling the neural substrates of associative learning and memory: A computational approach. *Psychological Review,* in press.

Granger, R.H. & Schlimmer, J.C. (1986) The computation of contingency in classical conditioning. *The Psychology of Learning and Motivation,* **20**, in press.

Hawkins, R.D., Carew, T.J., & Kandel, E.R. (1983) Effects of interstimulus interval and contingency on classical conditioning in *Aplysia. Society for Neuroscience Abstracts,* **9**, 168.

Hawkins, R.D., Castellucci, V.F., & Kandel, E.R. (1981a) Interneurons involved in mediation and modulation of gill-withdrawal reflex in *Aplysia.* I. Identification and characterization. *Journal of Neurophysiology,* **45**, 304-314.

Hawkins, R.D., Castellucci, V.F., & Kandel, E.R. (1981b) Interneurons involved in mediation and modulation of gill-withdrawal reflex in *Aplysia.* II. Identified neurons produce heterosynaptic facilitation contributing to behavioral sensitization. *Journal of Neurophysiology,* **45**, 315-326.

Hawkins, R.D. & Kandel, E.R. (1984) Is there a cell-biological alphabet for simple forms of learning? *Psychological Review,* **91**, 375-391.

Kupfermann, I. (1985) Learning. In *Principles of Neural Science: Second Edition.* Kandel, E.R. & Schwartz, J.H. (Eds.), Elsevier, New York.

Rescorla, R. (1968) Probability of shock in the presence and absence of CS in fear conditioning. *Journal of Comparative and Physiological Psychology.* **66**, 1-5.

Shepherd, G.M. (1983) *Neurobiology.* Oxford University Press, New York, 545-561.

PUTTING AFFECT INTO TEXT

Eduard H. Hovy
Yale University
Computer Science Department [1]

Abstract

How is affect communicated in language? Natural languages contain a large number of techniques for injecting affect into text, both explicitly and implicitly. This paper discusses some techniques that speakers use to slant their text, and describes what is required for a computer program to generate differently slanted versions of a single underlying representation.

1 Introduction

Affect — the speaker's sympathies and antipathies — can be communicated in language both explicitly and implicitly. Since an explicit statement of the speaker's opinion (such as "I don't like your shoes") may alienate the hearer, the speaker should skirt sensitive issues and achieve effects indirectly. Most languages have a large body of techniques for doing so. What are these techniques? This paper briefly describes how the program PAULINE (Planning And Uttering Language In Natural Environments) is able to produce differently slanted texts from a single underlying representation using affect.

PAULINE generates the examples used in this paper from the interpretations produced by the JUDGE program ([Bain 86]), a case-based expert system that models the sentencing behaviour of a judge. As input, JUDGE accepts the representation of a fight — a set of actions and resulting states — and as output to PAULINE it produces a set of interpretations. Each interpretation describes an action, its justifiability, and the culpability of the actor. Saying only the actions, a typical fight is:

```
FIRST, JIM BUMPED MIKE ONCE AND HURT HIM. THEN MIKE SMACKED JIM, HURTING
HIM. NEXT, JIM HIT MIKE ONCE. THE RESULT WAS THAT HE KNOCKED HIM DOWN.
AFTER THAT, MIKE SMACKED JIM SEVERAL TIMES AND KNOCKED HIM DOWN. JIM
SLAPPED MIKE SEVERAL TIMES. THE RESULT WAS THAT HE HURT HIM. AFTER THAT,
MIKE STABBED JIM. AS A RESULT, JIM DIED.
```

Clearly, in any real account of the fight, Jim's version will to differ appreciably from Mike's. Each speaker will make the decisions that slant the text in his favour. PAULINE makes affect-related decisions during each of its three stages: topic collection, topic organization into phrases and sentences, and topic realization into text.

[1]This work was supported in part by the Advanced Research Projects Agency monitored by the Office of Naval Research under contract N00014-82-K-0149.

2 Computing Affect

In order to communicate his opinions implicitly by slanting the text, the speaker must be able to determine which aspects of it he is sympathetic to, which aspects he dislikes, and which he does not care about. PAULINE uses three values of affect: GOOD, BAD, and NEUTRAL. Here, *affect* simply denotes something akin to "like". (Three values are sufficient to give the program interesting behaviour. In this regard it is similar to the narrative summarization work in [Lehnert 82].)

PAULINE gets its affects from two sources: The first source is the user, who lists one or more representation elements as *sympathies* or as *antipathies*. (For example, when PAULINE is to defend Mike, the concept "Mike" is GOOD and the concept "Jim" is BAD.) The second source is the intrinsic affect associated with generic representation types. (For example, in neutral context in the JUDGE domain, the concepts "hit" and "die" are BAD, the concept "unintentionally" is GOOD, and all other concepts, such as "Jim" and "Mike", are NEUTRAL. Similar information is used by the JUDGE program to determine its interpretation of each action.)

In order to determine its opinion about any arbitrary piece of input representation, PAULINE needs the ability to combine its affects with the intrinsic affects and to propagate the results along the relations between concepts. Though their exact form depends on the design of the representation, the **basic rules of affect propagation** are:

1. **affect is preserved when combined with NEUTRAL**
2. **like affects combine to GOOD**
3. **unlike affects combine to BAD**
4. **the combined affect inverts for certain relations between affect-bearing concepts, (e.g., the conceptual *patient* of a BAD act). This requires special rules**

This works as follows: assume the current topic is the action ACT-6:

```
#{ACTION-UNIT  =  ACT-6
      [ACT :  HIT]
      [ACTOR :  MIKE]
      [TO :  JIM]
      [FORESEEABILITY :
            #{STATE  =  STATE-10
                [TYPE :  PHYSICAL-INTEGRITY-VIOLATION]
                [ACTOR :  JIM]
                [DEGREE :  SERIOUS-TEMPORARY]}]
      [NUMBER :  SINGLE]
      [DEGREE :  HARD]
      [INTENTIONALITY :  PRESENT]
      [RESULT :  #{STATE  =  STATE-11
                    [TYPE :  PHYSICAL-INTEGRITY-VIOLATION]
                    [ACTOR :  JIM]
                    [DEGREE :  KNOCK-DOWN]}]}
```

Figure 1

(this is a slightly modified and pruned version of the actual JUDGE representation). Stated neutrally, ACT-6 reads

(1) MIKE INTENTIONALLY HIT JIM HARD ONCE AND KNOCKED HIM DOWN

Assume PAULINE is sympathetic to Mike. In order to determine its opinion of ACT-6, the program combines the intrinsic affect for the type of ACT-6, HIT, (BAD) with its affect for Mike (GOOD) (noting that, since Mike is the ACTOR, the affect doesn't invert), to get the affect BAD. That is to say, in ACT-6, Mike looks bad and, when speaking, the action should be omitted if possible or mitigated if not. One of the ways PAULINE could say this is:

(2) MIKE JUST TAPPED JIM ONCE

If, on the other hand, PAULINE's sympathies are for Jim, it combines the affect for HIT (BAD) with its affect for Jim (GOOD), giving BAD. Since Jim fills the role TO (the conceptual role *patient*), this result must be inverted, and so the final affect is GOOD. That is to say, ACT-6 is GOOD for the case against Mike. In this case PAULINE should enhance the topic. Furthermore, the RESULTant state, STATE-11, carries the same affect, because Jim (GOOD) suffers a PHYSICAL-INTEGRITY-VIOLATION (intrinsically BAD). This result was intentionally (INTENTIONALITY PRESENT) caused by Mike (BAD). These three affects GOOD, BAD, and BAD combine to produce GOOD for Jim in STATE-11, causing it to be enhanced too. Thus, when defending Jim, PAULINE produces sentences such as

(3) MIKE PURPOSELY SMASHED JIM AND KNOCKED HIM DOWN

3 Generating with Affect

3.1 Topic Organization into Phrases

Before it says anything, the generator performs a number of planning tasks. One task is to cast sentence topics together into phrases. In addition to the planning criteria used by other generators (such as focus in [McKeown 82] and hearer knowledge in [Cohen 78] and [Appelt 81]), PAULINE uses affect to control the juxtaposition of sentence topics, since certain phrases are very useful for conveying affect implicitly. The following phrases can be called *enhancers*:

(a) "*Not only* did Pete play the game, *but* he hit five home runs"
(b) "Pete played the game; *what's more*, he hit five home runs"

Clearly, these phrase forms imply that Pete's playing and his home runs carry the same affective value (either both GOOD or both BAD), and in fact that the value is to be strengthened due to their juxtaposition. In contrast, the following sentences carry no such cumulative affective import:

(c) "Pete played the game and he hit five home runs"
(d) "When Pete played the game he hit five home runs"

When an enhancing phrase juxtaposes two affect-laden sentences, the affect is strengthened; when it juxtaposes an affect-laden sentence with a neutral one, the affect is imputed to the latter. Thus, in addition to stressing affective concepts, a speaker can strengthen his case by imputing affect to neutral concepts too! This is, for example, what PAULINE does to produce

(4) NOT ONLY DID JIM EXPECT NO THREAT FROM MIKE ANY LONGER, BUT HE COULD FORESEE THAT HE WOULD HURT HIM IF HE REALLY SLAPPED HIM

when defending Mike. Here Jim's not perceiving a threat from Mike is simply NEUTRAL, but his ability to foresee the BAD result of his action, coupled with the fact that he did it anyway, is BAD for him. However, when juxtaposed in this way, *both* sentences seem BAD for Jim — exactly what PAULINE wants.

Similarly, *mitigators* are phrases with weakening effect. When a mitigating phrase juxtaposes two sentences carrying opposite affect, the resulting affect is that of the first sentence, weakened; when it juxtaposes an affect-laden sentence with a neutral one, the opposite affect is imputed to the latter. In the following sentences, if "John whipped the dog" carries BAD affect, then, if we know nothing more, "he remembered the cat" becomes GOOD:

(e) "*Although* John remembered the cat, he whipped the dog"
(f) "John remembered the cat. *However*, he whipped the dog"

(The first part could just as well have been made BAD by using an enhancer: "*Not only* did John remember the cat, *but* he whipped the dog".)

A number of constraints must be met before two topics can be juxtaposed in an enhancer or mitigator phrase:

1. **Two-predicate enhancer and mitigator phrases can only be used when the parts carry consistent affects; that is,**
   - **in enhancer phrases the two predicates must carry like affect**
   - **in mitigator phrases the two predicates must carry unlike affect**
   - **or else one predicate must be NEUTRAL**
2. **Two-predicate enhancer and mitigator phrases can only be used when the topics in both parts focus on the same concept**
3. **The predicates in two-predicate enhancer and mitigator phrases should match in as many aspects as possible**

3.2 Realization

Sentence subject selection: Since the sentence subject is a prominent position, it must be chosen with care. Grosz [Grosz 77], Sidner [Sidner 79], and McKeown [McKeown 82], among

others, describe rules for choosing subjects in order to produce flowing text. These rules are often underspecific; McKeown's algorithm simply picks the default (a predefined entry for each predicate) when a number of focus candidates exist with the same number of implicit links to the potential focus list. Affect can be used as an additional criterion for subject choice — either at a low level, simply to help winnow out candidates, or at a high level, to help slant the text very strongly. When PAULINE has the goal to convince the hearer or to make known its affects, it use, in addition, the following rule:

> 4. **The new focus should be selected from GOOD candidates for sentences with GOOD affect and from BAD candidates for sentences with BAD affect**

For example, when PAULINE is defending Jim, the following sentence topic is BAD for him; by this rule, Mike, an antipathy, is chosen as subject, thereby de-emphasizing Jim's role:

(5) MIKE ONLY WAS BUMPED BY JIM ONCE

Sentence Inclusion: Adverbs: Some adverbial stress words specifically function as enhancers or mitigators; in the JUDGE domain, PAULINE uses the following:

intentionality: "intentionally", "purposely" — "accidentally"
degree: "hard" — "lightly" (hit)
number: "repeatedly" — "once" (stabbed)
stress: "really" — "only", "just"

When these words are used to modify concepts that do not already carry affect, they seem strange, forcing the hearer to postulate affect; an additional constraint is:

> 5. **Adverbial stress words can only be used to enhance or mitigate expressions that carry some affect already**

During the realization of a sentence, the program collects the aspects of the topic it can legitimately say as adverbs (it cannot, for example, misrepresent the topic to say "lightly" when the DEGREE is HARD). From these, it selects at most two (the two least recently said types), since when affective adverbs are overused the effect is unnatural, to give, for example:

(6) MIKE *JUST* HIT JIM *ONCE*
(7) JIM COULD FORESEE THAT HE WOULD HURT MIKE IF HE *REALLY* SLAPPED HIM. HE SLAPPED HIM *REPEATEDLY* AND HURT HIM

Word Choice: Verbs: Often an action can be described by a number of verbs of which most carry some affect. For example, PAULINE accesses "hit" from the representation element HIT and then uses its affective values to discriminate to "tap" as a mitigator and "smash" as an enhancer (the lexicon contains discrimination nets, as in [Goldman 75]):

(8) JIM JUST *TAPPED* THAT JERK MIKE ONCE
(9) JIM PURPOSELY *SMASHED* MIKE AND KNOCKED HIM DOWN

Table 1: Defending Mike

FIRST, *THAT JERK* JIM BUMPED MIKE *HARD* AND HURT HIM. THEN MIKE *JUST TAPPED* JIM *ONCE.*	enhancing noun, stress mitigating adverbs, verb
AFTER THAT, JIM DID NOT EXPECT THAT MIKE WAS GOING TO HURT HIM ANY LONGER; *ALSO;* JIM COULD FORESEE THAT HE WOULD INJURE MIKE IF HE PURPOSELY STRUCK HIM. HE *REALLY SMASHED* HIM. THE RESULT WAS THAT HE INJURED HIM.	NEUTRAL and BAD in enhancer phrase enhancing stress enhancing verb
NEXT, MIKE HIT JIM, KNOCKING HIM DOWN.	suppress all but action
NOT ONLY DID JIM EXPECT NO THREAT FROM MIKE ANY LONGER, *BUT* HE COULD FORESEE THAT HE WOULD HURT HIM IF HE REALLY SLAPPED HIM. HE SLAPPED HIM *REPEATEDLY* AND HURT HIM.	enhancer phrase enhancing adverb
MIKE REALIZED THAT JIM *REALLY* HURT HIM; *ALSO*, HE EXPECTED THAT JIM'S GOAL WAS TO HURT HIM. JIM *WAS STABBED.*	enhancing stress, phrase mitigating passive
AS A RESULT, JIM *ONLY* DIED.	mitigating (!) stress

Table 2: Defending Jim

JIM COULD NOT EXPECT THAT MIKE WOULD BE HURT IF JIM *ACCIDENTALLY* BUMPED HIM; ALSO, A REASONABLE PERSON COULD NOT FORESEE THAT IF HE BUMPED HIM HE WOULD HURT HIM. HE HAD NO INTENTION TO BOTHER MIKE. MIKE *ONLY WAS BUMPED* BY JIM ONCE. THE ACTION WAS AN ACCIDENT.	mitigating adverb mitigating stress mitigating passive
THEN MIKE REALIZED THAT JIM HURT HIM. *IN ADDITION*, MIKE DID NOT EXPECT THAT JIM WAS GOING TO HURT HIM ANY LONGER. MIKE'S GOAL WAS TO INJURE JIM. MIKE COULD FORESEE THAT HE WOULD INJURE HIM IF HE PURPOSELY HIT HIM ONCE. HE HIT HIM. THE RESULT WAS THAT HE INJURED HIM. HE REQUIRED JUSTIFICATION FOR CAUSING HIM TO BE INJURED. THE ACTION WAS AN ESCALATED RETALIATION.	enhancer phrase include all topics that are BAD for Mike...
NEXT, JIM REALIZED THAT MIKE INJURED HIM. JIM *JUST TAPPED* MIKE ONCE. THE ACTION WAS A SIMPLE RETALIATION.	mitigating stress mitigating verb
...	etc...
AS A RESULT, JIM DIED.	

4 Conclusion

In summary, consider the generation of a complete fight, from the set of interpretations of each action PAULINE gets from JUDGE. In table 1, PAULINE's sympathies are for Mike, so that it stresses Jim's part of the fight; in table 2, from the same input, PAULINE defends Jim.

5 Acknowledgements

Thanks to Bill Bain for JUDGE, and to Jeff Grossman and Ashwin Ram for lots of hard work on PAULINE.

6 Bibliography

1. Appelt, D.E.,
Planning Natural Language Utterances to Satisfy Multiple Goals, Ph.D. dissertation, Stanford, 1981.

2. Bain, W.M.,
Case-based Reasoning: A Computer Model of Subjective Assessment,
Ph.D. dissertation, Yale University, forthcoming 1986.

3. Cohen, P.R.,
On Knowing What to Say: Planning Speech Acts, Ph.D. dissertation, University of Toronto, 1978.

4. Grosz, B.J.,
The Representation and Use of Focus in Dialogue Understanding, Stanford Research Institute Technical Report no. 151, 1977.

5. Goldman, N.M.,
Conceptual Generation, in **Conceptual Information Processing**, Schank, R.C. (ed), North-Holland Publishing Company, 1975.

6. Lehnert, W.G.,
Plot Units: A Narrative Summarization Strategy, in **Strategies for Natural Language Processing**, Lehnert, W.G. & Ringle, M.H. (eds), Lawrence Erlbaum Associates, 1982.

7. McKeown, K.R.,
Generating Natural Language Text in Response to Questions about Database Queries, Ph.D. dissertation, University Of Pennsylvania, 1982.

8. Sidner, C.L.,
Toward a Computational Theory of Definite Anaphora Comprehension in English, Massachusetts Institute of Technology Technical Report AI-TR-537, 1979.

Illusory Conjunctions of Objects and Forms: Integration Errors in a Very Short-term Store

Helene Intraub

Department of Psychology
University of Delaware

ABSTRACT

Two experiments tested the predictions of an integrative buffer model of visual processing, regarding the illusory conjunction of components of rapidly presented displays. Color pictures of objects were presented at a rate of 9/s, in the same spatial location. Experiment 1 used a modified report procedure to test the hypothesis that Stroop-like response competition, during naming, not a perceptual error, resulted in the high confidence "illusory conjunctions" reported in previous research. Subjects were provided with the name of a picture in advance and reported "yes" or "no" to indicate if that picture was the one in the frame. Contrary to the response competition hypothesis, high confidence errors occurred frequently under these conditions. Experiment 2 tested the hypothesis that the direction of migration (preceding or following picture) is the result of a difference in the sequential allocation of attention to the frame first or to its "host" picture first on different trials. As predicted by the integrative buffer model, subjects were faster in detecting the frame when they confidently reported it around the preceding picture than around the following picture in the sequence, and reaction times associated with correct reports fell between the two.

INTRODUCTION

The purpose of these experiments was to test the predictions of a model of the early stages of scene processing that Intraub (1985a) proposed might account for temporal migration and illusory conjunctions of components of visual displays. Temporal migration is a type of illusory conjunction of visual components that occurs when stimuli are presented in rapid succession (e.g., 9-20 items per second) in the same spatial location. Although components of a single display are simultaneously presented they are sometimes reported, with high confidence, as having occurred separately. Subjects will report a component as having been an integral part of the preceding or following display in the sequence. There have been reports of color migration among letters (McLean, Broadbent & Broadbent, 1983), letter case among words (Lawrence, 1971), form migration among letters, words, numbers and pictures (Gathercole & Broadbent, 1984; Intraub, 1985a) and object migration among scenes (Intraub 1985b).

The model proposed to account for these phenomena, is based upon a model of the early stages of visual processing, in which a very short-term conceptual/visual buffer memory plays a central

role (Potter, 1976; Avons & Phillips, 1980). Several experiments using search tasks or memory tasks have provided evidence for a very short-term post-categorical store that is distinct from the icon (Avons & Phillips, 1980; Intraub, 1981, 1984; Loftus & Ginn, 1984; Potter,1976). Intraub (1985a) proposed that this conceptual short-term store may play a role in the integration of visually presented information.

According to this view, migration occurs when identification time is slow relative to presentation rate because one display is still being analyzed in the buffer at the same time that processes are initiated on the next new display. Thus a black outline frame migrates among pictures, but parts of the pictures themselves do not (Intraub, 1985a) because the former display requires more identification time. Similarly, to obtain the same level of frame migration among different types of stimuli, a faster presentation rate must be used for stimuli that are relatively easy to identify, (e.g., numbers and letters) as compared with those that are more difficult (e.g., colored objects and words; Gathercole and Broadbent, 1984; Intraub 1985a).

The reason that components sometimes migrate to the preceding display and sometimes to the following one, can be explained in terms of the allocation of attention. Consider Intraub's (1985a) task in which subjects must report which object was presented with a black frame around it. The preceding picture in the sequence is in the short-term buffer when the picture with the frame is presented. Although presented simultaneously, the frame and picture are not rapidly integrated because they are not meaningfully related. As a result, if the subject pays attention to the frame first, it may become integrated with processing of the previous picture, which is still in the visual/conceptual buffer. If the subject pays attention to the target picture first, then the frame, which is now in the buffer, may become integrated with the next picture in the sequence.

Two experiments tested different aspects of the integrative buffer model. Experiment 1 used a modified report procedure and showed that high confidence migration reports cannot be attributed to difficulties in naming rapidly presented displays. Experiment 2 tested the validity of the attention allocation hypothesis as an explanation of differences in the direction of migration, using a reaction time task in conjunction with a standard temporal migration task.

EXPERIMENT 1

To determine if high confidence error reports obtained in temporal migration experiments are an artifact of the naming requirement during high speed presentation, the report procedure was changed from the unconstrained naming procedure typically used in temporal migration experiments (which requires subjects

to access and articulate the proper name for the target) to a "yes-no" detection task. A picture was named by the experimenter and the subject's primary task was to answer "yes" or "no" to indicate whether or not the frame was around the specified picture. If high confidence error reports to -1 and +1 pictures are the result of response competition during naming, then no high confidence error reports to specified -1 and +1 pictures should be obtained using this procedure. Specification of -2 and +2 pictures by the experimenter in some sequences, served as "catch trials" to allow a measure of "yes" response bias.

Method

Subjects and Apparatus. Subjects were 18 individually-tested undergraduates. They were seated approximately 1.7 m from a rear projection screen. The image was projected from an adjacent room using a Visual Instrumentation Corporation Selecta-frame 5, data analyzer 16mm projector at silent speed (18 frames per second). The size of the field was 14 x 20 cm, which subtended a visual angle of approximately 5^{o} x 8^{o}.

Stimuli. The stimuli were twelve objects that were cut out from magazines and photographed on a gray background. These are the same stimuli that were used in Intraub (1985b: Experiments 2 and 3). The objects were: a car, a hot air balloon, a suitcase, an organ, a chair, a tractor, a goblet, an American flag, a stove, a pair of eyes, a movie projector, and a truck.

Design. Each sequence contained all twelve pictures, with one of the twelve in the black frame. Each picture was photographed with the black frame around it six times, yielding 72 sequences. On the six occasions that a picture was the target it appeared with three different pairs of flanking pictures (-1 and +1 pictures), such that on one occasion the order was ABC and on the other it was CBA (B is the picture with the frame). Subjects were divided into three equal groups. What differed among the three groups, was the position of the specified picture (-1, target, or +1) in a given sequence. The position of the specified picture in each sequence was counterbalanced across subject group. Twelve additional sequences were filmed to serve as "catch trials" (-2 and +2 pictures specified).

Procedure. Before the experiment, subjects were familiarized with the pictures and received practice naming them. Following this, they were told that prior to each sequence, the experimenter would name a picture while they looked at the fixation point. Their task was to respond "yes" or "no", followed by a confidence rating (sure, pretty sure, not sure, guess), to indicate whether or not the frame was around the specified picture. On trials eliciting a "no" response they were asked to report the picture that they saw the frame around, along with a second confidence rating.

Results

The results support the claim that the migration effect obtained with pictures and forms (Intraub, 1985b) is not due to difficulty in naming rapidly presented pictures. All subjects persisted in reporting the frame around temporally adjacent pictures in the sequence, with high confidence. As in Intraub (1985b) the yes/no procedure yielded the same general pattern of results as the unconstrained naming procedure (Intraub, 1985a, 1985b).

Table 1 shows the percent of trials in each specification condition to which the subject reported "yes" with high confidence. Collapsing over minus and plus positions, a repeated measures ANOVA showed a significant main effect of position, ($F_{(1,34)} = 61.48$, $p < .001$). High confidence "yes" responses decreased as the specified picture was further removed from the target. A planned comparison of the percent of "yes" responses in the -1/+1 condition and the -2/+2 position shows that the large number of erroneous "yes" responses to -1/+1 pictures cannot be accounted for by a guessing bias because there were significantly fewer "yes" responses to -2/+2 pictures ($p<.001$). The means for the -1/+1 and -2/+2 conditions, were 44.9 (SD = 21.1) and 13.8 (SD = 18.0), respectively.

Whether the relatively small number of "yeses" to -2/+2 pictures, reflects a guessing bias or contained actual integration errors is being addressed in other research. The major point being made here is that a strong migration effect was obtained with these sequences using the yes/no task.

Table 1

Mean percent of total possible responses in each specification condition that were high confidence "yes" responses (Exp 1)

	Specification Condition				
	-2	-1	0	1	2
Mean	14.9	43.6	63.3	46.1	13.8
SD	17.0	23.1	18.7	24.9	21.3

EXPERIMENT 2

The purpose of Experiment 2 was to test the attention allocation assumption of the model using a reaction time task in conjunction with the standard temporal migration task. According to the attention allocation assumption, -1 errors occur when subjects attend to the frame before the "host" picture and +1 errors occur when the subjects attend first to the "host" picture and then to the frame. If the assumption is true, then reaction time to frame detection (measured with a key press) should be faster when subjects report the frame around the -1 picture, than around the +1 picture. The times obtained on trials in which the frame is reported around the host picture should neither be faster than the -1 time nor slower than the +1 time.

Method

Subjects and stimuli. The subjects were 27 undergraduate volunteers who were paid $3.00 for their participation. The filmed sequences were the same as in Experiment 1.

Apparatus. The apparatus was the same except for the addition of a hand-held key press which the subject depressed with his or her thumb. Reaction times were measured using an Apple II Plus computer that was interfaced with the digital frame counter of the variable speed data-analyzer projector. Reaction times from the onset of the frame to the key press were accurate to 1 msec.

Procedure. Subjects were familiarized with the pictures, and were given practice naming them. Following this, they were told that their task was to press the key in the shortest time possible in response to the frame, and then to immediately report which picture it had occurred with, along with a confidence rating. Each subject was provided with 24 practice sequences followed by 168 experimental sequences. After each group of approximately 28 sequences, subjects received a brief break and viewed a relatively slow presentation of the 12 pictures which they were asked to name.

Results

Table 2 shows the percentage of pictures reported as occurring with the frame as a function of the picture's position in the sequence and confidence rating. The table shows that addition of the reaction time task did not alter the pattern of results typically obtained with these and other pictorial stimuli using the unconstrained naming procedure (Intraub, 1985a, 1985b).

Table 2

Percent of responses reporting the frame around the correct scene (0), preceding scenes (-), or following scenes (+) in the sequence as a function of confidence level (Experiment 2)

	Position of the reported scene in the sequence						
Confidence	-3	-2	-1	0	1	2	3
Sure	0	1	8	81	9	0	1
Pretty Sure	2	2	19	58	15	1	3
Not Sure	7	6	23	36	18	3	6
Guess	12	7	19	28	15	6	13

The mean reaction time to frame detection as a function of the position of the reported picture in shown in Table 3. As may be seen in the table, the pattern of results followed the prediction of the model. Reported position of the frame significant affected reaction time ($F(2,52) = 19.35$, $p < .0001$).

Table 3

Mean reaction time to detecting the frame as a function of the position of the picture the subject reported seeing it with when the subject reported high confidence in the response (Exp 2)

	Position of the reported picture		
	-1	0	+1
Mean	327	332	353
SD	42	40	46

Frame detection was faster on trials in which subjects reported the frame around the -1 picture than around the +1 picture ($t_{(26)} = 5.33$, $p < .005$). This comparison was also significant by sign-test, with 22 of the 27 subjects showing faster RTs to -1 pictures.

Conclusions

The results of these experiments are consistent with the integrative buffer model proposed by Intraub (1985a). According to this view of illusory conjunctions, they occur because information is still being processed in a conceptual/visual buffer, at the time that the new display is presented. Parts of the pictures themselves do not migrate at these speeds because components of the pictures can be quickly integrated (e.g., eyes in a face). This integrative buffer would not necessarily play a role in integrating the information from successive fixations (see Pollatsek, Rayner & Collins, 1984), because it seems to be working during the a time interval comparable to a single fixation. Indeed if film speed is reduced so that rate of picture presentation approximates the average eye fixation frequency of about 3 or 4 per second, no frame migration occurs. These phenomena seem more related to temporal integration times in perception, and provide a new means with which to study the perceptual organization of the components of simultaneously presented visual information.

REFERENCES

Avons, S.E. & Phillips, W.A. (1980). Visualization and memorization as a function of display time and poststimulus processing time. Journal of Experimental Psychology: Human Learning and Memory, 6, 407-420.

Gathercole, S. E. & Broadbent, D. E. Combining attributes in specified and categorized target search: Further evidence for strategic differences. Memory and Cognition, 12,329-337.

Intraub, H. (1985a). Visual dissociation: An illusory conjunction of pictures and forms. Journal of Experimental Psychology: Human Perception and Performance, 11, 431-442.

Intraub, H. (1985b). Illusory Conjunctions: Integration errors in a very short-term store. Psychonomic Society Convention, Boston, MA, November, 1985.

Intraub, H. (1985c). Conceptual masking of briefly glimpsed photographs In Groner, R., McConkie, G.W. & Menz, C. (Eds.), Eye movements and psychological processes. North Holland Publishing Co.

Lawrence, D.H. (1971). Two studies of visual search for word targets with controlled rates of presentation. Perception and Psychophysics, 10, 85-89.

Loftus, G.R. & Ginn, M. (1984). Perceptual and conceptual masking of pictures. Journal of Experimental Psychology: Learning, Memory and Cognition, 10. 435-441.

McLean, J.P., Broadbent, D.E. & Broadbent,M.H.P. (1983). Combining attributes in rapid serial visual presentation tasks. Quarterly Journal of Experimental Psychology, 35, 171-186,

Pollatsek, A, Rayner, K. & Collins, W.E. (1984). Integrating pictorial information across eye movements. Journal of Experimental Psychology: General, 113, 426-442.

Potter, M.C. (1975). Meaning in visual search. Science, 187, 965-966.

Potter, M.C. (1976). Short-term conceptual memory for pictures. Journal of Experimental Psychology: Human Learning and Memory, 2, 509-522.

The literate listener: Effects of spelling on syllable judgments

Jola Jakimik
Department of Psychology, University of Wisconsin-Madison

Rebecca Treiman
Department of Psychology, Wayne State University

ABSTRACT

We investigated the effects of spelling knowledge on the representation of spoken language. In the experiment, subjects first saw the written form of a nonsense word, then heard it and judged the number of syllables. For the identical acoustic tokens, the number of judged syllables varied with the accompanying spelling. The effect of spelling was stronger for one syllable pronunciations than for two syllable pronunciations. The results are discussed in relation to the role of spelling knowledge in listeners' representations of phonology.

INTRODUCTION

The alphabetic principle of our writing system links the written and spoken forms of words. Awareness of the phonological structure of speech has been shown to play an important role in learning to read and write (e.g., Treiman & Baron, 1981). We were concerned with influences in the opposite direction; that is, with effects of orthographic knowledge on listeners' notions of words' phonological forms. Specifically, we asked whether the number of syllables people think a spoken word contains depends on how they think the word is spelled.

Some evidence that knowledge of an alphabetic writing system shapes one's conception of spoken language was provided by Morais, Carey, Bertelson, and Alegria (1980). These researchers compared literate and illiterate adults on their ability to segment spoken words into phonemes. Morais et al. had their subjects add or delete phonemes from words and nonwords. They found that the illiterate adults had difficulty in manipulating phoneme-sized units. Morais et al. concluded that learning to read is important for the ability to think of speech as composed of separate units.

Further evidence that knowledge of spelling shapes one's conception of the component sounds of a word was provided by Ehri and Wilce (1980). Ehri and Wilce had fourth grade children count the sounds in familiar words. They compared pairs like rich and pitch which have equivalent pronunciations, but different spellings. Ehri and Wilce found that the children counted more sounds in the words with more letters in their spelling. Since children may have had different mental representations of the pronunciations of these existing words, Ehri and Wilce conducted a second study in which they taught children spellings for novel words. For example, children were taught either ZICH or ZITCH. In a subsequent session, they counted the sounds in the novel words. Again, children counted more sounds in the words with extra letters.

Knowledge of the spelling of a word influenced children's judgments of how many sounds the spoken form contained, both for preexisting spellings and newly learned ones.

These demonstrations of an influence of spelling on mental representations of phonology have dealt with early developmental influences. We sought to demonstrate such an influence in adult listeners. We all continue to learn new words and must represent their phonological properties. Does knowledge of spelling influence these representations, and in what way? In our experiment, we varied both the pronunciation of the new words and their spelling. We presented the written forms first.

Instead of looking at judgments of the phonemic structure of novel spoken words, we looked at judgments regarding syllabic structure. The spelling of a word conveys information about several levels of linguistic structure, including syllable composition. Compare the two spellings FLOUR and FLOWER. We have the intuition that the former is a one syllable word and the latter a two syllable word, even though they can be pronounced identically.

METHOD

Materials and Design

The critical items were 16 pairs of nonwords. The nonwords were based on pairs of real words like rule - jewel and owl - towel. Nonwords were constructed by combining the rhyme portions of the words with new onsets (i.e., single consonants and clusters). Two nonword pairs were formed from each word pair. For example, rowl - rowel and spowl - spowel were formed from owl - towel. The first members of the pairs have "one syllable spellings" and the second have "two syllable spellings." Thus, each nonword pair represents two possible pronunciations and two possible spellings.

In addition, there were 10 one syllable fillers (e.g., blee), 8 two syllable fillers (e.g., tarrid), and 16 three syllable fillers (e.g., repliment).

Two lists were constructed from these stimuli. Each list contained 16 critical nonwords and 34 fillers. The lists were constructed such that one member of each critical pair occurred on each list. For example, rowl and spowel occurred on one list, and rowel and spowl on the other. The 50 nonwords were grouped into 10 sets of 5 items each.

The lists of nonwords were recorded by the first author. The speaker attempted to pronounce each critical item in accordance with its spelling on a given list. Each trial began with the word "ready", then after a 3 to 4 second pause the nonword was pronounced. About one second later, the speaker said "turn the page" and there was a 5 second pause before the next "ready." This sequence was repeated for five items. After the last item in a set, the speaker said "Now recall the words in any order." There was a 30 second silent period on the tape. Each tape began with a practice set of five real

words.

The taperecorded materials were accompanied by booklets of written materials. The booklets contained the written forms of the nonwords, typed in large capital letters in the center of a page, followed by answer sheets for the syllable counting task. Following the answer sheet for the fifth item in each set was a sheet with five recall clues, consisting of the first consonant or consonants of the nonwords. At the beginning of each set of five items, there was a sheet with the instruction "Turn this page when you hear ready."

There were two versions of the written materials, with the one syllable spelling in one version, and the two syllable spelling in the other. For example, ROWL and SPOWEL appeared in one version; ROWEL and SPOWL appeared in the other.

Each tape was paired with each version of the written materials. Thus, the written form of each critical item was either the same as its pronunciation on a given tape (match), or was the alternate spelling (mismatch). Four groups of listeners received the four combinations of recorded and written materials, so that the spelling manipulation was between subjects.

Procedure

Subjects were tested individually in a sound-treated room. They were told that the experiment investigated how people remember new words, and that they would see and hear nonsense words and then recall them. At the beginning of each set of five, subjects waited for the word "ready," and then turned the cover sheet to see the first printed nonsense word. They then heard the item pronounced. Then they turned the page, and had 5 seconds to mark the number of syllables. When they finished, they turned to the next printed nonsense word. In this way, subjects viewed the printed form of the nonsense word for several seconds before they heard it spoken. After marking the number of syllables in the fifth item in each set, subjects recalled the last five items, into the microphone, using the clues provided in the booklet. After 30 seconds, they went on to the next set.

Subjects

Forty undergraduates at the University of Wisconsin-Madison participated for course credit or for payment. All were native speakers of English.

Spelling and Pronunciation Judgments

Two additional sets of judgments were collected from other subjects. One group of forty subjects only saw the spellings of the nonsense words, and indicated how many syllables each had. A second group only heard the taperecorded nonsense words (re-recorded from the tape used in the main experiment), and judged the number of syllables.

RESULTS

Only the results from the syllable judgment task are reported here. The recall results will be presented in a later report.

Subjects' responses were scored as consistent or inconsistent with the intended number of syllables in the pronunciation. For the fillers, there were only 9 disagreements (out of 1360 responses) with the intended number of syllables. For the critical items, all but one of the inconsistent responses corresponded to the other (one or two) possible number of syllables.

Table 1 presents the percentage of responses consistent with the pronunciations of the nonsense words. When the number of syllables in the spelling matched the pronunciation, 84% of the responses agreed with the intended number of syllables. When the spellings and pronunciations mismatched, only 31% agreed with the pronunciation. The effect of spelling on syllable judgments was much stronger for one syllable pronunciations than for two syllable pronunciations. When a two syllable spelling accompanied a one syllable pronunciation, subjects generally judged that the word contained two syllables. When a one syllable spelling accompanied a two syllable pronunciation, one and two syllable judgments were about equally common.

These patterns were supported by statistical tests across subjects and items. The effect of spelling was highly significant in the subjects analysis, $F(1,38) = 127.94$, $p < .001$, and in the items analysis, $F(1,30) = 304.85$, $p < .001$; min $F'(1,63) = 90.12$, $p < .001$. The interaction between spelling and pronunciation was significant in both the subjects ($F(1,38) = 7.49$, $p < .01$) and items ($F(1,30) = 24.30$, $p < .001$) analyses; min $F'(1,58) = 5.73$, $p < .05$.

TABLE 1
Percent of responses consistent with pronunciation

	Spelling	
	Match	Mismatch
Pronunciation		
One Syllable	79	11
Two Syllable	89	51

TABLE 2
Percent agreement with intended number of syllables

One Syllable Spelling	86
Two Syllable Spelling	90
One Syllable Pronunciation	70
Two Syllable Pronunciation	70

Table 2 presents the agreement between subjects' judgments and the intended number of syllables for the written and spoken forms alone. Syllable judgments based solely on the spellings of the nonsense words agreed with our intentions almost 90% of the time. The amount of agreement was essentially equivalent for the one and two syllable spellings. Syllable judgments based solely on hearing the nonsense words agreed with our intentions 70% of the time. Although there was less agreement with regard to pronunciations than with regard to spellings, the attempts at one and two syllable pronunciations were equally successful.

DISCUSSION

When adult listeners counted syllables in a novel spoken word, the number depended on the spelling they had just read. To take a particular instance, the one syllable pronunciation of "shule" was judged as monosyllabic by 95% of listeners who only heard it. With the spelling SHULE, 9 out of 10 listeners judged it as having one syllable. With the spelling SHEWEL, 8 out of 10 listeners judged it as having two syllables. The tendency of our adult listeners to count extra syllables in accord with the spellings of novel words is similar to Ehri and Wilce's (1980) finding that children counted extra sounds in words with additional letters.

On what were our subjects' syllable judgments based? The presence of an interaction between spelling and pronunciation suggests that both spelling and pronunciation were used. Before elaborating on this aspect of the results, two uninteresting interpretations of the interaction must be ruled out. If we consider the judgments based on pronunciation alone, we cannot attribute the interaction between spelling and pronunciation to greater ambiguity in the number of syllables in one syllable pronunciations. The one and two syllable pronunciations were comparable in terms of agreement on the number of syllables. Neither can the interaction be attributed to weaker bias of the one syllable spellings. Overall, the alternative spellings were equally strongly biased for the intended number of syllables.

The greater power of spelling over one syllable pronunciations than over two syllable pronunciations can be interpreted by postulating a distinction between a word's actual pronunciation and its "ideal" or canonical pronunciation. A listener's notion of a word's ideal pronunciation is affected by the word's spelling as well as by its phonological and morphological structure. We further postulate that ideal pronunciations tend

to be more complex and differentiated than actual pronunciations. It is natural for listeners to set up a two syllable ideal pronunciation (e.g., "shewel") when they only hear one syllable ("shule"); it is less natural for them to set up a one syllable ideal when they hear two syllables. This asymmetry reflects the fact that polysyllabic words are often shortened (e.g., "general" becomes "gen'ral"), while it is less common for syllables to be added. Thus, listeners were more willing to insert an unheard syllable based on the spelling, than to ignore a heard syllable.

Further support for the distinction between actual and ideal pronunciations, and for the idea that ideal pronunciations are more differentiated than real pronunciations, comes from a study by Taft and Hambly (1985). These researchers asked subjects to monitor known words for targets presented auditorily. According to linguists' phonemic analyses, "safari" begins with /səf/, but the majority of Taft and Hambly's literate listeners took it to begin with /sæf/. Taft and Hambly concluded that orthography influences the mental representations of words. Their listeners may be said to have thought that "safari" really contains an unreduced vowel, regardless of the acoustic-phonetic information that is usually present. The ideal representation differs, in this case, from the phonemic representation. It may be considered a metaphonological representation -- a representation whose units are available to conscious awareness and manipulation.

Cowan, Braine, and Leavitt (1985) provide some further data on the nature of metaphonological representations. They studied people who "talk backward," reversing the order of units in words. Some subjects reversed the spellings of words, others reversed sounds. But even subjects who manipulated sounds re-ordered units which were defined by orthography. For example, the consonant cluster /ks/ was treated as a unit only when it was spelled with an X. Cowan et al. argue that knowledge of spelling acts as an important source of constraint in the development of metaphonological knowledge.

Our results support this claim and, together with other recent findings (Donnenwerth-Nolan, Tanenhaus, & Seidenberg, 1981; Jakimik, Cole, & Rudnicky, 1985; Seidenberg & Tanenhaus, 1979), imply a very tight connection between the written and spoken forms of words. Orthographic knowledge, it appears, influences performance in a variety of tasks that do not logically require it.

REFERENCES

Cowan, N., Braine, M. D. S., & Leavitt, L. A. (1985). The phonological and metaphonological representaion of speech: Evidence from fluent backward talkers. Journal of Memory and Language, 24, 679-698.

Donnenwerth-Nolan, S., Tanenhaus, M. K., & Seidenberg, M. S. (1981). Multiple code activation in word recognition: Evidence from rhyme monitoring. Journal of Experimental Psychology: Human Learning and Memory, 7, 170-180.

Ehri, L. C., & Wilce, L. S. (1980). The influence of orthography on readers' conceptualization of the phonemic structure of words. Applied Psycholinguistics, 2, 371-385.

Jakimik, J., Cole, R. A., & Rudnicky, A. I. (1985). Sound and spelling in spoken word recognition. Journal of Memory and Language, 24, 165-178.

Morais, J., Carey, L., Alegria, J., & Bertelson, P. (1979). Does awareness of speech as a sequence of phones arise spontaneously? Cognition, 7, 323-331.

Seidenberg, M. S., & Tanenhaus, M. K. (1979). Orthographic effects on rhyme monitoring. Journal of Experimental Psychology: Human Learning and Memory, 5, 546-554.

Taft, M., & Hambly, G. (1985). The influence of orthography on phonological representations in the lexicon. Journal of Memory and Language, 24, 320-335.

Treiman, R., & Baron, J. (1981). Segmental analysis ability: Development and relation to reading ability. In G. E. MacKinnon & T. G. Walker (Eds.), Reading research: Advances in theory and practice (Vol. 3). New York: Academic Press.

Modifying Explanations to Understand Stories*

Alex Kass
Yale University Department of Computer Science

ABSTRACT

We describe a system that learns new schemas by modifying old ones, in order to understand anomalous events in stories that it reads. We discuss how these schemas (called Explanation Patterns [Schank 86]) are structured in order to make them modifiable, and how the understanding process applies and modifies them. This model bridges the gap between two previous models of understanding, which were based on either application of prestored schemas or understanding-time inference chaining. By employing modifiable schemas, our model is more flexible than the former and more efficient than the latter.

Introduction

The crucial problem in story-understanding is inference. Inference is needed, among other things, to tie different sentences together, to fill in ellipses, and to resolve ambiguity. The story-understanding literature contains several models of the inference process, but for the purposes of this paper we can divide these models into two principal categories: One group operates via the application of prestored schemas; the other builds inference chains from scratch at understanding time. Both models have something to offer, but each has serious drawbacks as well. In this paper we argue that in order for a system to overcome these drawbacks it must be able to adapt its schemas to new situations. We discuss a new model (implemented in a computer program called SWALE) that is based on modifiable schemas, called Explanation Patterns (henceforth XPs).

Previous Work — Inference Chaining and Schema Application

The inference-chaining approach to story understanding is exemplified by Rieger's Conceptual Inferencer [Rieger 75] and Wilensky's PAM [Wilensky 78]. Rieger's system was completely bottom up; his program chained forward from each input sentence and noticed where the chains intersected. Although this simple model was able to generate many useful inferences it was too unconstrained. The combinatorial explosion of inferences caused the irrelevant to overwhelm the useful. Wilensky was able to constrain things somewhat by having knowledge of plans and goals guide the inference process. PAM would attempt to match inputs to known goals of the actor, and to backward chain to those goals if they couldn't be matched directly. Guiding the inference process in this way helps matters considerably but it remains the case that PAM had to do a lot of work to understand each story. Furthermore, because it didn't store the inference chains it built, it had to do as much work to understand a story it had seen a hundred times as it did to understand it the first time.

A response to the inference-chaining school is the schema-application approach, as practiced, for example, by Charniak's Ms. Malaprop [Charniak 72], and Cullingford's SAM [Cullingford 78]. These programs avoid most understanding-time inference by using prestored schemas that contain the expectations needed to understand the story. For these programs, the inference process is reduced to matching input from the story against a schema in memory. They are therefore very efficient at handling stories that closely match their schemas, but fail badly unless there is a very close match.

*This work is supported in part by the Air Force Office of Systems Research under contract 85-0343

Our Approach — Schema Modification

Given that schema-based programs tend to be brittle but efficient, while the inference-chainers are more flexible but less efficient, one might be tempted to propose that there are two different modes of story-understanding; that a complete model would combine SAM and PAM, using scripts when they were applicable, and resorting to inference-chaining when necessary. Such an either or system would be a good idea if stories were generally totally novel or totally old-hat. The fact is that most are neither; interesting stories we read are usually reminiscent of things we understand well, but not exactly like them — they are near misses and we don't want our model to have to understand these from scratch. Rather, we want to be able to tweak our old schemas to make them applicable. This way the understander learns new schemas incrementally when old ones fail. Old schemas serve as the starting points for creating the new, and the wheel does not need to be reinvented in order to build a slightly different schema. SAM-like scripts, however, are not good candidates for modification because they don't encode enough of the causal reasoning that went into building them to make it clear to a tweaking mechanism what modifications are reasonable to make. In order to make the tweaking idea work, we need to specify what these modifiable schemas should look like and what the process will be that modifies and applies them. These are exactly the goals of the SWALE project, which we will spend the rest of this paper describing.

XPs — Modifiable Schemas

XPs differ from scripts in that scripts are meant to provide an overall view of some group of events (such as a doctor's visit or a meal at a restaurant) while XPs are designed to be a detailed trace of the reasoning that was used to resolve a particular problem that could arise in such a group of events (such as the waiter not bringing your food). An XP is a set of beliefs and a set of belief-support relations which link the beliefs together in an inference network. The belief-support links specify which beliefs depend on which others, and what the type of the dependency is. This is what makes XPs modifiable. The belief supports indicate *why* a given belief is in the XP. This tells the modification process what the effect of deleting or changing beliefs will be.

Some of SWALE's XPs

The best way to give a feel for what XPs are like is to describe some the XPs we have equipped the SWALE program with. Consider the following story: "Swale, a successful 3-year old race horse, was found dead in his stall a week after winning the Belmont Stakes race." Most people who read this story find the death to be an anomalous event, which they feel the need to explain. What follows are some death-related XPs that we have built into SWALE, Some of these XPs are clearly relevant to Swale's death while others are connected only in a more fanciful way. Our goal is to have SWALE propose as many interesting explanations as possible by applying the XPs we have indexed in its XP library and tweaking them to create new variations of these XPs; we are much more interested in having the system develop interesting hypotheses than in having it avoid bad ones.

The Jim Fixx XP: Joggers jog a lot. Jogging results in physical exhaustion because jogging is a kind of exertion and exertion results in exhaustion. Physical exhaustion coupled with a heart defect can cause a heart attack. A heart attack can cause death.

The Janis Joplin XP: Being a star performer can result in stress because it is lonely at the top. Being stressed-out can result in a need to escape and relax. Needing to escape and relax can result in taking recreational drugs. Taking recreational drugs can result in an overdose. A drug overdose can result in death.

Too Much Sex XP: Too much sex can kill you.

Preoccupation XP: Being preoccupied about something can cause you to be inattentive. Being inattentive can result in walking into traffic. Walking into traffic can result in you being hit by a vehicle. Being hit by a vehicle can cause death.

Despondent Suicide XP: Thinking about something you want but that you don't have can make you despondent. Being despondent can result in suicide.

The idea is that when an anomaly is encountered while reading a story or having an experience in the real world, the understander uses features of the anomaly to retrieve XPs that might explain the anomaly. For any anomaly an understander encounters, there are four possible states of readiness its XP library might be in:

1. An XP can be retrieved from memory that applies perfectly to the anomaly. In this case it is easy

to explain the anomaly. It is a simple case of schema-application.

2. None of the XPs retrieved from memory apply directly, but some relevant XPs can be modified (we call this tweaking) to create new XPs that are applicable. This case is trickier, but it is also more important, since, at the end of the process the system has created and learned a new XP.

3. No XP is retrieved that can be tweaked to apply. An XP must be built from scratch out of the primitive inference rules. This is the kind of situation for which PAM was designed. Under our current model, understanders often give up on problems for which they are so unprepared unless the anomaly is quite important. Although a complete model should include a component to handle this case, the SWALE program currently does not.

4. The system does not even have the inference rules to allow it to build an appropriate XP. In this case the system simply is not equipped to understand the anomalous experience.

The SWALE Process Model

My claim is that most interesting experiences fall into category 2; we understand them by retrieving an existing XP and modifying it to fit the new situation. This is what the SWALE program does. Thus SWALE is both an understanding program and a learning program. Its actions are driven by the goal of discovering an explanation that will help it understand an anomalous event, and in doing so it learns new explanations, and stores them for future use. After an anomaly is detected the algorithm involves the following:

XP SEARCH: Search the XP library for an XP that may apply to the anomaly.

XP EVALUATION: Attempt to apply XPs. If successful then skip to XP INTEGRATION.

XP TWEAKING: If unable to apply XPs directly, attempt to tweak them into XPs that might apply better. If successful send tweaked XPs back to XP EVALUATION.

XP INTEGRATION: If results accepted, integrate into memory making appropriate generalizations.

The program is divided into three modules. The main driver is called the Accepter and is responsible for detecting anomalies, evaluating explanations, and generalization. One sub-system is devoted to retrieving XPs and another to tweaking them. David Leake is working on the Accepter, Chris Owens on the Retriever, and I am working on the Tweaker. Figure 1 presents a simple description of the SWALE architecture.

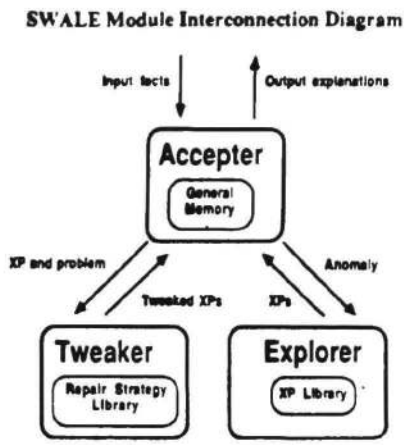


Figure 1: SWALE Module Interconnection diagram

Unfortunately, there isn't enough room here to describe the entire SWALE program. There are, of course, important issues that arise throughout the processing of the story. For example, the reader might wonder how SWALE notices anomalies, how it searches for relevant XPs and how it evaluates explanations. These are all interesting issues that we simply don't have room to touch on here — this paper is really only about how XPs that have been retrieved and evaluated as near misses can be modified. Some of the other issues are discussed in [Leake and Owens 86]. A somewhat more detailed (albeit out of date) discussion of the entire SWALE program appears in [Kass, Leake and Owens 86].

Tweaking XPs to Make Them Fit

When the Accepter rejects an XP as a near miss it passes the XP and the reason for rejection to the Tweaker. The goal of the Tweaker is to generate a new XP (or set of XPs) that might fit better. These are then sent back to the Accepter for re-evaluation.

The high level control structure of the Tweaker is very simple. There are two main sub-steps:

Strategy Retrieval: Retrieve XP REPAIR STRATEGIES from a library of such strategies maintained by the Tweaker. Retrieval of an appropriate set of XP

REPAIR STRATEGIES relies on using the XP FAILURE TYPE, created during the evaluation stage, as an index into the XP REPAIR STRATEGY library. Each XP REPAIR STRATEGY is stored in the library along with a failure pattern. The retrieval step essentially involves matching the XP FAILURE TYPE, against these patterns and collecting the strategies associated with patterns that successfully match. These are the strategies that will be applied in the application step.

Strategy Application: Each strategy is a program designed to map an XP that suffers from some problem to a set of modified XPs that don't suffer from the problem. Apply each retrieved strategy in turn, collecting and returning any resulting XPs. Of course, the details of what goes on during this phase of processing is completely determined by the nature of the XP REPAIR STRATEGY. Some of the XP FAILURE TYPEs generated in the current version of the program, and some of the associated XP REPAIR STRATEGIES are discussed below. This list not intended to be exhaustive by any means; we've just begun to build up SWALE's library of strategies. However, it should convey the basic form that these things take.

In a sense, the SWALE Tweaker represents the application to understanding of the same philosophy that Hammond's CHEF program [Hammond 84] applied to planning. CHEF uses goal-failure configurations to index plan-repair strategies; SWALE uses XP FAILURE TYPEs to index XP REPAIR STRATEGIES. Hammond's theoretical goal was a content-theory of plan repair; ours is a content-theory of explanation repair. Of course, the idea of using failures to index repair strategies traces back at least to HACKER [Sussman 75].

Some XP FAIURE TYPES

NORMATIVE-FILLER-VIOLATION: This failure indicates that the explanation hypothesizes that a role in an action description be filled by an actor who is not a member of the categories that the Accepter expects to fill the role.

When the **JIM FIXX XP** is retrieved it generates this failure description because it calls for SWALE be a jogger, but the program expects joggers to be humans.

SCRIPT-LINE-VIOLATION: This failure is similar, but not identical to the one above. It indicates that the explanation called for an actor to fill a role in a script that it is not valid for it to fill because one of the lines in the script is an action that the actor is actually incapable of performing.

When evaluating the **JANIS JOPLIN XP** this failure is generated because the XP hypothesizes that Swale was the actor in the recreational drugs script, but this script involves injecting oneself, which Swale is not capable of.

SCHEDULING-VIOLATION: This failure complains that an explanation hypothesizes that an action will occur at a particular time but the program has reason to believe that it should have happened at a different time (earlier or later) instead.

For example, the **TOO MUCH SEX XP** calls for Swale to be having sex around the time of his death. The program knows, however, that race horses are kept celibate until sent to stud after their racing days are over. It therefore generates this failure.

UNCONVINCING-SUPPORT-LINK: This failure is quite different from those above. The Accepter isn't complaining about any paritcular belief, but rather about the jump from one to another. The Accepter expects to have inference rules to back up each link in an XP. When a link isn't satisfactorily supported this failure is generated.

Some XP REPAIR STRATEGIES

Associated with each XP FAILURE TYPE is one or more XP REPAIR STRATEGIES. The best way to explain these is to describe some examples. We give some brief descriptions below. Examples of how some are used appears in the description of the SWALE run near the end of the paper.

SUBSTITUTE ALTERNATIVE THEME: This is a fairly general strategy. It is a candidate for fixing any **INVALID-ACTION** failure.

The notion here is to find out which line (or lines) in the inappropriate script was actually important in the XP (*ie.* which one supports other beliefs in the XP), and to search for a theme associated with the actor that can substitute. The new theme must be one that is appropriate for the current actor, and one which has the necessary line(s) in it. In other words, the XP carries within it the knowledge of why a particular belief is important to the XP, and this strategy uses this information to find another theme that can fit in in an analogous way.

SUBSTITUTE EQUIVALENT ACTION: This is quite like the above strategy, and can fix the same set of failures.

Sometimes there are no themes associated with the actor that can substitute for the action that he Accepter has found objectionable. In this case, another way to search for a substitution is to look at generalizations of the objectionable action, and then find other specifications of those generalizations. Any of these other specifications that support the inferences supported by the original action are candidate substitutions.

SUBSTITUTE ANTICIPATION: This rather specialized strategy applies to some **SCHEDULING-VIOLATION** failures when the action in question was expected to happen later than called for in the XP.

When this happens it is reasonable to entertain the notion that thinking about the future event might play a role in the explanation. Thus this strategy substitutes the belief that the actor was thinking about the event for the belief that the event actually occurred. Sometimes this makes sense, although often it doesn't. It is up to the Accepter, rather than the Tweaker, to decide.

FIND CONNECTING XP: This provides a way to fix **UNCONVINCING-SUPPORT-LINK** failures.

It attempts to repair such a problem by finding another XP that will connect the two beliefs in question. It works by calling the Explorer. If no XP can be found to connect the beliefs directly it will try a limited amount of causal chaining. After each chaining step it will call itself recursively, using the newly inferred belief. This is used to find connections between thinking about sex and death.

A Brief Description of a SWALE Run

SWALE is actually a running computer program. There is no room here to present the actual output from the system (which is rather verbose). The following is a brief paraphrase of a SWALE trace in which it develops some variations on explanations contained in its library when it is set to work on the story about Swale's death. This is intended to give the reader a rough idea of the way that SWALE's processing proceeds:

- Use routine search to find XPs concerning premature death in animals. Find **DEATH FROM ILLNESS XP**. This XP can't apply, since Swale wasn't sick. Because this is a severe failure do not try to tweak.
- Look for XPs indexed by unusual features of Swale. Racehorses are in top physical condition; death + top condition retrieves **JIM FIXX XP**. The Evaluator rejects **FIXX XP** because Swale can't be a jogger.
- Try to Tweak. The problem was a **DEFAULT-FILLER-VIOLATION**, so try **SUBSTITUTE ALTERNATIVE THEME**. Swale's known themes are horse-race and eat-oats. The horse-race theme is selected because it involves running, which was the aspect of jogging playing a role in the XP. The tweaked XP is: Since Swale had a heart defect, the exertion from running overtaxed his heart.
- Evaluate the new XP. It's reasonable, but since the heart defect can't be confirmed, continue looking for other XPs. Other strategies fail to find more XPs, so try folkloric explanations of death. Pull up the old wives' tale **TOO MUCH SEX XP**. The evaluator notices that racehorses aren't allowed to have sex while racing, but they do have a lot of sex when they retire to the stud farm. This is a **SCHEDULING-VIOLATION**
- Tweak. The tweaking strategy, **SUBSTITUTE ANTICIPATION** applies to this fault. Could Swale have died just from thinking about life on the stud farm?
- The new XP is unconvincing. There's no link from thinking about sex to death. Tweak.
- Fault is **UNCONVINCING-SUPPORT-LINK**, use the strategy **FIND CONNECTING XP**. Possible effects of thinking about sex are, excitement, and depression (if you're thinking about not having it). Distraction can be linked to death by two XPs, Excitement can cause death by heart-attack. Depression can cause death from suicide.
- Search continues but no more XPs are found. Each of the new explanations depends on conditions which can't be confirmed. Since the Fixx XP was the possibility located most directly by the Explorer, the tweaked version of Fixx is accepted as the most likely explanation. The causally-significant feature Fixx and Swale shared was that they did physical exertion. The XP is generalized to apply to actors who have an exertion theme and this is installed in memory for future use.

Some Explanations Created by SWALE

The Jim Fixx Reminding Explanation: Swale had a congenital heart defect. The exertion of running in horse races strained his heart and brought out the latent defect. He had a heart attack and died.

The Drug Overdose Explanation: Swale's owner was giving him drugs to improve his performance. He accidentally gave him an overdose, which killed him.

The Stud Farm Pair

SWALE's processing brings it to consider the idea that thinking about sex too much caused Swale's demise. It then attempts to imagine ways in which this might have occurred and develops the explanations that follow:

The Sexual Excitement Explanation: Swale was thinking about his future life on a stud farm. Since he was an excitable creature, thinking about the prospects proved to be too much strain for his heart. He had a heart attack and died.

The Despondent Suicide Explanation: Swale was thinking about his forced chastity during his racing career. He became despondent and killed himself.

Conclusion

Explanation Patterns represent frozen inference chains in a way that preserves the reasoning pattern used to develop them. This allows a model using XPs to bridge the gap between schema appliers and inference chainers. SWALE understands stories in terms of its preestablished schemas, but those schemas are not rigid in the way that scripts are. When a schema is a near miss, SWALE it in in order to make it apply. Of course, the further the story strays the more tweaking will be needed, so this model predicts that there will be a continuum of difficulty, from straight application through more and more major tweaking, to building the explanation from scratch.

SWALE is still in its formative stages but we are very excited by the preliminary results. The program possesses an important trait that previous understanders did not: When its knowledge structures fail, the program attempts to build new hypotheses for understanding the events and then stores these for future use. This is an important component of flexible, human-style understanding.

Acknowledgements

All of the other members of the SWALE team, Roger Schank, Chris Riesbeck, David Leake and Chris Owens helped develop the ideas in this paper. Chris Riesbeck and David Greenberg provided helpful comments on preliminary drafts.

References

[Charniak 72] Charniak, E., *Towards a Model of Children's Story Comprehension,* Technical Report 266, MIT Artificial Intelligence Lab, 1972.

[Cullingford 78] Cullingford, R., *Script Application: Computer Understanding of Newspaper Stories,* Ph.D. Thesis, Yale University, 1978. Research Report #116.

[Hammond 84] Hammond, K., *Indexing and Causality: The organization of plans and strategies in memory,* Technical Report 351, Yale University Department of Computer Science, December 1984.

[Kass, Leake and Owens 86] Kass, A. M. and Leake, D. B. and Owens, C. C., *SWALE: A Program that Explains,* 1986. In [Schank 86].

[Leake and Owens 86] Leake, D. B. and Owens, C. C., Oragnizing Memory for Explanation, *Proceedings of the Eighth Annual Conference of the Cognitive Science Society,* Cognitive Science Society, Lawrence Erlbaum Associates, 1986.

[Rieger 75] Rieger, C., Conceptual Memory and Inference, *Conceptual Information Processing,* North-Holland, Amsterdam, 1975.

[Schank 86] Schank, R.C., *Explanation Patterns: Understanding Mechanically and Creatively,* 1986. Book in press.

[Sussman 75] Sussman, G.J., *Artificial Intelligence Series,* Volume 1: *A computer model of skill acquisition,* American Elsevier, New York, 1975.

[Wilensky 78] Wilensky, R., *Understanding Goal-Based Stories,* Ph.D. Thesis, Yale University, 1978. Research Report #140.

COMPUTER STRATEGY OF

DECISION MAKING UNDER TIME PRESSURE *

Daniel LACOMBE

Département d'éducation physique, Université de Montréal

Claude SARRAZIN

Département d'éducation physique, Université de Montréal

Claude ALAIN

Département d'éducation physique, Université de Montréal

The study focuses on the cognitive processes underlying decision making in a dynamic context. The purpose is to formally define the predecisional structure and to describe the cognitive strategy a subject employs when solving a decision problem under time pressure. A model was elaborated defining the cognitive-decisional strategy of a defensive player (D) in squash when selecting a motor act in response to his opponent's eventual shot. (Sarrazin et al., 1983). Computer simulation based on protocol analysis was applied aiming at verifying the inner validity and logic of the proposed model. Results points to a viable predecisional information structure established on predefined methods used to reach a specific preparation state. These methods are sequences of goals and operators that are stored in property listed and are activated by transistory knowledge states in working memory. The computer program could also account for a substantial part of the variation in the length and accuracy of processing. Discrepancies observed between the program decisions and the ones reached by expert players led us to question the use of a normative decision rule in a dynamic context.

Previous research in the field of psychomotor learning and performance has dealt with either mechanisms governing movement or the perceptual factors involved in performing a motor task. It is interesting to note, however, that investigations into the processes underlying decision making in a context subject to time pressure, such as sport, has been virtually ignored (Whiting,

* We gratefully acknowledge the assistance of Robin Michel.

1979). This dearth or empirical research is probably attributable to methodological difficulties identifying cognitive processes in a fast-paced environment. These shortcomings have led us to explore the relevance of conceptual frameworks and methodological tools related to cognitive psychology (Einhorn & Hogarth, 1981; Kleinmuntz & Kleinmuntz, 1981) in gaining insight into the cognitive processes involved in choosing a motor act.

The study reported hereafter is formulated in light of problem solving theory and methodology (Newell & Simon, 1972). The decision maker is placed in a problem solving situation. He should be viewed as a symbolic information processing system employing certain cognitive operations as adapting mechanisms in order to cope with a complex environment (Hogarth, 1981). In addition, verbal protocols can provide significant information about deliberately selected processes underlying decision making (Kellog, 1982; Payne, 1982). Finally, it is possible to establish some functional equivalence between performance patterns on a computer and a human being during a given task (Card, Moran & Newell, 1980; Bhaskar & Simon, 1977).

Through the use of verbal protocols and computer simulation, the present study investigated the cognitive processes involved in solving decision problems under time pressure. The aims were to verify the inner validity of a proposed model of decision making in squash (Sarrazin & al. 1983) and to evaluate the extent in which this model accounts for the variability in speed and accuracy of the motor reaction.

Task analysis and decision making model

The decision task facing a defensive player (D) in squash was considered. It consisted in choosing the best preparation state in order to react appropriately to an opponent's impending shot. A preparation state is regarded

as a physical bias toward the chosen response. Previous results (Alain & al. 1983) revealed that expert squash players considered three different preparation state: 1) total preparation according to which D totally favors a unique response to be executed by his opponent without taking into consideration any other possible responses. 2) Partial preparation whereby D primes one response without excluding the possibility that an alternate reaction may be required. 3) Neutral preparation according to which D's bias is the same for each of the possible response.

An integral part of the simulation task was to provide a detailed description of the structural characteristics of the task environment. Data collection and verbal protocol analysis of D attending an opponent's shot in squash competition were completed. The methods and theoretical constructs underlying these stages are reported elsewhere by Alain & al. (1983) et Sarrazin et al. (1983). This led to the formulation of a model of D's decisional behavior.

In the course of reacting to repetitive decisions made in face of time pressure, an expert squash player will formally define a problem space. This problem space consists of a highly organised internal representation of the task environment. Operators and goals would then be applied to varying knowledge states in order to construct a search sequence within this space. A production system consisting of conditional statements spells out the logic guiding this search. It describes a set of methods (sequence of goals and operators) and a selection rule that D uses to choose the best possible state of preparation. This choice would be accomplished through the extention of the subjective expected utility principle (S.E.U.) (Coombs, Dawes &

Tversky, 1970). That is, D. would assign varying S.E.U. values to each preparation state and subsequently choose the one with the highest value. Assignment of these values is achieved by computing various sources of information related to the three following functions: 1) The subjective probability assigned to each possible shot of the offensive player. 2) The time pressure attributed to each shot, and finally 3) the utility value that D assigns to each preparation state he might choose. The various sources of information related to the task of choosing the best preparation state are, for instance, the respective positions of the players on the court, play habits of the offensive player, the opponent's ability to aim the ball to the chosen position, the estimated time required to reach the ball, the time available to reach the position the ball is expected to touch the court, and the significance or score of the game.

In sum, data collected thus far have led to the formulation of a conceptually viable model of decision making in squash. However in order to further corroborate the viability of this model, this study was designed to verify the inner logic. The goals, operators, methods and selection rule were integrated into a computer simulation program. In this regard the following sections outline how the information processing approach was applied to the study of decision making in sport.

Computer representation of a decision making problem

The basic structures of the simulation program were elaborated and operationalized on a PDP-10) computer using a recent version of UCI-LISP (Meehan, 1979). The resulting LISP program comprised a set of production rules governed by an adapted translation of Winston's (1977) production system interpreter. This interpreter is an internal aspect of processing which

TABLE 1

Production System

COMPUTER TRACE	PSYCHOLOGICAL STEPS
P_1 : Heuristic = First	P_1: FIND THE FIRST PREDICAT T OF AN ITERATION AND EVALUATE IT.
Itération 1 Prédicat = (EQ(Preparation-P)(Quote Note)) Action = Print solution Production = Stop	ITERATION 1 = IF D POSSESSES ALL THE NECESSARY INFORMATION TO CHOOSE A SPECIFIC STATE OF PREPARATION IN HIS PRODUCTION MEMORY THEN PRINT THE CHOICE OF A SPECIFIC STATE OF PREPARATION AND STOP THE PROCESSING.
Itération 2 Predicat = T Action = Nothing Production = P_2	ITERATION 2 = BY DEFAULT, THE PREDICAT IS T AND D SETS THE GOAL OF EVALUATING THE NEXT PRODUCTION
P_2 : Heuristic = First	P_2: FIND THE FIRST PREDICAT T OF AN ITERATION AND EVALUATE IT.
Itération 1 Predicat = (EQ(Preparation-P)(Quote Uncertain)) Action = Nothing Production = P_3	ITERATION 1 = IF D IS UNCERTAIN OF THE PREPARATION STATE TO CHOOSE THEN D SETS THE GOAL OF FINDING THE BEST POSSIBLE PREPARATION
P_9 : Heuristic = All	P_9 : EVALUATE EVERY ITERATION THAT HAS A PREDICAT T
Itération 1 Predicat = (EQ(POS-P)(Quote Desire)) Action =(DPOS) Production = Stop Itération 2 Predicat = (EQ(ORI-P)(Quote Desire)) Action = (DORI) Production = Stop Itération 3 Predicat = (EQ(ID-P)(Quote Desire)) Action = (ID) Production = Stop Itération 4 Predicat = (EQ(Habd-P)(Quote Desire)) Action = (DHABITUD) Production = Stop Itération 5 Predicat = (EQ(HABL-P)(Quote Desire)) Action = (DHABILET) Production = Stop Itération 6 Predicat = (OR(EQ(POSP)(Quote Note)) (EQ(ORI-P)(Quote Note))(EQ(ID-P)(Quote Note)) (EQ(HABD-P)(Quote Note))(EQ(HABL-P)(Quote Note))) Action = (ASPS) Production = Stop	ITERATION 1, 2, 3, 4, 5, 6 = THE SAME PRINCIPLE IS GOVERNING ALL OF THESE ITERATIONS IF D IS UNCERTAIN OF THE STATE OF PREPARATION TO CHOOSE AND D WANTS TO CHANGE OR TO GIVE A VALUE TO THE EXPRESSION POSITION (OR ORIENTATION, SHOT, HABITS, ABILITY) ON SET OF EXPRESSIONS UNCLEAR, THEN D APPLIES THE HIGHER LEVEL OPERATOR DPOS (OR DORI, ID, DHABITUD, DHABILET) AND STOP THE EVALUATION OF THIS ITERATION.

also consists of a lisp program. Simple rules verify the presence of certain antecedent conditions then initiate specific actions. For illustrative purposes, table 1 presents 3 production rules from the total set of productions which are outlined in the process of choosing a specific preparation state.

The complete set of production rules encompasses 8 general production rules which are broken down into 22 iterations. A production rule is defined as one or more iterations, an iteration being composed of a predicat-action-production triad. Each triad constitutes a specific strategy and the whole series of 22 iterations comprises D's global cognitive strategy. This strategy determines D's predecisional information structure and also contributes to the actual choice of a specific preparation state. The predicat is a simple function that verify whether or not the goal conditions and the stored problem characteristics from the property list are met. As such it essentially determines which iteration is to be evaluated by the interpreter. As soon as a predicat is dubbed true, the action associated with it is initiated.

The actions of an iteration are either inexistant (nothing), as in simply establishing a goal (iteration 2 of P1, table 1), or constitutes applications of operators involved in D's decision process, namely lower and higher level operators. Their respective definitions characterizes the grain of the anlysis. The lower level operators consist of the few basic processes used by D to generate, manipulate, store, or retrieve symbolic expressions from property lists. These property lists are simple pair lists (name-value) associated with an identifier placed in D's production memory. The basic processes define the higher level operator used by D to solve the decision

problem. Each higher level operator is applied to a knowledge state. This produces a new knowledge state which brings D closer to his choice of a specific preparation state.

Methods and task characteristics

A sequence of goals and higher level operators in the program represents a particular method of reaching a specific preparation state. The program's simplest method is the one associated with the goal of retrieving a preparation choice already located in production memory. However, if D does not have a pre-established preparation choice, then he must use a more elaborate sequence of goals and operators. The simulation trace in figure 1 illustrates the result of a complete assessment of iterations. In this case, after computing the differente sources of information for two specific shots, D selects according to the S.E.U. rule a total preparation for a pass shot.

Preference for one method over another depends on the specific information D possesses when he becomes a defensive player as well as other processing characteristics emerging from the context. For instance, figures 2 and 3 illustrate the influence of situation repetition on D's process duration. Figure 2 shows a relatively long processing time (production rules P1, P2..) attributable to D's engagement in the production system without a well defined internal representation of the task environment. Therefore, in order to make a choice the player must use more computational phases. Figure 3 illustrates a situation later on in the game. D's progressive involvement in the game has lead him to identify the offensive player's habits and abilities as well as other information pertaining to this situation. As a consequence, the values within certain property lists remains the same when facing repetitive decisions. Given this initial knowledge state, D

C. SIMULATION TRACE OF DECISION PROCESS

```
* (SETQ ACTIONFLG T)
* (SETQ %%TRACE¬P T)
T
* (EXE¬P P1)

Operator
  DPOS              = (10 . 6)

Operator
  DORI              = (6 . FACE)

Operator
  1D                = (BOAST PASS¬SHOT)

Operator
  DHABITUD          = ((BOAST 0.600)(PASS¬SHOT 0.799))

Operator
  DABILIT           = ((BOAST WEAK)(PASS¬SHOT STRONG)

Operator
  ASPS              = ((BOAST 0.357)(PASS¬SHOT 0.643))

Operator
  PR¬TEMP           = ((BOAST LOW)(PASS¬SHOT HIGH))

Operator
  ASPR              = ((BOAST (1.0 0.799 1.0 0.899 1.0))(PASS¬SHOT
                       (0.0 0.899 0.300 0.799 0.300)))

Operator
  DPOINT            = (2 2 1 1 1)

Operator
  DSTR              = (2 2 2 2 2)

Operator
  DFAT              = (1 1 1 1 1)

Operator
  ASVU              = (5 5 4 4 4)

Operator
  SEU               = (1.785 4.322 2.200 3.343 2.200)

Operator
  CHOICE            = SEU VALUE NO 2

TOTAL PREPARATION FOR BOAST       = 1.785
TOTAL PREPARATION FOR PASS¬SHOT   = 4.322
PARTIAL PREPARATION FOR BOAST     = 2.200
PARTIAL PREPARATION FOR PASS¬SHOT = 3.343
EQUAL PREPARATION                 = 2.200

Selected motor reaction is
TOTAL PREPARATION FOR PASS¬SHOT   = 4.322
```

Fig. 1. Simulation trace of D's decision process

A. INTERNAL REPRESENTATION OF TASK ENVIRONMENT

DEFENSIVE PLAYER:

POSITION	10
ORIENTATION	FACE
SCORE	8
TIREDNESS	EXHAUSTED
STRATEGY	FAKE
HABITS	((BOAST 0.600)(LOB 0.199)(DROP-SHOT 0.000) (PASS-SHOT 0.799))
ABILITY	((BOAST WEAK)(LOB WEAK)(DROP-SHOT WEAK) (PASS-SHOT STRONG))
TIME-PRESSURE	((BOAST LOW)(LOB LOW)(DROP-SHOT HIGH) (PASS-SHOT HIGH))
PS-EFFICACY	((TP AUGMENTED)(PP AUGMENTED)(EP DIMINISHED))
PS-ENERGY-COST	((TP NORMAL)(PP NORMAL)(EP EXHAUSTING))

ATTACKING PLAYER:

SCORE	2

B. COMPUTATIONAL PHASES

PRODUCTION RULES	ITERATIONS	PRODUCES
P1	2	(P2)
P2	1	(P3)
P3	2	(P4)
P4	1	(P5)
P5	1	(P9)
P9	1	(STOP)
P9	2	(STOP)
P9	3	(STOP)
P9	4	(STOP)
P9	5	(STOP)
P9	6	(STOP)
P5	2	(P15)
P15	1	(STOP)
P15	2	(STOP)
P5	3	(P17)
P17	1	(STOP)
P17	2	(STOP)
P17	3	(STOP)
P17	4	(STOP)
P5	4	(P3)
P3	1	(P1)
P1	1	(STOP)

Fig. 2

PRODUCTION RULES	ITERATIONS	PRODUCES
P1	2	(P2)
P2	1	(P3)
P3	2	(P4)
P4	1	(P5)
P9	1	(P9)
P9	2	(STOP)
P9	3	(STOP)
P9	4	(STOP)
P9	5	(STOP)
P9	6	(STOP)
P5	2	(P15)
P15	1	(STOP)
P15	2	(STOP)
P5	4	(P3)
P3	1	(P1)
P1	1	(STOP)

Fig. 3. Influence of situation repetition on D's decision process

needs to process less information and less computational phases. The length of the process also appears to increase according to uncertainty (Hick, 1952; Hyman, 1953). In the proposed strategy, increasing the quantity of information to process (for instance, more shots introduced in the problem space) will lead to computation of numerous productions, which in turn increase processing time.

Further analysis of program function have led us to question the use of the S.E.U. rule in the choice stage of the decision making process. According to the program, computation is characterized by a small sequence of dimensional processing in the assesment stage followed by a more complete holistic processing when selecting the best possible preparation stage (Walsten, 1980; Payne, 1982). In this regard, the normative S.E.U. rule used requires that D possesses all the available information on all alternatives and dimensions in order to compute S.E.U. values and subsequently choose a preparation state. This implies that D must be placed in an ideal situation where he feels no information overload and has the necessary time to process all related information. Under time pressure the vast majority of decision making will be made without a complete evaluation of the available information, therefore a normative rule is likely to be used in a dynamic context.

Furthermore, in examinating the structural organisation of the lower level operator involved in the choice rule used in this program, it also becomes apparent that the computational demands are considerable, time consumming, and presumably overloading short term memory. These claims that are suported by the results of Kleinmuntz & Kleinmuntz (1981) done while observing an extreme computational complexity of the bayesian strategy as opposed to

generate and test or heuristics strategies. Therefore, the intricate nature of computing S.E.U. values suggests that the choice stage is made under simple rules, such as lexicographic ordering (Fishburn, 1974), or through elimination by aspects (Tversky, 1972).

The above results emphasize the need to consider computational effort, demands on memory, and speed of execution when assessing a choice strategy in a fast-paced environment.

Alain, C., Lalonde, C., & Sarrazin, C. (1983). Decision making and information processing in squash competition. In H. Reider, K. Bos, H. Mechling et K. Reischle (Eds.), Motorik and Bewegunsforschung. Heidelberg: Verlag Karl Hofman Schorndorf, 196-202.

Bhaskar, R. & Simon, H.A. (1977). Problem solving in semantically rich domains: An example from engineering thermodynamics. Cognitive Science, 1, 193-215.

Card, K.S., Moran, T.P., & Newell, A. (1980). Computer text-editing: An information processing analysis of a routine cognitive skill. Cognitive Psychology, 2, 32-74.

Coombs, C.H., Dawes, R.M., & Tversky, A. (1970). Mathematical Psychology. Englewood Cliffs, New Jersey: Prentice Hall.

Einhorn, H.J., Hogarth, R.M. (1981). Behavioral decision theory: Processes of judgment and choice. Annual review of psychology, 32, 53-88.

Fishburn, P.C. (1974). Lexicographic order, utilities and decision rules: A survey. Management Science, 20, 1442-1471.

Hick, W.E. On the rate of gain of information. (1952). Quarterly Journal of Experimental Psychology, 4, 11-26.

Hogarth, R.M. (1981). Functionnal and dysfunctional aspects of judgmental heuristics. Psychological Bulletin, 90, 197-217.

Hyman, R. (1953). Stimulus information as a determinant of reaction time. Journal of Experimental Psychology, 45, 188-196.

Kellogg, R.T. (1982). When can introspect accurately about mental processse? Memory & Cognition, 10, 141-144.

Kleinmuntz, D.N., Kleinmuntz, B. (1981). Decision strategies in simulated environments. Behavioral science, 26, 294-305.

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

Payne, J.W. (1982). Contingent decision behavior. Psychological Bulletin, 92, 382-402.

Sarrazin, C., Lacombe, D., Alain, C. & Joly, J. (1983) Simulation study of a decision-making model of squash competition, phase one: the analysis of the protocol. Human movement science, 2, 279-306.

Tversky, A. (1972) Elimination by aspects: A theory of choice. Psychological Review, 19, 281-299.

Wallsten, T.S. (1980) Process and models to describe choice and inference behavior. In T.S. Wallsten (Ed.), Cognitive process in choice and decision behavior, Hillsdale (N.J.): Erlbaum.

Whiting, H.T.A. (1979) Subjective probability in sport. Dans C.R. Bell (Ed.) Uncertain Outcomes, Lancaster: MTP Press.

Winston, P. (1977) Artificial Intelligence. Reading, M.A. Addison-Wesley.

Organizing Memory for Explanation*

David B. Leake and Christopher C. Owens
Yale University

ABSTRACT

We present a mechanism for remembering explanations and re-using them to explain new episodes. This task requires a representation scheme for explanations, a dynamically organized memory, and a means of modifying old explanations to fit new facts. In this paper we focus on memory organization. We describe strategies for indexing and retrieving explanations, for using causal knowledge to select relevant features of episodes and for guiding generalization. We discuss work in progress on a computer implementation of this model.

INTRODUCTION

A task that people are able to perform, and one we have often required of Artificial Intelligence systems, is to make explanations. The ability to explain things has long been considered to be a yardstick for judging the depth to which a program has understood a situation. Yet recently we have begun to see explanation as more than just a task requiring understanding; we have begun to see explanation as an integral and necessary component of the understanding process itself [Schank 86].

In order to understand a situation, we must be able explain it to ourselves, that is, to connect it in a useful way with the rest of our knowledge. The type of explanation required for a given situation depends on the task which the explainer wishes to perform. For some AI programs, explanations relate facts to currently-active memory structures like scripts [Cullingford 78] or similar schemas guiding top-down processing. Such an approach explains a fact like "John left a tip" by, in effect, saying "because that is what typically happens in restaurant situations." Explanations can also relate actions to the goals they satisfy, as in the case of SHRDLU [Winograd 72], which could explain any of its actions by relating them back to the goals they satisfied. Explanation of less stereotyped information has required chaining together pieces of primitive causal knowledge into more elaborate explanatory chains. PAM [Wilensky 78] could relate an input fact to goals via known plans, for example explaining why a person would buy a gun if he intended to commit a robbery.

While building explanations by chaining together small pieces of causal knowledge increases flexibility, undirected backwards chaining causes a combinatoric explosion of connections to consider. Although not a problem in highly restricted domains, combinatoric explosion is rapidly aggravated by increasing knowledge. This violates our intuition that increasing knowledge and experience should facilitate building explanations, not make the process more difficult.

We (working jointly with Alex Kass) are developing a program to construct explanations of complicated real-world events, events for which people have no ready explanations. Making difficult explanations by building up connections from scratch would be overwhelming; we need to bring connections learned from relevant experience to bear on the process. To be able to retain these relevant experiences, we must have appropriate memory structures to store past explanations; for the information to be accessible, we must be able to form useful categories to organize them.

This paper describes a mechanism, implemented in our program, for categorizing explanations in

*This work is supported in part by the Air Force Office of Systems Research under contract 85-0343

memory, retrieving them where relevant, and for applying them to new situations.

NEW EXPLANATIONS FROM OLD

An episode for which we have been collecting anecdotal evidence is the death of the highly successful young racehorse Swale, who died one week after having won the prestigious Belmont Stakes horse race. Although few facts had appeared in the newspapers and although most people have little specialized knowledge to apply to explaining racehorse death, people were nevertheless able to quickly hypothesize a variety of plausible explanations.

One of these explanations was that Swale's death was like the ironic death of Jim Fixx, a prominent advocate of the health benefits of jogging. Fixx died while running, as a result of a congenital heart defect. According to this explanation Swale had a hidden heart problem that was exacerbated by the strain of his all-out effort to win the race.

Another explanation was that Swale had been killed by his owners to collect the insurance money. (This quickly turned out to be implausible when it was revealed that Swale had been under-insured.)

The questions raised by these explanations are: How did these people get reminded of Jim Fixx on the one hand and of the notion of insurance fraud on the other, how did they retrieve the explanations corresponding to these concepts, and how were they able to apply them to the circumstances of Swale's death?

Explanation and reminding are closely bound together. While [Schank 82] observed that one episode can remind people of another if the two have a thematically common explanation, we claim that the process can run in the other direction as well: that explanations can be constructed from a reminding. If, during the course of processing an episode, an understanding system can be reminded either of some other episode whose explanation is relevant or of an explanation template that has been used to explain similar situations in general, then the system can apply the old explanation to the task of understanding the new episode.

The knowledge structure we use for storing explanations is the Explanation Pattern (XP), initially outlined in [Schank 86]. XPs contain a template of the kind of episode they are designed to explain, with causal annotation identifying the connections between the features of the episode. When a near-miss explanation needs modification to fit the current situation, this causal structure is essential to the revision process. How problems with near-miss explanations are identified, and how the revision or "tweaking" of explanations is done, are beyond the scope of this paper; they are described in more detail in [Kass 86] and in [Kass, Leake and Owens 86].

Our explanation algorithm is as follows:

ANOMALY DETECTION Attempt to fit story into memory. If successful DONE; otherwise an anomaly has been detected.

XP SEARCH Search for an XP that can be applied to explain the anomaly.

XP ACCEPTING Attempt to apply XPs. If successful then skip to XP INTEGRATION.

XP TWEAKING If unable to apply XPs directly then attempt to tweak them into XPs that might apply better. If successful send the tweaked XPs back to XP ACCEPTING.

XP INTEGRATION If any results are accepted, integrate results back into memory making appropriate generalizations.

This paper focuses primarily on XP Search and XP Integration.

EXPLANATION MEMORY

A system that re-uses old explanations to understand new episodes must have a way of storing old explanations and a means of finding them when appropriate to help in understanding new situations. This memory must fail gracefully: if it can't find an explanation exactly suited to the current episode, the near-misses must serve as departure points for new explanation creation. Continuing the Swale example, explanation patterns about death of racehorses would obviously be relevant, as would explanation patterns about deaths of athletes, about deaths of the young and famous, about destructions of important income-producing properties, and about other bad things that have happened to racehorses. How can one find these explanation

patterns so that they might be proposed as candidates?

Although we agree with [Gentner and Landers 85] that access to relevant knowledge structures (in our case Explanation Patterns, in her case analogies) can be based upon surface features, we believe strongly that all features of an episode are not used equally in a search for applicable structures. An intelligent process must select the features to be used as indices into memory, and the search process must complement the processes by which explanation patterns have been indexed.

As an example of the latter kind of indexing process, here are some indexing strategies which can be used to decide how to store XPs in memory. We suspect that these indexing methods are a few among what will ultimately be many.

Indexing rule 1: Index an XP via features participating in the anomaly that the XP is designed to resolve. For example, Jim Fixx's death was surprising because his health seemed outstanding; the Jim Fixx XP described above can be indexed under the combination: (Death + Excellent Physical Condition)

Indexing rule 2: Index an XP via a feature playing a role in the chain of causation contained within the explanation. Using this strategy, the Jim Fixx XP could be indexed under: (Death + Heart Defect) or (Death + Jogging)

Indexing rule 3: Index an XP via any highly unusual feature of the episode that it originally explained, whether or not that feature played a causal role in the explanation. If you get a flat tire during a blizzard, another instance of car problems in blizzards may remind you to get your tires retreaded.

Indexing rule 4: Index an XP via features defining membership in a commonly-stereotyped group. (A group that has been previously defined for other purposes.) For example, one of the explanations we collected for Swale's death centered around other deaths of famous young star performers, with remindings of Jimi Hendrix and Janis Joplin. This explanation pattern is indexed under (Death + Successful Star Performer)

Once XPs have been indexed using the above methods, complementary retrieval strategies can be applied to new episodes in order to find relevant XPs. Following are some retrieval strategies associated with the above; again we expect this list to grow.

Retrieval Strategy 1: Consider directly-indexed XPs. We claim that for a given initial categorization of an event or for a given anomaly type, there are likely to be indexed a small number of immediate candidate explanation patterns addressing the situation. Under death, for example, we might find old age and sickness.

Retrievel Strategy 2: If an XP fails to fit, consider the features that have caused it to fail. If we try old age as an explanation of Swale's death, it fails because Swale was young. So, we look to see what we have indexed as explanations of early death. Similarly if we try death from sickness, we find a contradiction with Swale's excellent condition inferred from his recent racing victory. This leads us to examine explanation patterns indexed under death and excellent condition, which yields the Jim Fixx reminding.

Retrieval Strategy 3: Consider extreme or unusual features of the current episode. Swale, for example, was of great monetary value compared with the normative instance of a racehorse. Death and great monetary value can index a variety of obvious XPs, such as being killed for the insurance money.

GENERALIZING AND FORMING CATEGORIES

Within an XP memory, it's important to generalize explanation patterns into categories based on shared causal structure. When XPs are retrieved, near misses will be useful if causally-related XPs are stored under similar indices– conceptually near each other.

There initially seems to be a problem with circularity here: one reason for doing explanation is to select the relevant features for categorizing an episode, but these features are the ones needed in order to retrieve a correct XP. However, this apparent circularity is avoided by having static methods to initially select a set of candidate features for use as indices, even though these indices may retrieve irrelevant or wrong XPs. Once a set of XPs is available as a starting point, dynamic methods

modify them into final explanations embodying a reasonable set of features from the episode.

Early attempts to organize a memory full of knowledge structures used straightforward inductive category formation. An example of this approach was IPP [Lebowitz 80]. It read newspaper stories about terrorist attacks, using MOPs [Schank 82] to provide expectations during the story understanding process. By extracting features shared across multiple stories, the program formed generalized MOPs representing the features common to those episodes. These new categories could then be used in understanding future stories. This method is a realistic approximation to one kind of generalization people actually do. For example, from reading newspaper stories about Italy when the Red Brigades were active, IPP formed the generalization that the usual victims of kidnapping in Italy were businessmen.

PROBLEMS WITH INDUCTION

But systems using the type of induction described above have a number of problems (which are discussed further in [Schank, Collins and Hunter 86]). Inductive learning systems depended upon having multiple episodes available. Their idea was that, as many episodes were processed, features resulting from presumably randomly-distributed noise would tend to cancel each other out, leaving generalizations containing only important features. But looking only at frequency of feature appearance has two problems. One is that people can in fact form generalizations in one trial, which is not accounted for by a feature frequency method. The second is that, without any metric of feature importance, induction learning systems, when faced with small numbers of episodes, tend to form silly generalizations, for example: "Terrorist attacks in India always kill exactly two people." People form erroneous generalizations, but not of this variety.

Furthermore, the very notion of saving common features across multiple episodes depends on a representation scheme in which some features are clearly part of the episode and others are not. One does not want to take a complete description of the state of the world at the time of two separate episodes and generalize all the features that were true at both times. Instead, one wants to take two episode representations and generalize the features they have in common. Limiting the feature space this way is a reasonable strategy for certain domains, particularly technical reasoning where the number of factors is known and limited. But for general-purpose understanding, limiting the representation scheme to one in which each episode has only a few features results in an unnaturally sparse and impoverished knowledge base. A key component of the understanding task is focusing attention: deciding, among all the things that happen to be true of the world at a particular time, which of them are reasonably part of a given episode and which are not. A representation that tries to solve this a priori does not help much towards a psychologically interesting theory of understanding, nor does it have the potential to ultimately yield sophisticated automated reasoners.

USING CAUSAL LINKS

Within our model, the causal annotation of XPs is crucially important to the tasks of modifying and generalizing explanations. When we try to relate the Jim Fixx explanation to Swale's death, we don't simply consider every aspect in which Jim Fixx and Swale were similar. Instead, we use the causal structure of the Jim Fixx explanation to see that the relevant fact about Jim Fixx was that he did regular strenuous exercise.

In our search through aspects of Swale, we don't spend any search time considering, for example, that Jim Fixx was also an author or that joggers typically get up early in the morning. Those facts are present and could be accessed by some different XP, but because they are not causally connected with this particular view of Jim Fixx's death they do not participate in this attempt to explain Swale's death. However, if we tried to apply an XP about ironic deaths of famous people to Jim Fixx, then the fact that he wrote books advocating jogging would certainly be considered in any attempt to generalize his death with other events (such as the death of the famous natural foods advocate who reputedly died of stomach cancer.) XPs provide a specific, causally connected view of a class of events; trying to apply an XP to a new event tightly constrains the features that will be considered as candidates for inclusion in a description and explanation of the new event.

Once we have accepted a modified XP as the explanation of an event, we attempt to generalize the old

and revised XPs in order to form a category that subsumes both explanations. Causal links within the original explanations constrain the generalization process: any generalized condition must support the same causal chains as the conditions it subsumes. Our explanation of Fixx's death differs from that of Swale since it depended on the actor jogging, while Swale's depended on his running in horse races. To generalize these premises, we need to find not only an abstraction they share (such as both being participation in outdoor activities) but one that fits each causal pattern. Since physical exertion is the abstraction that satisfies this requirement, it is the generalization selected. Our new explanation is that an actor in apparently good physical condition may die if he does a lot of exertion and has a hidden heart defect.

Thus in our system, the formation of generalizations is driven by the need to accomodate new events in memory. There is a difference between the knowledge which is accessible to the system in the form of basic causal knowledge, and the more complex patterns which our program applies to new situations. Similarly we maintain the difference between knowledge that is simply present in the system and knowledge that is accessible under a particular index. Rather than trying to generalize an explanation pattern as far as possible without regard to whether the generalization is needed to deal with our domain, we generalize an XP only when it can no longer accomodate the data.

CONCLUSIONS

The mechanism we have described allows an understander to retrieve and apply relevant old explanations to new situations, and to use information from the explanations to guide its categorization of episodes in memory. This process accounts for certain kinds of concept formation in small numbers of trials, and avoids some of the faults of straight induction. It also avoids some of the difficulties faced by other explanation systems (e.g., [Segre and Dejong 85]) in that the understander is not required to build up an explanation from scratch each time a situation is encountered. Instead, our system can resolve new problems more efficiently as its library of explanation patterns grows with experience.

Our theory of explanation bears on focus of attention in a situation: we suggest some strategies for selecting features to use as indices when searching memory for relevant patterns. Even when these strategies yield no directly-applicable pattern, our system fails gracefully in that it can use a near-miss pattern as the starting point for building a new explanation. Our system is still in its preliminary stages; further work on explanation memory will involve refining and extending the strategies used to search for XPs and to evaluate whether a given XP is satisfactory.

ACKNOWLEDGEMENT

The work described in this paper was done jointly by the authors, Alex Kass, Roger Schank and Christopher Riesbeck, all at Yale. We would like to thank Chris Riesbeck for his helpful comments on drafts of this paper.

REFERENCES

[Cullingford 78] Cullingford, R., *Script Application: Computer Understanding of Newspaper Stories,* Ph.D. Thesis, Yale University, 1978. Research Report #116.

[Gentner and Landers 85] Gentner, D. and Landers, R., Analogical Reminding: A Good Match is Hard to Find, *Proceedings of the IEEE 1985 International Conference on Systems, Man and Cybernetics,* IEEE, 1985, pp. 607ff.

[Kass, Leake and Owens 86] Kass, A. M. and Leake, D. B. and Owens, C. C., *SWALE: A Program that Explains,* 1986. In [Schank 86].

[Kass 86] Kass, A. M., Modifying Explanations to Understand Stories, *Proceedings of the Eighth Annual Conference of the Cognitive Science Society,* Cognitive Science Society, Lawrence Erlbaum Associates, 1986.

[Lebowitz 80] Lebowitz, M., *Generalization and Memory in an Integrated Understanding System,* Ph.D. Thesis, Yale University, October 1980.

[Schank, Collins and Hunter 86] Schank, R.C. and Collins, G. and Hunter, L., *Transcending Inductive Category Formation in Learning,* Behavioral and Brain Sciences, 9/4 (1986). In Press.

[Schank 82] Schank, R.C., *Dynamic Memory: A Theory of Learning in Computers and People,* Cambridge University Press, 1982.

[Schank 86] Schank, R.C., *Explanation Patterns: Understanding Mechanically and Creatively,* 1986. Book in press.

[Segre and Dejong 85] Segre, A. M. and DeJong, J., Explanation-Based Manipulator Learning: Acquisition of Planning Ability through Observation, *Proceedings of the IEEE 1985 International Conference on Robots and Automation,* IEEE, 1985, pp. 555ff.

[Wilensky 78] Wilensky, R., *Understanding Goal-Based Stories,* Ph.D. Thesis, Yale University, 1978. Research Report #140.

[Winograd 72] Winograd, T., *Understanding Natural Language,* Academic Press, New York, 1972.

A NEURAL NETWORK THEORY OF FRONTAL LOBE FUNCTION

Daniel S. Levine
Department of Mathematics
University of Texas at Arlington
Arlington, TX 76019

INTRODUCTION

The frontal cortex is six-layered only in primates and is the neocortical area best connected to the hypothalamus. For these reasons and many others, Fuster (1980, p. 144) stated:

> "The central notion...is that the prefrontal cortex plays a role in the temporal structuring of behavior. The prefrontal cortex is thought to be essential for the synthesis of cognitive and motor acts into purposive sequences."

This article attempts to integrate this qualitative notion with existing neural network theories of motivation and cognition.

Grossberg (1975) discusses the striving for balance between two subsystems in a network. The attentional system seeks stable response to fluctuating sensory cues by focusing attention on important subclasses of cues. The arousal system enables adaptation to unexpected events and new reinforcement contingencies. Frontal lesions often change the balance between attention and arousal.

REVIEW OF BEHAVIORAL RESULTS

Delay Tasks, Perseveration, and Novelty

In delayed response (Jacobsen, 1935), an animal first sees food placed under one of two or more identical covers. After one-half to two minutes in which all covers are hidden, the animal must choose which cover to lift for food. Intact chimpanzees, monkeys, dogs, and cats perform this task easily, but frontally lesioned primates perform it poorly.

The delayed response deficit does not reflect short-term memory loss. Konorski and Lawicka (1964) found that most delayed response errors of frontally lesioned dogs involved perseveration of the response made on the previous trial, indicating that memory of cues had not been abolished. Interfering tasks between trials weakened perseveration.

Frontal monkeys also perform poorly on delayed alternation (Stamm, 1964) and delayed matching to sample (Spaet and Harlow, 1943). In delayed alternation, food is placed, concealed from an animal, alternately under the left and right of an identical pair of containers, and each time the animal must look again for food after a delay. Frontal monkeys tend to repeatedly choose one container that was once rewarding, even in the face of errors. In delayed matching to sample, the animal is first presented with a "sample" or visual stimulus, then after a delay is presented with a configuration of stimuli that includes the original one. The animal is then rewarded for choosing the sample correctly.

These results suggest that perseveration is a general consequence of frontal lesions. Milner (1964) confirmed this idea by asking frontal-lobe patients to sort cards based on any one of three criteria (color, shape, or number shown on the card). The patients were not told which criterion was correct, but at each placement were told whether their choice was right or wrong. Frontal patients initially deduced the correct strategy. However, when the experimenter changed the criterion, the patients preserved their earlier strategy. In the same vein, frontal patients asked to draw, in succession, a cross, two circles, and a triangle often draw four crosses instead (Luria and Homskaya, 1964).

In spite of perseverative tendencies, frontally damaged animals show increased preference for novel stimuli over familiar ones, regardless of reward value. Pribram (1961) gradually increased the number of objects. When a new object was introduced the peanut was placed under it. Frontal animals showed less tendency than normals to perseverate their choice of the object under which the peanut had been previously placed.

EEG Data

Walter (1964) and Walter et al. (1964) recorded a negative-going potential shift, the contingent negative variation (CNV), in humans anticipating a motor response. The CNV originates in the frontal lobes and spreads thence to other areas of neocortex. A similar potential change, also of frontal origin, accompanies a rhesus monkey's anticipation (Donchin et al., 1971).

In normal subjects, verbal instructions to await a visual or tactile signal cause enhancement of potentials the signal later evoked in the corresponding sensory cortex (Luria, 1969). Frontal patients, however, lack this potential change.

Dorsal Versus Ventral Frontal Cortex

The dorsal part of the frontal cortex has reciprocal connections with secondary sensory cortices. The ventral (or orbital) part has reciprocal connections (some via the mediodorsal thalamus) with the hypothalamus and limbic system. Hence:

> "...lesion studies indicate that the cortex of the dorsal and lateral prefrontal surface is primarily involved in cognitive aspects of behavior. The rest of the prefrontal cortex, medial and ventral appears to be mostly involved in affective and motivational functions...." (Fuster, 1980, p. 74).

Nevertheless, dorsal and ventral regions are extensively interconnected. This article will view these two areas as parts of a system, one part primarily motivational and the other cognitive, but both related to goal-directed behavior.

THE MODEL: MOTIVATIONAL ASPECTS

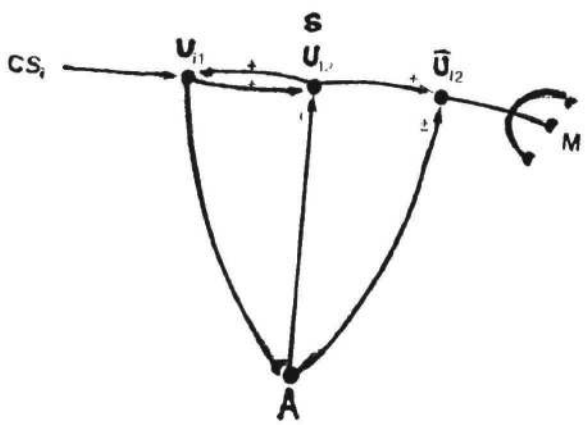
Both perseveration and enhanced novelty reaction can be seen as parts of deficit in drive-related incentive motivation. In other words, the frontal lobes serve to bias the organism's behavior toward actions that have current reward value, as opposed to actions that were once rewarding and have become motor habits, or actions that are exploratory in purpose. We shall now review how attention and arousal interact in some model neural networks.

Grossberg (1971) discussed the synchronization problem of classical conditioning, how the conditioned stimulus (CS) and an unconditioned stimulus (US) can become associated even when those stimuli are presented with different time lags on different trials. The solution of this problem involved "arousal" cell which include drive representations. Also, to permit secondary conditioning, it was found necessary to have two sensory representation stages for each stimulus.

Figure 1 reviews a general network, variants of which appear in Grossberg (1971, 1975, 1982) and Levine (1983). In Figure 1, the i th conditioned stimulus CS_i excites the cell population U_{i1} of its representation. Sensory representations are denoted generically by S. After receiving the CS_i input, U_{i1} sends signals to stage U_{i2} of the i th sensory representation and to all the arousal representations.

FIGURE 1

General Network for Classical or Operant Conditioning



Semicircles denote modifiable, arrows non-modifiable synapses. A is unconditionally activated by US, becomes activated by CS. Excitatory ("+") synapses from A to U_{i2} to U_{i1} lead to selective attention to stimuli conditioned to positive or negative arousal. Excitatory or inhibitory ("±") synapse from A to U_{i2} leads to enhancement of movement by stimuli conditioned to positive arousal and suppression of movement by stimuli conditioned to negative arousal. (Modified from Levine, 1983; reprinted with permission from Elsevier Science Publishing Company, Inc.).

The arousal representations (A in Figure 1) include, for example, A_h which subserve hunger and A_f which subserve fear. A given arousal population sends signals back to level U_2 of S only if it receives a large sensory input from level U_1 and a large drive input (such as hunger level or electric shock level). The synapse $U_{i1} \rightarrow A$ is always strong for an unconditioned stimulus, and is strengthened during learning for a conditioned stimulus.

Suppose that a hungry animal lifts a given cup, causing food to appear. Then A_h is excited and creates a positive $A_h \rightarrow S$ signal to all recently active sensory representations, such as those of the cup (S_c) and of proprioceptive feedback from the lifting response (S_{ir}). The U_{i2} stages of S_c and S_{ir}, having received U_{i1} and A_h inputs, can fire and send signals to M. The $A_h \rightarrow S_{ir}$ connection supplies positive incentive motivation for the motor act of lifting the cup.

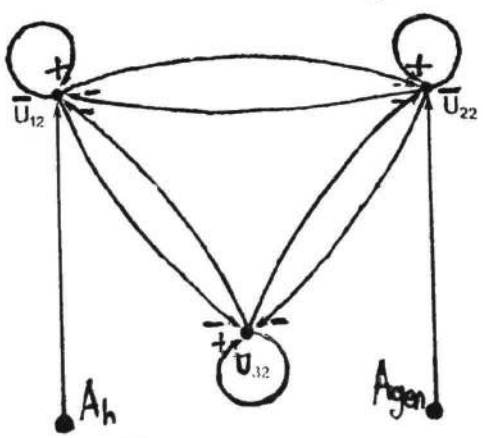
If lifting the cup leads to shock instead of food, then inhibitory $A_f \rightarrow S_{ir}$ connections supply negative incentive motivation for cup lifting, which can suppress $S_c \rightarrow M$ and $S_{ir} \rightarrow M$ firing. Negative incentive motivation can also come from frustration if expected reward is absent (Grossberg, 1975).

A CS conditioned to a given drive activates $A \rightarrow \bar{U}_{i2}$ positive incentive motivations, enabling signals from U_{i2} to M. Such signals influence $U_{i2} \rightarrow M$ synaptic habit strengths. Habit strengths are also influenced, less strongly, by repeated performance of a motor act even without current reward. Finally, habits can be influenced by the reward value of novelty (Berlyne, 1969). Response-contingent change (whether up or down) in light intensity in a rat's cage can reinforce bar pressing. Grossberg (1975) explained this effect using a nonspecific arousal locus (A_{gen}) that excites all the drive representations (A_j in Figure 1).

Figure 2 shows three stimulus representations, $\bar{U}_{12}$ which is excited by drive-specific incentive motivation because the stimulus is conditioned to that

FIGURE 2

Competition Between the Representations $\bar{U}_{i2}$ of Three Conditioned Stimuli



$\bar{U}_{12}$ is excited by reward, $\bar{U}_{22}$ by novelty, $\bar{U}_{32}$ by habit.

drive, $\bar{U}_{22}$ which is excited by nonspecific arousal because the stimulus is novel, and $\bar{U}_{32}$ which is initially active because the stimulus is one to which the animal has developed the habit of going. The representations of these objects are translated into target motor patterns via the $\bar{U}_{i2} \rightarrow$ M connection of Figure 1. Self-excitation and mutual inhibition between the $\bar{U}_{i2}$ creates competition for short-term storage.

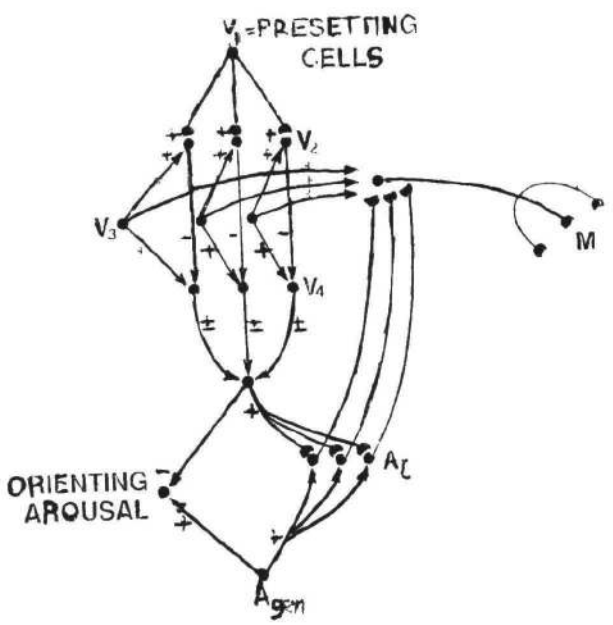
Figure 3 shows how unexpected events excite and expected events inhibit nonspecific arousal. The nonspecific arousal source is reminiscent of reticular areas, and the presetting cells, perhaps, of cerebellar areas.

My hypothesis that the frontal lobes are part of a major incentive motivational pathway is consistent with known anatomy. The A cells are reminiscent of drive-related areas of the hypothalamus. The frontal cortex is the only neocortical area known to have reciprocal monosynaptic connections with the lateral and preoptic hypothalamus (Nauta, 1971).

Thus a frontally lesioned animal can learn a response that leads to food reward, since some hunger-related incentive motivation still exists. Once that response has been established, however, even if reward ceases, perseveration occurs because incentive motivation for a competing response is weakened. Also, negative incentive motivation from frustration, which would normally occur when food is no longer found, is diminished.

FIGURE 3

A Network Where Expected but not Unexpected Patterns Inhibit Orienting Arousal



Activities of presetting cells represent a stored expected pattern. (Modified from Levine, 1983; reprinted with permission from Elsevier Science Publishing Company, Inc.)

If the new response involves a novel stimulus, perseveration is overcome by novelty. The approach to a novel object is stronger than the motor habit of approaching the familiar object. Also, the hunger-related incentive motivation exciting approach to the familiar object is weaker than in normal monkeys.

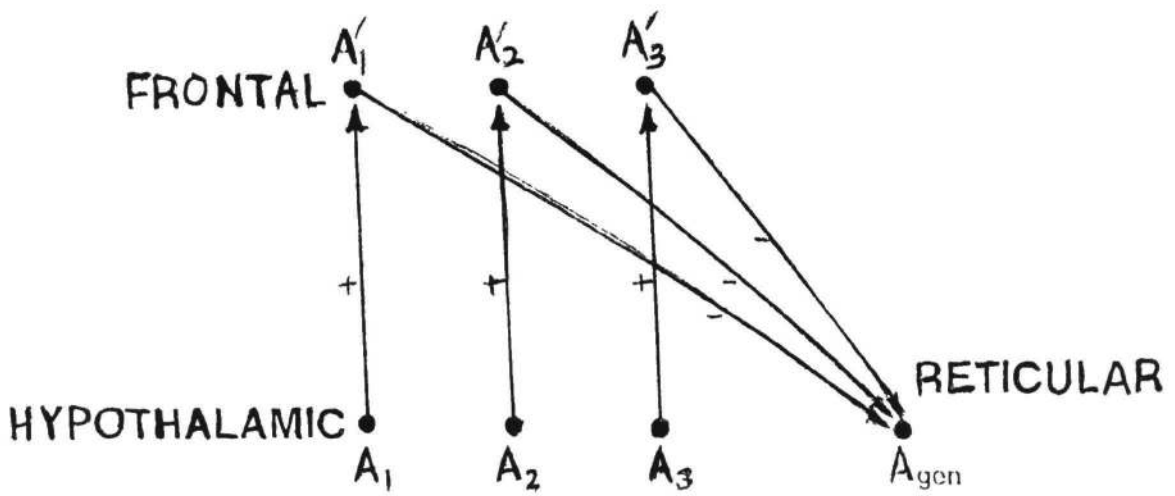
These results suggest a physiological prediction illustrated in Figure 4. Each "hypothalamic" A cell locus excites its own "frontal" representation A′, which inhibits the "reticular" nonspecific arousal locus. Frontally lesioned animals have trouble suppressing orienting reactions (Fuster, 1980, p. 61), which also supports this hypothesis.

THE MODEL: COGNITIVE ASPECTS

Weakening of incentive with frontal lesions is amplified by loss of neural preparation for expected sensory consequences of movement. Such preparation arises from the coding of sequences that include representations of stimuli, responses, and reinforcements in particular orders. It includes corollary discharge (Teuber, 1964), the compensation that the retina makes for eye movements.

FIGURE 4

Hypothesis for Frontal Participation in Incentive Motivation



The contingent negative variation accompanies expectation of one stimulus S_2 while another stimulus S_1 is present. This wave therefore depends on internal representations for S_1 and S_2 separately and for the temporal sequence S_1S_2. Further evidence that such sequences, and longer ones, are represented in frontal cortex is that frontally lesioned monkeys are easily distracted from sequential tasks (Grueninger and Pribram, 1969).

Grossberg (1978) discussed coding of "higher-order chunks" longer sequence representations in short-term memory models. He stated (p. 325) the following rule:

> "Self-Similar Coding Rule: Other things being equal, higher-order chunks have greater STM activity and longer duration than lower-order chunks."

This rule promotes goal-directed behavior, because longer stimulus sequences predict events better than shorter sequences.

The behavioral data suggest that the self-similar coding rule occurs particularly at the frontal cortex. Electrophysiological results from the dorsal (Fuster et al., 1982) and in the ventral frontal cortex (Rosenkilde et al., 1981). Both frontal areas in monkeys contain different types of cells whose activities change in correlation with each event in a delayed matching to sample sequence (sample/cue, choice stimuli, instrumental response, reinforcement). Some cue-sensitive cells respond to particular cue features such as color or location. Moreover, cells with similar properties may be organized into columns.

The dorsolateral frontal area known as the frontal eye field also shows variety in cell responses. Some cells in this area of monkeys discharge during but not before eye movement (Bizzi, 1968 and Bizzi and Schiller, 1970). Other frontal eye field cells fire before saccades, falling into three categories:

> "Visual activity occurred in response to visual stimuli whether or not the monkey made saccades. Movement activity preceded purposive saccades, even those made without visual targets. Anticipatory activity preceded even the cue to make a saccade if the monkey could reliably predict what saccade he had to make" (Bruce and Goldberg, 1985,p. 603).

The self-similar coding rule can best be understood by considering how short-term memory biases can develop in the selective coding of features. That issue was studied by Grossberg and Levine (1975). Their network was an on-center off-surround field, that is, each population excited itself and inhibited the others. The network (without biases) had been introduced by Grossberg (1973) to explain how noise can be suppressed and significant parts of a pattern contrast-enhanced. The activities x_i of the populations satisfied a system of differential equations of the form

$$dx_i / dt = -Ax_i + (B_i - x_i)(f(x_i) + I_i) - x_i \sum_{k \neq i} f(x_k) \qquad (1)$$

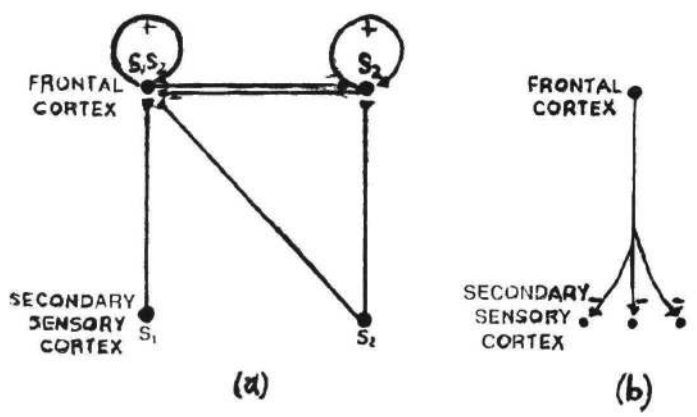
where I_i are outside inputs, f is a monotone increasing function reflecting averaged neuronal input-output transformations, and B_i denotes maximum possible

activity. If B_i is interpreted as number of cell sites, self-excitation of a population is proportional to number $B_i - x_i$ of "inactive sites," and inhibition of a population by the others is proportional to number x_i of "active sites."

Grossberg and Levine (1975) discussed how populations with larger B_i tend to suppress activities of populations with smaller B_i. Differences in B_i often arise from developmental and attentional biases. It is consistent that populations coding longer temporal sequences should have higher B_i values, perhaps reflecting more inputs from an earlier processing stage.

I conjecture that the dorsal frontal cortex contains on-center off-surround fields of populations coding chunks of all orders (see Figure 5a). Frontal afferents could also influence other on-center off-surround fields at the sensory cortices themselves (Figure 5b). The network of Grossberg (1973) has a quenching threshold, that is, an intensity below which stimuli are suppressed. Quenching threshold is lowered by tonic inhibition, leading to sharper decisions between stimuli. Hence I also predict that dorsal frontal cortex tonically inhibits secondary sensory cortex, thereby increasing the masking of irrelevant stimuli by relevant ones.

FIGURE 5



A) Mechanism for self-similar coding. Frontal representation for sequence S_1S_2 receives inputs from more populations than does representation for stimulus S_2 alone. Also, in frontal on-center off-surround field, larger B_i in equation (1) causes bias in favor of S_1S_2 (as represented by darker self-excitatory arrow). B) Tonic inhibition supplied to secondary sensory cortex by frontal cortex.

THE BEHAVIORAL PICTURE

The frontal cortex integrates sensory information from the neocortex with visceral information from the hypothalamus and limbic system. Hence frontal damage leads to "interoceptive agnosia" (Nauta, 1971,p. 182), including distractibility, lack of foresight, and inappropriate behavior. Frontal patients have been reported, for example, to urinate in public or tell off-color jokes at a funeral. Such behavior suggests disconnection between "reptilian" (instinctual), "old mammalian" (emotional), and "new mammalian" (rational) brains (MacLean, 1964).

Interfacing between the "three brains" seems to depend on orbito-dorsal connections within frontal cortex. Little is known about the structure of such connections. The flexibility of motivational responses and the known connections of dorsal frontal cortex with other neocortical areas and orbital frontal cortex with limbic areas suggest that orbito-dorsal connections should be nonspecific and modifiable in both directions. This should facilitate motivationally-based decisions between competing long-term plans.

REFERENCES

Berlyne, D. E. (1969). The reward-value of indifferent stimulation. In: Reinforcement and Behavior, J. T. Kapp, Editor, Academic Press, pp. .

Bizzi, E. (1968). Discharge of frontal eye field movements during saccadic and following eye movements in unanesthetized monkeys. Experimental Brain Research 6:69-80.

Bizzi, E. & Schiller, P. H. (1970). Single unit activity in the frontal eye fields of unanesthetized monkeys during eye and head movement. Experimental Brain Research 10:151-158.

Bruce, C. J. & Goldberg. M. E. (1985). Primate frontal eye fields. I. Single neurons discharging before saccades. Journal of Neurophysiology 53:603-635.

Donchin, E., Otto, D., Gerbrandt, L. K. & Pribram, K. (1971). While a monkey waits: Electocortical events recorded during the fore-period of a reaction time study. Electroencephalography and Clinical Neurophysiology 31:115-127.

Fuster, J. M. (1980). The Prefrontal Cortex; Anatomy, Physiology, and Neuropsychology of the Frontal Lobe. Raven Press.

Fuster, J. M. (1985). The prefrontal cortex and temporal integration. In: Cerebral Cortex, E. G. Jones and A. Peters, Editors, Volume 4, Plenum Press, pp. 151-177,

Fuster, J. M., Bauer, R. H. & Jervey, J. P. (1982). Cellular discharge in the dorsolateral prefrontal cortex of the monkey during cognitive tasks. Experimental Neurology 77:679-694.

Goldberg, M. E. (1980). Cortical mechanisms in the visual instantiation of movement. Experimental Brain Research 41:A32-A33.

Grossberg, S. (1971). On the dynamics of operant conditioning. Journal of Theoretical Biology 33:225-255.

Grossberg, S. (1973). Contour enhancement, short-term memory and constancies in reverberating neural networks. Studies in Applied Mathematics 52:217-257.

Grossberg, S. (1975). A neural model of attention, reinforcement and discrimination learning. International Review of Neurobiology 18:263-327.

Grossberg, S. (1978). A theory of human memory: Self-organization and performance of sensory-motor codes, maps, and plans. Progress in Theoretical Biology 5:233-374.

Grossberg, S. (1982). Processing of expected and unexpected events during conditioning and attention: A psychophysiological theory. Psychological Review 89:529-572.

Grossberg, S. & Levine, D. S. (1975). Some developmental and attentional biases in the contrast enhancement and short-term memory of recurrent neural networks. Journal of Theoretical Biology 3:341-380.

Grueninger, W. E. & Pribram, H. H. (1969). Effects of spatial and non-spatial distractors on performance latency of monkeys with frontal lesions. Journal of Comparative and Physiological Psychology 68:203-209.

Hirsch, H. V. B. & Spinelli, D. N. (1970). Visual experience modifies distribution of horizontally and vertically oriented receptive fields in cats. Science 168:869-871.

Jacobsen, C. F. (1935). Functions of the frontal association area in primates. Archives of Neurology and Psychiatry 33:558-569.

Konorski, J. & Lawicka, W. (1964). Analysis of errors of prefrontal animals on the delayed-response test. In: The Frontal Granular Cortex and Behavior, J. M. Warren and K. Akert, Editors, McGraw-Hill, pp. 271-294.

Levine, D. S. (1983). Neural population modeling and psychology: A review. Mathematical Biosciences 66:1-86.

Luria, A. R. (1969). The origin and cerebral organization of man's conscious action. Evening lecture at The XIX International Congress of Psychology, London. Moscow University Press.

Luria, A. R. & Homskaya, E. D. (1964). Disturbance in the regulative role of speech with frontal lobe lesions. In: The Frontal Granular Cortex and Behavior, J. M. Warren and K. Akert, Editors, McGraw-Hill, pp. 353-371.

MacLean, P. D. (1964). Man and his animal brains. Modern Medicine (February 2), pp. 95-106.

Milner, B. (1964). Some effects of frontal lobectomy in man. In: The Frontal Granular Cortex and Behavior, J. M. Warren and K. Akert, Editors, McGraw-Hill, pp. 313-334.

Nauta, W. J. H. (1971). The problem of the frontal lobe: A reinterpretation. Journal of Psychiatric Research 8:167-187.

Pribram, K. H. (1961). A further experimental analysis of the behavioral deficit that follows injury to the primate frontal cortex. Experimental Neurology 3:432-466.

Rosenkilde, C. E., Bauer, R. H. & Fuster, J. M. (1981). Single cell activity in ventral prefrontal cortex of behaving monkeys. Brain Research 209:375-394.

Spaet, T. & Harlow, H. F. (1943). Problem solution by monkeys following bilateral removal of the prefrontal areas. II. Delayed reaction problems involving use of the matching-to-sample method. Journal of Experimental Psychology 32:424-434.

Stamm, J. S. (1964). Retardation and facilitation in learning by stimulation of frontal cortex in monkeys. In: The Frontal Granular Cortex and Behavior, J. M. Warren and K. Akert, Editors, McGraw-Hill, pp. 102-125.

Teuber, H.-L. (1964). The riddle of frontal lobe function in man. In: The Frontal Granular Cortex and Behavior, J. M. Warren and K. Akert, Editors, McGraw-Hill, pp. 410-444.

Walter, W. G. (1964). Slow potential waves in the human brain associated with expectancy, attention and decision. Archiv fuer Psychiatrie und Nervenkrankheiten 206:309-322.

Walter, W. G., Cooper, R., Aldridge, V. J., McCallum, W. C. & Winter, A. L. (1964). Contingent negative variation: An electric sign of sensori-motor association and expectancy in the human brain. Nature 203:380-384.

Views From a Kill

James H. Martin

Berkeley Artificial Intelligence Research
Computer Science Division
University of California, Berkeley

ABSTRACT

Metaphor is a problem for natural language knowledge acquisition systems. Experts will make utterances based upon domain metaphors which the acquisition system may not possess. An approach is presented which uses knowledge about previously understood metaphors to process new uses. This approach is contrasted with several formal proposals for metaphor understanding which do not use explicit knowledge about metaphors. A system for representing metaphorical knowledge, as part of a general knowledge representation language, has been built. A knowledge acquisition system, UCTeacher, is described which can process newly encountered metaphors using knowledge of the domain and explicit knowledge about how similar metaphors have been used before. A detailed example from the system is presented.

Introduction

Metaphors are a widespread phenomena in natural language. The vast majority of metaphors are conventional parts of the language and are easily understood. Explicit knowledge of conventional metaphors is what makes these metaphors so easy to understand. Metaphor becomes a problem only when the hearer does not already have knowledge of the underlying metaphor that the utterance is based on.

This is exactly the situation faced by UCTeacher. UCTeacher is a natural language knowledge acquisition system for the the UNIX Consultant (Wilensky 1984). UC is a knowledge based consultant system that answers users questions about the UNIX operating system. Experts on UNIX can use UCTeacher to give more knowledge to UC simply by telling it the new information in English. One major problem for UCTeacher is learning new extensions to known metaphors during the knowledge acquisition task. For more information on other aspects of knowledge acquisition and UCTeacher see (Martin, 1985).

Consider the following examples from the UNIX domain:

1) You can kill a process by typing '^C'.

2) You can get into lisp by typing 'lisp' to the shell.

3) To leave the mail program type 'exit'.

* This research was sponsored in part by the Defense Advance Research Projects Agency (DOD), Arpa Order No. 4031, Monitored by Naval Electronic System Command under Contract No. N00039-C-0235, and by a GTE Laboratories Fellowship.

4) Run a file through the spell program to check for spelling mistakes.

Each of these examples contains a metaphor which is a specialization of a very general metaphor as applied to the concept COMPUTER PROCESS: example one involves viewing a process as an active agent that can be killed, examples 2 and 3 involve the metaphor that an interactive computer process is an environment that one can enter and leave, example 4 is an instance of a conduit/pipe metaphor. Experts in the domain of interest will frequently use such metaphors when giving new information to the system. The problem faced by UCTeacher is to find a way to understand these utterances given the fact that it does not yet possess the metaphors underlying them.

Our Approach

Knowledge about previously understood conventional metaphors is used directly in learning the new use of an old metaphor. Take examples 2 and 3 from above. At first it is not clear how the terms 'get into' and 'leave' can be applied to UNIX programs. The system has no knowledge of the fact that programs can be thought of as environments. It is the fact that there is a general container/environment metaphor in English which has been conventionally used in a number of other ways that allows it to understand these new uses. The basic strategy will be to identify the metaphor being used and then try to find conventional uses of that metaphor that are similar to the current situation. UCTeacher then analogically maps one of these known uses to the current situation. This strategy will be effective to the extent that new uses of conventional metaphors are closely related to other previously understood uses.

Previous Work on Metaphor

There have been two major approaches to the metaphor problem by the AI community. The first approach views metaphors as analogies. The problem of understanding a metaphor seen as a problem of analogically mapping information from a source domain to a target domain. Winston (1980), Carbonell (1981) and Gentner (1983) have all proposed various mechanisms for deciding how to selectively map information. The second approach has been inspired by the work of Lakoff and Johnson (1980). They assert that much of ordinary language is based on a relatively small set of systematic underlying conceptual metaphors. In this view conventional metaphors are not simple idioms nor are they the result of analogical reasoning. They reflect a set of underlying knowledge structures that are structured using conceptual metaphors. Carbonell (1980) has made a proposal that direct mappings be used to represent this type of metaphorical knowledge. He further proposed that these direct mappings could be used for analyzing metaphors. Jacobs (1985) implemented a system using similar knowledge for the purpose of generating utterances containing conventional metaphors.

Representing Knowledge about Metaphors

The conventional metaphors of English are represented as structured mappings in an abstraction hierarchy. Included in this abstraction hierarchy are metaphors which are directly related to word senses. Consider the verb kill. The literal definition can be paraphrased as 'cause the death of a living thing'. Clearly the 'process' in example (1) does not fit neatly into this definition. Now consider the following examples:

5) The Mets killed the Dodgers.

6) The senate killed the immigration bill.

7) The Islanders killed the penalty against them.

8) He killed the conversation when he came into the room.

9) A holding penalty killed the 49er's drive.

10) My editor told me to kill the last three paragraphs of my story.

The above examples contain related senses of the verb kill each with a very specific meaning. Among these senses are termination, defeat and deletion. These senses are represented by specific mappings between the KILL concept and the target meaning. Each of these mappings is in turn an instance of an abstract metaphor. In the example section it will be shown how these mappings can be used to process example (1).

The knowledge representation language that is being used to represent these mappings is KODIAK (Wilensky 1984). KODIAK is an extended semantic network language in the spirit of KL-ONE (Brachman, 1979). A unique feature of KODIAK is the semantic relation called VIEW. A VIEW is structured association between two concepts that asserts that one concept can be thought of in terms of another concept without asserting that the two concepts are related via a more abstract category. VIEWs are the tool that are being used to represent metaphorical mappings.

Metaphor and Knowledge Acquisition

The task faced by the knowledge acquisition system is to process utterances like those in examples 1 through 4, given the fact that the system's knowledge of both the facts and metaphors of the domain is incomplete. UCTeacher upon encountering an unknown metaphor uses the hierarchy of abstract metaphors and specific instances of known metaphors to come to a correct construal of the utterance. In particular the hierarchy is used to suggest plausible mappings and specific instances of metaphors are used in the creation of new mappings and concepts.

Metaphor Extension Algorithm

A four stage process is used in order to come to a correct construal of an utterance containing a new metaphor.

- Exploit the metaphor abstraction hierarchy to limit the search for a new mapping.
- Examine specific metaphoric word senses from the current example that have the mapping found in the first stage as an ancestor in the hierarchy.
- Analogically map one of these senses to the current situation.
- Create new VIEWs to connect the new meaning to the literal meaning.

The first step restricts the search to the most specific metaphorical mapping that can account for the current problem. The second step finds other ways that this metaphor has been used previously. The third step maps the meaning of one the previous uses onto the current situation. Creating the views in the final step allows this metaphor to be preocessed directly in the future. The following section describes an example of this processing in the current UCTeacher system.

An Example from UCTeacher

Consider the following working example from the UCTeacher system:

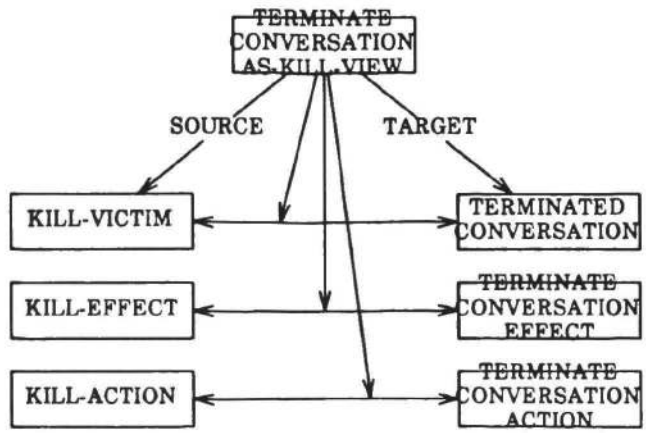
```
You can kill a process by typing '^C'.
```

The situation is such that the system has never heard the verb 'kill' applied to the concept UNIX-PROCESS. Moreover the UNIX-PROCESS concept is in direct violation of what the system knows can fill the role of the victim of a kill. In addition there are no known metaphorical VIEWs that could have been used to resolve this constraint violation. The system now attempts to create a new metaphorical mapping based on some old mapping in an attempt to resolve this constraint.

The first step performed by UCTeacher is to descend down the hierarchy to find as specific a mapping as possible that would cover this violation. Specifically it attempts to find a mapping with a source concept that covers the category living thing and a target category that allows a computer process. In this case the most specific VIEW that covers this is a VIEW from abstract entities to 'person'. Note that there are more specific VIEW's below this one but they violate the constraint on the viewed-thing being a computer-process, indicating that there are no known personifications of this concept as yet. The view that was found, PERSONIFICATION, will be used to guide the search for candidate VIEW's in the next step.

The second step is to consider pre-existing VIEWs from the kill-victim concept. These correspond to various specific metaphoric senses of the verb 'kill'. However only those VIEWs that are members of the category found in step one are considered. For each of these VIEWs an attempt is made to match the viewed-thing of this VIEW to the target concept 'UNIX-PROCESS'. The VIEW that most closely matches the target concept will be used as a plausible source VIEW.

UCTeacher uses a hierarchical matching process to try to find a closely matching candidate VIEW. This hierarchical matching process attempts to abstract up the KODIAK hierarchy from the role that the viewed thing plays in the candidate VIEW until it finds a common ancestor with the target concept. In this case until it finds a common ancestor with UNIX-PROCESS. The VIEW which is selected is the TERMINATE-CONVERSATION-AS-KILL VIEW. This VIEW is represented below.

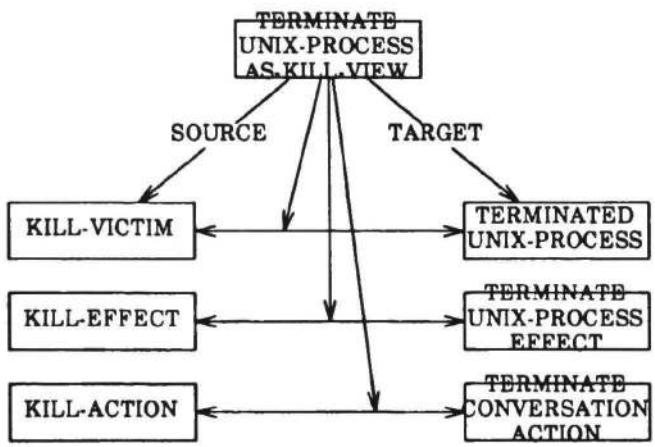


Killing a Conversation

This VIEW represents a mapping from a terminated conversation to the victim of a kill and corresponds to a word sense of kill indicating termination. The associated mappings are also shown relating the various actions and effects. These associated mappings plus the primary mapping on the victim constitute the COMPLEX-VIEW which represents this particular word sense.

The common parent concept that is found between TERMINATED-CONVERSATION and UNIX-PROCESS is TERMINATED-PROCESS. The process referred to here is the abstract notion of an ongoing sequence of actions with effects. It is important to note here that the system did not match the concept CONVERSATION against UNIX-PROCESS directly. It only abstracts on the role that CONVERSATION plays in in the context specified by the VIEW. This role is explicitly represented in the knowledge base by the concept TERMINATED-CONVERSATION. In this way the system is able to match only on those aspects of a concept that are relevant to the current context as defined by the VIEW. In this case what is relevant about the concept CONVERSATION is the fact that it can be terminated. It is this fact that needs to be mapped over to the target concept UNIX-PROCESS.

In the final phase of processing UCTeacher creates a new VIEW using the source view as a template. This new VIEW can be used directly in the future by both the analysis and generation components of UC. The following figure represents the views created for the new concepts.



Killing a Process

Conclusions

An expert using a natural language knowledge acquisition system will make utterances containing metaphors from the domain of interest. This poses a problem for the knowledge acquisition system since it may not possess the necessary metaphors for the domain. The answer to this problem is to give the system explicit knowledge about metaphors and a mechanism that can extend known metaphors to new domains. This approach is parsimonious with that suggested by Lakoff, among others. This knowledge about metaphors takes the form of a hierarchy of abstract metaphors and specific instantiations of metaphors with their meanings.

References

Brachman, R. J. et al., "Research in Natural Language Understanding". BBN report No. 4374, Cambridge, Ma. 1979

Carbonell, J.G., "Metaphor: A Key to Extensible Semantic Analysis" **Proceedings of the 18th Meeting of the Association for Computational Linguistics,** 1980

Carbonell, J. G., "Invariance Hierarchies in Metaphor Interpretation", **Proceedings of the Third Meeting of the Cognitive Science Society,** Cognitive Science Society, pp. 292-295, August 1981.

Gentner, D., "Structure Mapping: A Theoretical Framework for Analogy", **Cognitive Science** Vol. 7, No. 2, pp 155-170. 1983.

Jacobs, P. "A Knowledge-Based Approach to Language Production". PhD. Thesis. University of California, Berkeley, Report No. UCB/CSD 86/254, August 1985.

Lakoff, G. and Johnson, M., "Metaphors We Live By", University of Chicago, 1980.

Martin, J., "Knowledge Acquisition through Natural Language Dialogue" **Proceedings of the 2nd Conference on Artificial Intelligence Applications** Miami, Florida, December 1985

Wilensky, R., "KODIAK: A Knowledge Representation Language". **Proceedings of the 6th National Conference of the Cognitive Science Society,** Boulder, CO, June 1984

Wilensky, R., Arens, Y. and Chin, D., "Talking to Unix in English: An overview of UC". **Comm. ACM,** Vol. 27, No. 26 pp. 574-593, June 1984

Winston P., "Learning and Reasoning by Analogy", **Comm. ACM,** Vol. 23, No. 12, pp 689-703, December 1980

Operating Principles and Personal Theories: An Outline of Their Roles in Early Lexical Development

Carolyn B. Mervis
Department of Psychology
University of Massachusetts, Amherst

Very young children, like adults, form object categories on the basis of similarity among exemplars. But this statement is simplistic: The notion that categories are formed on the basis of similarity is not constrained adequately (Murphy & Medin, 1985). For example, consider the triplet horse, zebra, barber pole. Almost everyone would agree that the horse and the zebra were the most similar pair. In this case, similarity is being defined according to general form attributes or biological attributes. However, if the attribute "striped" were given sufficient weight, then the zebra and the barber pole would be the most similar pair. The type of similarity that provides the basis for category assignments must be specified. Murphy and Medin (1985) have argued that adults use personal theories to determine the types of similarity that will be noticed. Personal theories generally refer to informal theories or beliefs about the world that a person can state explicitly, or can answer questions about. These theories may be general, or may be restricted to specific domains. In addition to personal theories, both children and adults use operating principles or, in Slobin's (1973) words, basic self-instructions, to determine which types of similarity will be noticed. These operating principles usually are implicit and apply generally, rather than to particular domains. Infants and very young children are able to use operating principles prior to personal theories, and the use of operating principles continues throughout the lifespan.

In this paper, I describe several operating principles that I believe are used during late infancy and early childhood. I also discuss two of the earliest personal theories used by young children. I consider the effects these operating principles and personal theories have on early conceptual and lexical development and on the evolution of children's initial categories to conform to the adult standard. The proposed operating principles and personal theories are consistent with the results of a longitudinal study of early lexical development of normally developing children and children who have Down syndrome (Mervis, 1984, in press a, in press b) and the results of a diary study of my son Ari's early lexical development (Mervis, in press a).

ACQUISITION OF THE FORM-FUNCTION PRINCIPLE

By age 7 months, normally developing infants have acquired several principles that are critical for early conceptual development. First: Categories generally are concerned with whole objects (cf. Hofsten & Spelke, 1985). Second: Categories should be formed, and category assignment decisions made, based on similarity of attributes (e.g., Cohen & Younger, 1983). Third: The most important static attributes are those relevant to determining form (Olson & Strauss, 1978).

These infants treat attributes of the type generally used in categorization decisions as independent of one another (Cohen & Younger, 1983). By 10 months (the beginning of stage 5 of the sensorimotor period), infants have begun to notice correlations among attributes of the type generally used in categorization decisions. Furthermore, these infants weight correlated attributes more heavily than attributes that occur equally frequently within the category but are not correlated with one another (Cohen

& Younger, 1983), suggesting that the infants are sensitive to the importance of correlations for categorization. Infants notice both static (form) and dynamic (function, action) attributes, and begin to realize that form attributes usually have correlated function attributes, and that function attributes usually have correlated form attributes (e.g., Piaget, 1954). Thus, by age 10 months, infants have acquired a crucial principle concerning categorization, the form-function principle. This principle is: The form and function of objects generally are correlated. This attribute correlation should be used as the basis for categorization. Objects that have similar clusters of form attributes and correlated function attributes should be assigned to the same category. A corrolary to this principle is that correlated attributes are more important in determining conceptual coherence than are equally frequent but uncorrelated attributes. However, as Murphy and Medin (1985) have pointed out, the correlated attributes principle, while providing more constraint on categorization than a simple similarity principle, still is inadequate. Additional principles are necessary to determine which correlations actually are noticed.

Form-function correlations are most obvious for basic level categories (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). Not surprisingly, therefore, children's first categories are basic level categories (e.g., Mervis, 1983; Rosch, et al., 1976). Adult categories, particularly basic level ones, also tend to be formed based on such correlations. However, children's initial basic level categories often will not correspond to the adult-basic level category labeled by the same word. Such differences are to be expected; only the principles governing the determination of basic level categories are universal (e.g., Dougherty, 1978; Mervis, 1984; Rosch, et al., 1976). An object generally affords more than one set of form-function correlations. Therefore, everyone will not necessarily attend to the same correlations for a given object. The actual categories formed on the basis of the form-function principle will vary because different groups notice or emphasize different attributes of the same object as a function of different experiences or different degrees of expertise. Very young children often do not share adults' knowledge of culturally appropriate functions of objects and the correlated form attributes, leading children to de-emphasize attributes of an object that are important from an adult perspective (e.g., Mervis, 1981, 1984, in press a; see Carey, 1982, for a similar position concerning verb concepts). For example, a very young child does not have a concept of money or of saving money. Therefore, when confronted with a spherical bank, the child should ignore the slot and the keyhole. At the same time, children may notice a function and its correlated form attributes for that object that adults ignore. In such cases, children would emphasize attributes of an object that are unimportant to adults. To continue the spherical bank example, the child would notice that the object rolls and is round. Tversky and Hemenway (1984) have suggested that children often ignore small attributes that have functional significance for an adult, in favor of large attributes of the same object that afford an alternative function. In the spherical bank example, the child would ignore the slot and keyhole in favor of the round shape. (Note that different labeling practices of mothers when talking to young children rather than to adults are not responsible for these category differences; see Mervis, 1984, in press a, in press b.)

ACQUISITION OF THE ADDITIONAL CATEGORY PRINCIPLE

Children's categories that differ from the corresponding adult categories eventually must evolve to conform to the adult standard. For very young children, the first step in this process generally is the formation of a new category that includes certain members of one of the child's already-existing categories. The principle that guides this process is as follows: Assign a previously assigned object to an additional, newly formed category only if you are given concrete evidence of a new form-function attribute correlation that differs from the one on which you based your original assignment. (This new form-function correlation serves as the basis for the newly-formed category.) This principle is derived from the original form-function principle. Note that after application of this principle, the object is included in two (or more) categories; it remains a member of its old category, as well as being added to the new category.

Indications of the presence of a new correlation can take a number of forms; as part of all of these, an adult generally provides the adult-appropriate name for the object. As one possibility, the important attribute might be indicated by the child. The adult would be likely to respond by acknowledging the attribute that the child had indicated and then labeling the object with its adult-basic name. For example, if the child pointed out the wick of a round candle to the mother, she probably would respond by commenting on the wick and its function and then labeling the object, "candle."

Alternatively, an adult may point out the important attributes. Regardless of who identifies the relevant attributes, the most effective methods of introduction of the adult-basic name involve a clear indication of the new form-function correlation on which the new category should be based. Among the several methods used by adults, the most explicit indication of this correlation occurs when a concrete illustration of relevant form attributes and the correlated function attributes is provided, accompanied by a verbal description. For example, the adult might run a finger along the slot of a round bank, drop in a coin, and tell the child that this is a slot into which you put money. Use of concrete illustration without a verbal description or verbal description without concrete illustration is less effective, because the correlation is not made as explicit. The concrete illustration method is more effective than the verbal description method for very young children, because these children are more oriented to objects and action than to language. Finally, use of the adult-basic name alone is extremely unlikely to be effective. The metacognition necessary to realize that adult use of a new label implies the existence of a new form-function correlation is relatively sophisticated.

ACQUISITION OF THE EXPERT PRINCIPLE

At least two other important principles are acquired during the period of early lexical development: the expert principle and the origins principle. Acquisition of these principles occurs subsequent to acquisition of the principles already discussed. The new principles lead to changes in the perception of similarity. In this section, I consider the expert principle. The origins principle is considered in the following section.

For at least several months after the child has acquired the form-function principle, the role of purely linguistic input in categorization is relatively minor. However, once the child acquires the metacognition necessary to realize that adult use of a label for an object implies that the object is a member of the labeled category, the role of linguistic input can increase dramatically. This metacognition is based on acquisition of the

expert principle: There exist people (experts) who know more about categorization (or some types of categories) than you do. When these people apply a name to an object, they probably are correct. The expert principle requires acceptance of another person's authority on categorization issues, even when the child does not understand the basis for the other person's category assignment.

As noted previously, prior to acquisition of the expert principle, the child's categorization scheme is not influenced by adult use of adult-basic labels by themselves, without a concrete illustration or verbal description. Impressively, even after the child becomes aware that adults apply a different name to a given object (assign that object to a different category), the child is willing to let the two category assignments co-exist.

With regard to the addition of new categories, the expert principle takes two forms. The first is: Form a new category, and assign the relevant object(s) to it, if an expert labels the object with a word not included in your lexicon. The second is: Add another category assignment for a previously assigned object if an expert labels the object with a name different from the name of the category(ies) to which you had assigned the object. Both versions of this principle entail the child deciding to search for correlated attributes to serve as a category basis, just because an adult used a new label for a particular object. (It is possible for the child simply to learn the new name in reference to the object labeled, without searching for relevant attributes. In this case, however, the child will not generalize the name; it will serve essentially as a proper name for a particular object.)

Children who have acquired the expert principle, like adults, do not always accept the expert opinion (see Neisser, in press). However, in contrast to the pre-expert principle child, who is comfortable ignoring a contradictory adult label or allowing it to exist side by side with his or her own label, the child who acknowledges the expert principle cannot simply ignore contradictory input. The latter child will respond to the adult when a difference in labels occurs.

Once the child has acquired the two forms of the expert principle just described, two other forms should follow. Both involve deletion of objects from their assigned categories, ultimately yielding mutually exclusive basic level categories (Tversky and Hemenway's, 1984, mutual exclusivity of basic level categories principle). (More generally, the two forms should yield mutually exclusive coordinate categories.) *The first is: Delete an object* from its previously assigned category *if you label the object with that name* and an expert rejects your label. *The second is: Delete an object from its* previously assigned category *if an expert labels the object with a name* different from the one you would have applied, *and you have reason to believe* that the two names are coordinate.

The expert principle provides the child with a major improvement in his or her ability to acquire new categories rapidly, based on very little input. It is important to note, however, that the new categories the child forms on the basis of the expert principle are still based on the form-function principle as well. The important contribution of the expert principle is that, based on minimal input, it provides the child with an indication that a new category must be formed. The child still must search for a new form-function attribute correlation, as a basis for generalization. If the child cannot find such a correlation, then he or she will be unable to form a new category that includes more than the object to which the adult applied the new label.

ACQUISITION OF THE ORIGINS PRINCIPLE

The origins principle states that: Membership in a biological category is determined by lineage (ancestry, origins); an animal or plant belongs to the same category as its parents, and the progeny of the animal or plant also belong to the same category (e.g., Keil, in press). Acquisition of the origins principle requires acceptance of the expert principle (see Mervis, 1986).

Prior to acquisition of the origins principle, toddlers focus on observable attributes and categorize accordingly, based on the form-function principle. The available origins information (generally in the form of adjacency, as when an adult animal is depicted next to its offspring) is ignored. As Neisser (in press) has pointed out, such a focus usually will yield the same category scheme as a more sophisticated approach based on the origins principle. Sometimes, however, the two approaches yield different category assignments. For example, baby waterbirds generally are shaped more like ducks than like other types of waterbirds. In this case, the pre-origins principle child will categorize all baby waterbirds as ducks, even when he or she correctly categorizes an adjacent adult waterbird. After the child begins to acquire the origins principle, baby animals will be categorized according to parental species, whenever parental species is obvious. Categorization of juveniles will be more accurate when an adult category member is present than when the adult is absent. Prior to acquisition of the origins principle, categorization of juveniles will not vary as a function of the presence or absence of adult category members.

As is clear from Keil's (e.g., in press) research, development of the origins principle continues even during the grade school years. However, acquisition clearly begins during early childhood. This acquisition represents the earliest evidence of deviation from the form-function principle as a basis for categorization decisions. Eventually, for experts, categorization according to the origins principle will again conform to the form-function principle, but for a different set of correlations. This time, the correlations will involve genetics.

SUMMARY

In this paper, I have considered several operating principles as well as two personal theories that I believe are used during late infancy and early childhood. By age 10 months, infants have acquired the major categorization principle used across the lifespan: the form-function principle. However, infants often do not apply this principle to the same correlations as adults. Therefore, the two groups sometimes assign the same object to different categories. Category evolution, which must take place in order for the child's categories eventually to correspond to the adult standard, at first is based on the additional category assignment principle, which requires the adult to make a new form-function correlation obvious to the child. Eventually, category evolution is greatly influenced by the expert principle, which also subsumes the mutual exclusivity of basic level categories principle. The origins principle, which is based in part on the expert principle, allows the child to go beyond the form-function principle in determining the category assignments of juvenile category members. These principles clearly are important for the development of young children's categorization schemes. However, consideration of the role of operating principles and personal theories in category development has just begun. Further research and theorizing to determine additional principles and theories and their roles in category development and evolution are essential.

REFERENCES

Cardoso-Martins, C., & Mervis, C. B. (1985). Maternal speech to prelinguistic children with Down syndrome. American Journal of Mental Deficiency, 89, 451-458. (a)

Carey, S. (1982). Semantic development: The state of the art. In E. Wanner & L. R. Gleitman (Eds.), Language acquisition: The state of the art (pp. 347-389). New York: Cambridge University Press.

Cohen, L. B., & Younger, B. (1983). Perceptual categorization in the infant. In E. Scholnick (Ed.), New trends in conceptual representation: Challenges to Piaget's theory? (pp. 197-220). Hillsdale, N. J.: Erlbaum.

Dougherty, J. W. D. (1978). Relativity and salience in categorization. American Ethnologist, 5, 66-80.

Hofsten, C. von, & Spelke, E. S. (1985). Object perception and object-directed reaching in infancy. Journal of Experimental Psychology General, 114, 198-212.

Keil, F. C. (in press). Conceptual development and category structure. In U. Neisser (Ed.), Concepts and conceptual development: Ecological and intellectual factors in categorization. London: Cambridge University Press.

Mervis, C. B. (1984). Early lexical development: The contributions of mother and child. In C. Sophian (Ed.), Origins of cognitive skills (pp. 339-370). Hillsdale, N. J.: Erlbaum.

Mervis, C. B. (1986, April). Operating principles and personal theories: Their roles in early lexical development. Paper presented at the meeting of the New England Child Language Association, Cambridge, Massachusetts.

Mervis, C. B. (in press). Child-basic object categories and early lexical development. In U. Neisser (Ed.), Concepts and conceptual development: Ecological and intellectual factors in categorization. London: Cambridge University Press. (a)

Mervis, C. B. (in press). Early conceptual development of children with Down syndrome. In D. Cicchetti & M. Beeghly (Eds.), Down syndrome: The developmental perspective. New York: Cambridge University Press. (b)

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. Psychological Review, 92, 289-316.

Neisser, U. (in press). From direct perception to conceptual structure. In U. Neisser (Ed.), Concepts and conceptual development: Ecological and intellectual factors in categorization. London: Cambridge University Press.

Olson, G., & Strauss, M. (1978). Personal communication.

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

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. Cognitive Psychology, 8, 382-439.

Slobin, D. I. (1973). Cognitive prerequisites for the development of grammar. In C. A. Ferguson & D. I. Slobin (Eds.), Studies of child language development. New York: Holt, Rinehart and Winston.

Tversky, B., & Hemenway, K. (1984). Objects, parts, and categories. Journal of Experimental Psychology: General, 113, 169-193.

ACKNOWLEDGMENTS

I thank the mothers and children who participated in the longitudinal study discussed in this paper. I also thank Ari for his participation in the diary study. Chuck Clifton and Laurel Long provided valuable comments on previous versions of this paper. I am grateful to John Pani for his encouragement and critical discussion throughout this project. The research and theorizing presented in this paper were supported by the National Science Foundation, grants BNS 81-21169 and BNS 84-19Ø36. An expanded version of this paper may be obtained from Carolyn Mervis, Department of Psychology, University of Massachusetts, Amherst, Massachusetts Ø1ØØ3.

Studying the Problem Solving Behavior of Experts and Novices in Physics Via Computer-Based Problem-Analysis Environments *

Jose Mestre and William Gerace
Cognitive Processes Research Group
Department of Physics and Astronomy
University of Massachusetts
Amherst, MA 01003

ABSTRACT

The design and architecture of two user-controlled, computer-based problem analysis environments in classical mechanics are discussed. In the expert-like environment, the user analyzes problems according to a hierarchical concept schema consistent with how experts analyze novel problems in physics. In the second environment, the user searches a large equation data-base utilizing novice-like, surface feature keywords in order to locate the appropriate equation(s) to use in solving a problem. Cognitive and pedagogical implications of the research are discussed.

Novices and experts store and use domain-specific knowledge in distinctly different manners. Recall experiments reveal that experts "chunk" information into related clusters thereby facilitating recall, whereas novices display no evidence of chunking (Chase & Simon, 1973; Egan & Schwartz, 1979; Ehrlich & Soloway, 1982; Larkin, 1979). In contrast to novices, expert physicists use forward strategies, fewer equations, and shorter procedures to solve problems (Simon & Simon, 1978). Novice physics students also find it difficult to separate an abstract plan for solving a problem from the actual solution process, stating instead the equations that they would use, whereas experts outline an approach based on fundamental physics principles (Chi, Feltovich & Glaser 1981).

We are currently investigating the consequences of imposing two types of constraints on experts and novices engaged in solving problems in a branch of physics called classical mechanics. By "constraints" we mean that novices and experts will be asked to use one of two different computer-based, menu-driven environments to analyze problems before solving them. With the first analyzer, the user is asked to categorize the problem under consideration according to a hierarchical concept schema consistent with that used by experts to analyze problems. With the second analyzer, the user is asked to categorize the problem according to surface features consistent with those used by novices to solve problems.

---

* Work supported by National Science Foundation grant BNS-8511069. The contents herein do not necessarily reflect the position, policy or endorsement of NSF.

Jose Mestre & William Gerace

DESCRIPTION OF COMPUTER-BASED ANALYSIS ENVIRONMENTS

Hierarchical Analyzer

The Hierarchical Analyzer was designed to be consistent with research observations on expert physics problem solving behavior. When faced with a novel problem in which there is no obvious course of action, expert physicists first consider which basic principles apply, and then plan a solution strategy based on these principles. Consistent with this observation, Chi, Feltovich and Glaser (1981) found that, when asked to classify problems according to similarity of solution, novices use the problems' surface features whereas experts use the physical principle or law underlying the problem as the classification criterion. These findings suggest that an expert's knowledge-store can be described as a dense network containing clusters of related information with a hierarchical structure in which fundamental concepts occupy the top levels of the hierarchy, ancillary concepts occupy the middle portions, and domain-related facts and equations occupy the bottom levels of the hierarchy.

Two additional studies influenced the design of the Hierarchical Analyzer. Heller and Reif (1984) trained physics novices to generate a problem analysis called a "theoretical problem description." Force problems in classical mechanics were described in terms of concepts, principles and heuristics. When novices were induced to follow such descriptions, they improved substantially in their ability to construct problem solutions. Control novices, who received good grades in a classical mechanics course, were not able to generate appropriate descriptions of fairly routine problems. In a similar vein, Eylon and Reif (1984) investigated the effectiveness of imposing a hierarchical organization on the performance of different tasks in the domain of physics. They found that subjects who had received a particular physics argument organized in hierarchical form performed various recall and problem solving tasks better than subjects who had received the same argument non-hierarchically.

The Hierarchical Analyzer was designed to guide the user through a hierarchical analysis of a problem in terms of both concepts and heuristics. The user categorizes problems in terms of those principles and heuristics that can be used to construct a solution to the problem. This categorization is carried out by making selections from a number of menus. On any given menu, the selection made leads to another menu which is more specific than the previous menu, and which contains further menu selections consistent with the selection made on the previous menu.

The structure of the Hierarchical Analyzer resembles a flat-top pyramid. At the top of the pyramid are the four most fundamental concepts into which we have chosen to partition elementary classical mechanics. These are (1) Newton's Second Law or Kinematics, (2) Angular Momentum, (3) Linear Momentum, and (4) Work and Energy. After the initial classification, the user proceeds to menus containing ancillary concepts and useful heuristics, which occupy the middle levels of the pyramid. The menus occupying the bottom levels of the pyramid become increasingly detailed until reaching the end result of the analysis -- an equation(s) that has been dynamically constructed in accordance with the classification scheme selected.

To understand the structure and functioning of the Hierarchical Analyzer, it is perhaps best to give an example. Consider the following problem:

A small block of mass M slides along a track having both curved and horizontal sections as shown. The track is frictionless. If the particle is released from rest at height h, what is its speed when it is on the horizontal section of the track?

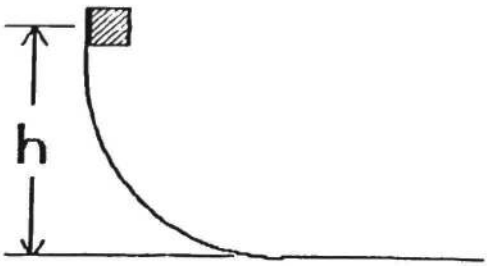


Figure 1 contains the series of menus and menu selections which appropriately analyze the problem (we have placed an asterisk next to the appropriate choice to facilitate discussions). Several features of Figure 1 should be noted. This problem can most easily be solved using work and energy principles, and thus menu item #4 is the appropriate selection. The second menu level is more specific and asks the user to describe the mechanical energy of the system. Selection #1, "Conservative system (Conservation of energy)," is the appropriate choice. Note the hints enclosed in parentheses to help the user decide which selection should be made. These hints guide the selection of choice (1) if conservative forces are present in the problem, or choice (2) if there are non-conservative forces present. At the third menu level, heuristics enter the hierarchy with the request to classify the changes in mechanical energy by considering one body at a time at some initial and final state. For the problem at hand, the block starts out with potential energy and ends up with only kinetic energy, so Selection #3 is the correct one. The fourth menu level asks the user for further classification of the changes in kinetic energy; in this case there is only a change in translational kinetic energy. At the fifth menu level, the user is asked to specify the boundary conditions (or conditions at the beginning and end points). Menu levels six and seven parallel of levels four and five for potential energy. At menu level number eight, the user is asked to specify whether there is more than one body in the system; in this case, the answer is "No." At menu level nine, the user is presented with a statement describing the principle s/he selected at the first menu level, and a statement about how this basic principle applies to the particular case at hand by elaborating on the restrictions imposed by the user's specific selections.

Note that the Analyzer does not provide the answer to a problem; it is a tool to aid the user in analyzing the problem. The user still must generate the equation that is the answer. For Problem #1, the user would have to perform some algebraic manipulations to obtain the correct answer, namely, $v=\sqrt{2gh}$. If the user has made an inappropriate selection at any menu level during the analysis, the end result would be an equation(s) that is consistent with the classification scheme selected, but inappropriate for use in solving the problem.

Three other features of the Analyzer should be noted. The "prompt line" at the bottom of menu levels 1-8 allow the user to choose options such as backing up to the previous menu in order to change a selection, entering a glossary to look up the definition of a term, or listing the menu selections made thus far. Second, if a problem lends itself to two different correct analyses, the Analyzer will allow for these two correct paths through the menu network. Finally, the user is given three choices at the final menu: 1) the problem is solved if the equation(s) given at the penultimate menu level are appropriate, 2) the user may review the equation(s) given in the penultimate menu level, or 3) if the problem being analyzed requires that two (or more)

FIGURE 1

Hierarchical Analyzer Menus & Choices

1

Which principle applies to this part of the problem solution?

1. Newton's Second Law or Kinematics
2. Angular Momentum
3. Linear Momentum
*4. Work and Energy

Please enter your selection: __

(B)ackup (M)ain menu (G)lossary (Q)uit (L)ist selections

2

Describe the system in terms of its mechanical energy

*1. Conservative system (conservation of energy)
2. Non-conservative system (work-energy exchange)

Please enter your selection: __

(B)ackup (M)ain menu (G)lossary (Q)uit (L)ist selections

3

Describe the changes in mechanical energy. Consider only the energy of one body at some initial and final state

1. Change in kinetic energy
2. Change in potential energy
*3. Change in potential and kinetic energies

Please enter your selection: __

(B)ackup (M)ain menu (G)lossary (Q)uit (L)ist selections

4

Describe the changes in kinetic energy

*1. Change in translational kinetic energy
2. Change in rotational kinetic energy
3. Change in translational and rotational kinetic energies

Please enter your selection: __

(B)ackup (M)ain menu (G)lossary (Q)uit (L)ist selections

5

Describe the boundary conditions

*1. No initial translational kinetic energy
2. No final translational kinetic energy
3. Initial and final translational kinetic energies

Please enter your selection: __

(B)ackup (M)ain menu (G)lossary (Q)uit (L)ist selections

6

Describe the changes in potential energy

*1. Changes in gravitational potential energy
2. Changes in spring potential energy
3. Changes in gravitational and spring potential energies

Please enter your selection: __

(B)ackup (M)ain menu (G)lossary (Q)uit (L)ist selections

7

Describe the boundary conditions

1. No initial gravitational potential energy
*2. No final gravitational potential energy
3. Initial and final gravitational potential energy

Please enter your selection: __

(B)ackup (M)ain menu (G)lossary (Q)uit (L)ist selections

8

Is there another body in the system which has not been examined?

1. Yes
*2. No

Please enter your selection: __

(B)ackup (M)ain menu (G)lossary (Q)uit (L)ist selections

9

The Work-Energy Theorem states that the work done on the system by all non-conservative forces is equal to the change in the mechanical energy of the system:

$$W_{nc} = E_f - E_i$$

According to your selections,

$W_{nc} = 0$ (Conservative system: mechanical energy conserved)

$E_f = (1/2\ Mv^2)_{1f}$

$E_i = (Mgy)_{1i}$

Please press any key to continue

10

*** Work and Energy ***

1. Problem solved
2. Return to Main Menu to continue solution
3. Review previous solution screens

Please enter your selection: __

different principles be applied in the analysis, the user may return to the main menu and continue. This third choice would be necessary if, for example, the block of the problem above suffered a totally inelastic collision with a second, stationary block on the level part of the track; in this case, the user would need to make a second path through the Analyzer choosing Linear Momentum at the first menu for the collision portion of the solution.

Formula-Centered Analyzer

The Formula-Centered Analyzer is intended to emulate the problem solving processes of novice physics students. It is flexible in that it could be used in a number of different ways by a number of different novices. Novices solving problems in physics tend to focus on finding the appropriate equation which can be manipulated to yield an answer. The most amusing evidence of this behavior is a typical "formula sheet" that students are allowed to take to physics exams -- it consists of a solid mosaic of equations. Further, novices appear to cue on a problem's surface features in deciding what equation to use. Surface feature cues take one of three possible forms: 1) problem types, such as "inclined plane" and "falling bodies," 2) variable names, such as "mass" and "velocity," and 3) physics terms, such as "potential energy" and "momentum."

The Formula-Centered Analyzer is a computer-based, sortable data-base made up of the equations in the first fourteen chapters of the commonly used classical mechanics text, Fundamentals of Physics by Halliday and Resnick (1974). This equation data-base contains over 150 equations, which the user can reduce by performing sequential sorts according to any one of three criteria: Variable Name, Problem Type or Physics Term. For example, for the problem above, the user may first choose to perform a sort according to the Variable Name "height," producing a list of those equations containing the variable "h." The user can then browse through the reduced equation list, or perform another sort. If the user chooses to perform another sort, for example using the Problem Type "sliding bodies," the data-base would be further reduced. After a few sorts, the number of equations would be reduced to a small, manageable number with specific properties, from which the user can select the one or two needed to solve the problem.

COGNITIVE AND PEDAGOGICAL RAMIFICATIONS

The two Analyzers described above are currently being used to study the problem solving behavior of both novices and experts. In particular, we are interested in whether or not novices exhibit distinctly different patterns of problem categorization and problem solving proficiency after prolonged use of one of the two Analyzers. With experts, we are interested in observing how they use both Analyzers; for example, do they use the Formula-Centered Analyzer as a hierarchical structure by only sorting according to Physics Term, or do they judiciously pick just the right combination of sorts to make their path to the desired equation as efficient as possible?

Although much is known about how experts and novices store domain-specific knowledge, and about how they solve problems, relatively little is known about the process of making the transition from novice to expert. We know that it

takes considerable time to become an expert, and that solving large numbers of problems is a necessary, but not sufficient condition for expertise. The work discussed in this paper is intended to address issues concerning the process of becoming an expert.

From a cognitive perspective, it is important to begin to understand how the transition from novice to expert occurs, and how it can be facilitated. Observing novices' and experts' problem solving behavior in the problem-analysis environments described above may shed some light on such questions as: 1) where along the expert-novice continuum a particular individual lies, and 2) the ability to ascertain a novice's "expert potential," or potential for becoming an expert in the domain in question. For example, the type of sort that an individual chooses to make when using the Formula-Centered Analyzer (that is, whether to sort only by variable name, or by physics term) may provide a good measure of an individual's position along the expert-novice continuum. More quantitative measures may be found in the length of time it takes a novice to adapt to the Hierarchical Analyzer, the number of specious analyses made and the ability to recognize a specious analysis.

From a pedagogical standpoint, the Hierarchical Analyzer could also provide the novice with the opportunity to actively participate in problem solving activities, while at the same time assimilating expert-like heuristics and methods for analyzing problems. In today's educational scenario, novices do not have an opportunity to observe experts engaged in problem solving activities for any prolonged period of time. When an expert physicist solves a problem for a novice, the solution is chosen for its clarity or elegance and often bears little resemblance to the process that the physicist used to solve the problem. The Hierarchical Analyzer could be a cost-effective tool for providing novices with real expert-like problem solving experiences.

REFERENCES

Chase, W.G. and Simon, H.A. (1973). Perception in chess. Cognitive Psychology, 4, 55-81.

Chi, M.T.H., Feltovich, P.J. and Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. Cognitive Science, 5, 121-152.

Egan, D.E. and Schwartz, B.J. (1979). Chunking in recall of symbolic drawings. Memory and cognition, 7, 149-158.

Ehrlich, K. and Soloway, E. (April, 1982). An empirical investigation of the tacit plan knowledge in programming. Research Report #236, Department of Computer Science, Yale University.

Eylon, B.S. and Reif, F. (1984). Effect of knowledge organization on task performance. Cognition and Instruction, 1, 5-44.

Halliday, D. and Resnick, R. (1974). Fundamentals of Physics. New York, NY: Wiley and Sons.

Heller, J.I. and Reif, F. (1984). Prescribing effective human problem-solving processes: Problem Description in Physics. Cognition and Instruction, 1, 177-216.

Larkin, J.H. (1979). Information processing models and science instruction. In J. Lochhead and J. Clement (Eds.), Cognitive Process Instruction. Philadelphia, PA: The Franklin Institute Press.

Simon, D.P. and Simon, H.A. (1978). Individual differences in solving physics problems. In R. Siegler (Ed.), Children's Thinking: What Develops? Hillsdale, NJ: Lawrence Erlbaum Associates.

COGNITIVE CONTROL OF AUTONOMOUS MOBILE ROBOTS: NESTED HIERARCHICAL INTELLIGENT MODULE

A. Meystel
Drexel University
Philadelphia, PA 19104, USA

On the Similarity Between Cognitive Controllers

An overview of the recent results in the area of Autonomous Control Systems for mobile robots suggests that different control systems show definite resemblance with each other: they have a hierarchy of knowledge-based decision-making units even when the system is equipped by a single actuator. It seems that all autonomous operations should be solved in such a similar (possibly, anthropomorphic) way. A concept of **hierarchical nested knowledge-based structure** is introduced in this paper which reflects the common properties of cognitive controllers, and an application of this concept is unfolded for a system of knowledge-based control of autonomous robots.

It is proven theoretically that nested hierarchies allow for an efficient knowledge organization as well as for correspondingly efficient knowledge processing. Theory of control oriented knowledge organization is being considered a part of a theory of **Autonomous Control Systems (ACS)** focused upon development of knowledge-based models of perception, memory, actuation, structures of algorithms, and design of systems for optimum motion of autonomous or semiautonomous systems. Theory of ACS implies that the similarities among the existing structures of autonomous robots (mostly, knowledge-based) reveal a number of inner mechanisms of goal oriented dealing with knowledge.

What Is Autonomous Control System?

Knowledge based ACS are defined as intelligent machines which should be able to operate in **completely or partially unknown environment** (or not yet recognized) with variable and/or uncertain traversability of the state space. This includes cases of obstacle strewn environment as well as

other situations based on incomplete and/or intrinsically imprecise information. Autonomous Control Systems serve as a substitute for a human operator in the multiplicity of cases where **the danger** for a human operator is expected, and also in a number of cases where the **intelligent duties** of the system require higher performance than can be provided by a human operator. It is assumed that ACS participate in goal-oriented activities, and the problems to be solved allow for its structuring in subproblems, tasks, etc. In other words, ACS is a *Golem-system.*

The structure of a typical Autonomous Control System in very general terms can be represented as shown in Figure 1. This structure contains the closed loop of a controller (sensors, perception, knowledge base, control

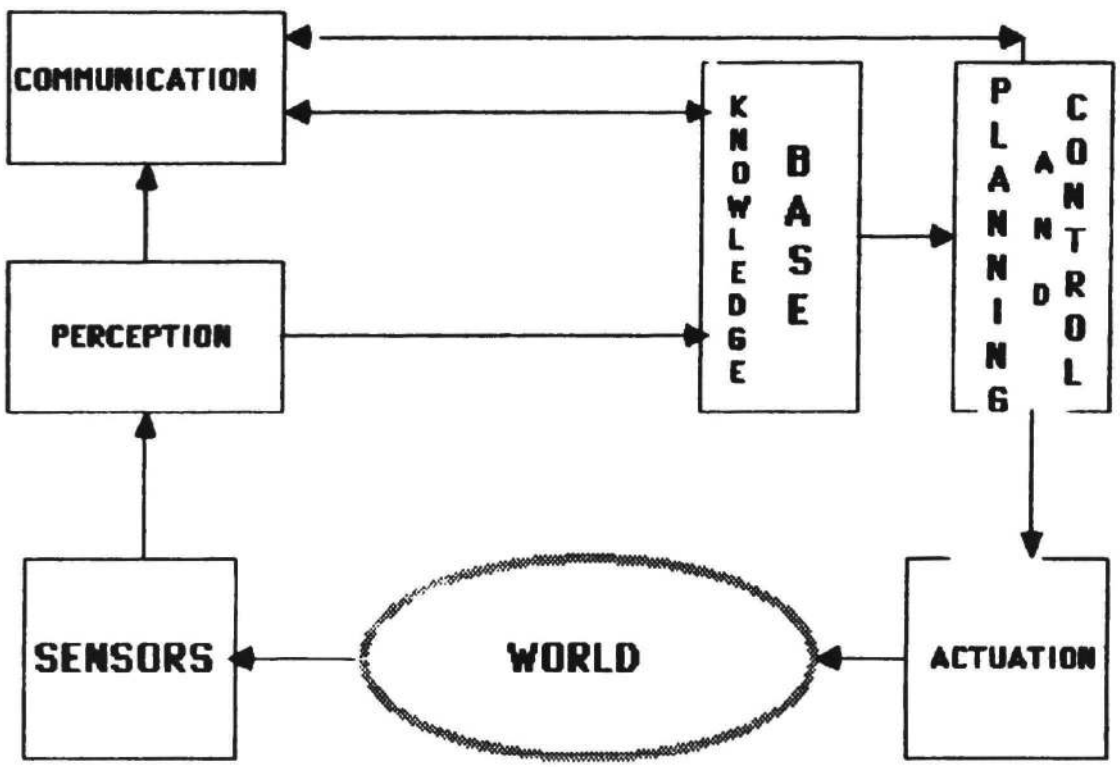


Figure 1. Structure of Autonomous Control System

and actuation loop closed through the world), and an external connection via communication link which serves to assign and reassign a task, to receive the results of the reconnaissance, to start the required ACS operation, to abort the operation, to update the ACS , and also to provide communication of several ACS units working as a team. The following features are critical: **real time operation, redundancies, space and weight constraints.**

Theory of ACS: Does It Exist?

Most of the virtual body of the theory of ACS control is implied by the results using "metaphorical" structures of cognition and incorporating the general results of the mathematical systems theory and theory of automata [M. ARBIB, 1969, 1972; R. E. KALMAN, 1969]. Some of the more recent papers help to refine the backgroung for the possible practical application [A.G. BARTO, 1981; J.J. HUPFIELD, 1985]. However, the problem of engineering design and manufacturing of ACS requires something more than a number of important general theoretical premises. The system of ACS turns out to combine a cluster of interrelated problems that must be solved only as a result of an effort with simultaneous and consistent consideration of topics treated usually in different scientific languages.

One important thing captures our attention in a variety of ACS realizations: they all are built in a hierarchical way. Structures of hierarchical intelligent control [G.SARIDIS, 1977, 1983] are potentially the proper tools for solving problems of control oriented manipulation with knowledge. Certainly, they should be given at least a rudimentary capability of performing **cognitive operations** which is usually done by various techniques of artificial intelligence, self-organizing automata, and neural networks. Part of these cognitive operations is learning which should change the quantitative evaluators in the list of rules (relations) as well as introduce new rules.

The model of dealing with cognitive operations in a form of perceptron-like fuzzy-state automaton was first introduced to simulate activities of human cerebellum (J.S. ALBUS, 1975), and then extended for application in a multiplicity of technological hierarchical structures (J.S. ALBUS, 1979, 1985). Similar methodologies are emloyed in a number of knowledge-based controllers (E.H. MAMDANI, S. ASSILIAN, M. BRAAE, D>A> RUTHERFORD, G.A. CARTER, H.R. van NAUTA LEMKE, J.J. OSTERGAARD, etc.).

Knowledge in Autonomous Systems

Knowledge Bases are usually considered as a source of well-formed formulas for man-machine decision-support systems of different kinds. Differential equations are not the only way of convenient world representation, and in the automata theory, we have a broadly developed basis for building various systems of control. Automata formalisms appear in a natural way, when the

struggle with nonlinearities, coupling, and cumbersome computations brings us to the idea of **table** (look-up).

Tables can be considered as ordered lists of clauses, (certainly, logic is presumed to be multivalued with fuzzy and/or probabilistic assigning of quantitative data). Generalization upon tables generate lists of logical (linguistical) rules. Goal oriented set of generalizations upon the lists of logical (linguistical) rules, leads to the hierarchical organization of information ("knowledge"). Goal oriented hierarchies created in this way, satisfy the following principle: *at a given level, the results of generalization (classes) serve as primitives for the above level.* Then each level of the hierarchy has its own classes and primitives; thus, it has its own vocabulary, and the algorithms assigned upon this vocabulary can be adjusted to the properties of the objects represented by the vocabulary. It is assumed that *metarules* providing the operations mentioned above, are part of the system.

It is proven that the ϵ-entropy of the knowledge organization, can be reduced using proper selection of **resolution reduction factor,** during the process of generalization. Several different knowledge hierarchies are shown to affect the operation of ACS: hierarchy of knowledge represented in the mechanical assembly, hierarchy of knowledge represented in the motion control system, in the system of sensors, and finally, in the **architecture of the knowledge processing system** *per se* (see Figure 2,a).

Each of the cones represents the resolutional hierarchies of knowledge at different levels of consideration. The domains covered by each of the "cones" communicate enabling translation among the domains. In Figure 2,b the structure of the **general knowledge** cone is shown for three levels of hierarchy. Finally, in Figure 2,c the consecutive decomposition is shown for a map of the world. Each of the consecutive images is obtained as a result of "**zooming**" procedure for the **area of attention**.

Hierarchical Nested Controller (HNC)

The ACS hierarchy is usually produced by the physical existence of a multiplicity of actuators (and/or sensors) as well as by the structure of problem, its decomposition in tasks. Then, the branching of hierarchical tree follows if the required operation (actuation) can be associated with a single

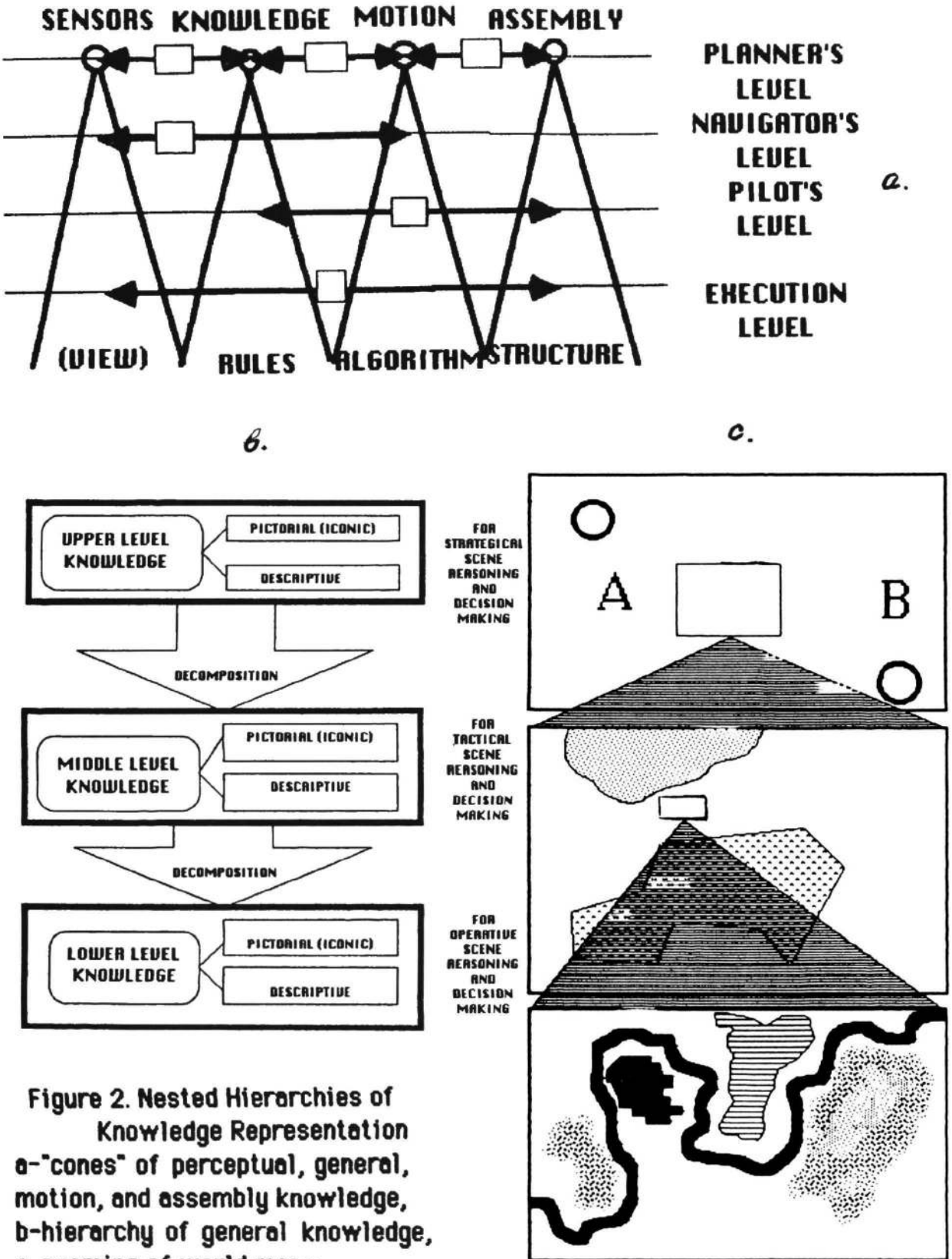


Figure 2. Nested Hierarchies of Knowledge Representation
a-"cones" of perceptual, general, motion, and assembly knowledge, b-hierarchy of general knowledge, c-zooming of world maps.

task decomposable in parts and those parts should be performed by a set of different actuators. **However, the hierarchy is retained even when only one actuator is being considered**. This would be a hierarchy of nested decision making processes over a nested hierarchy of world representation where each of the levels can be characterized exhaustively by the value of **resolution of knowledge representation**.

Hierarchical decision making process allows for using the limited computer power at each level of such hierarchy with no branching efficiently by consecutive zooming procedures.This hierarchical system of representation employs Minsky's "frames" or Samet's "quad-trees" in a "natural" way.. In this case, the tree hierarchy of intelligent control converts into a **hierarchical nested controller (HNC)**.

If HNC is acting under the above mentioned constraints the process of control allows for decoupling in parts dealing separately with information of different degree of resolution (easily interpreted as certainty, belief, etc. This means that the degree of "fuzziness" is different at different levels of the hierarchy, and in the nested hierarchy of the fuzzy-state automata **each automaton of the lower level (automaton-child) is enclosed in the corresponding parent-automaton**, and serves for clarification of its uncertainties.

A nested hierarchy of knowledge which is organized according to the degree of certainty and belief, implies a nested hierarchy of decision-making processes which in fact, leads to a similar nested hierarchical structure of **PLANNER-NAVIGATOR-PILOT** currently employed in some versions of mobile autonomous robots. ACS functioning depends upon a subset of cognitive operations in HNC associated with **motion planning, navigation, and control for autonomous mobile robots**. Thus, Figure 1 can be modified, and the refined structure of the system which reflects the HNC operation, is shown in Figure 3.

Nested Hierarchy of Knowledge in ACS

Clearly, the functional subsystems of ACS: "Perception", "Knowledge Base", and"Planning-Control" are intrinsically interlaced, and Figure 3 shows that they can be considered as an entity ("intelligent module"). This entity is built upon two interrelated knowledge bases: one, carrying the entity-relationships (ER) structure of the world, and another, defining

operations upon this structure. The operations are determined by the actual character of the primitives at a given level of consideration. These two interwoven knowledge bases constitute the background for the ACS operation.

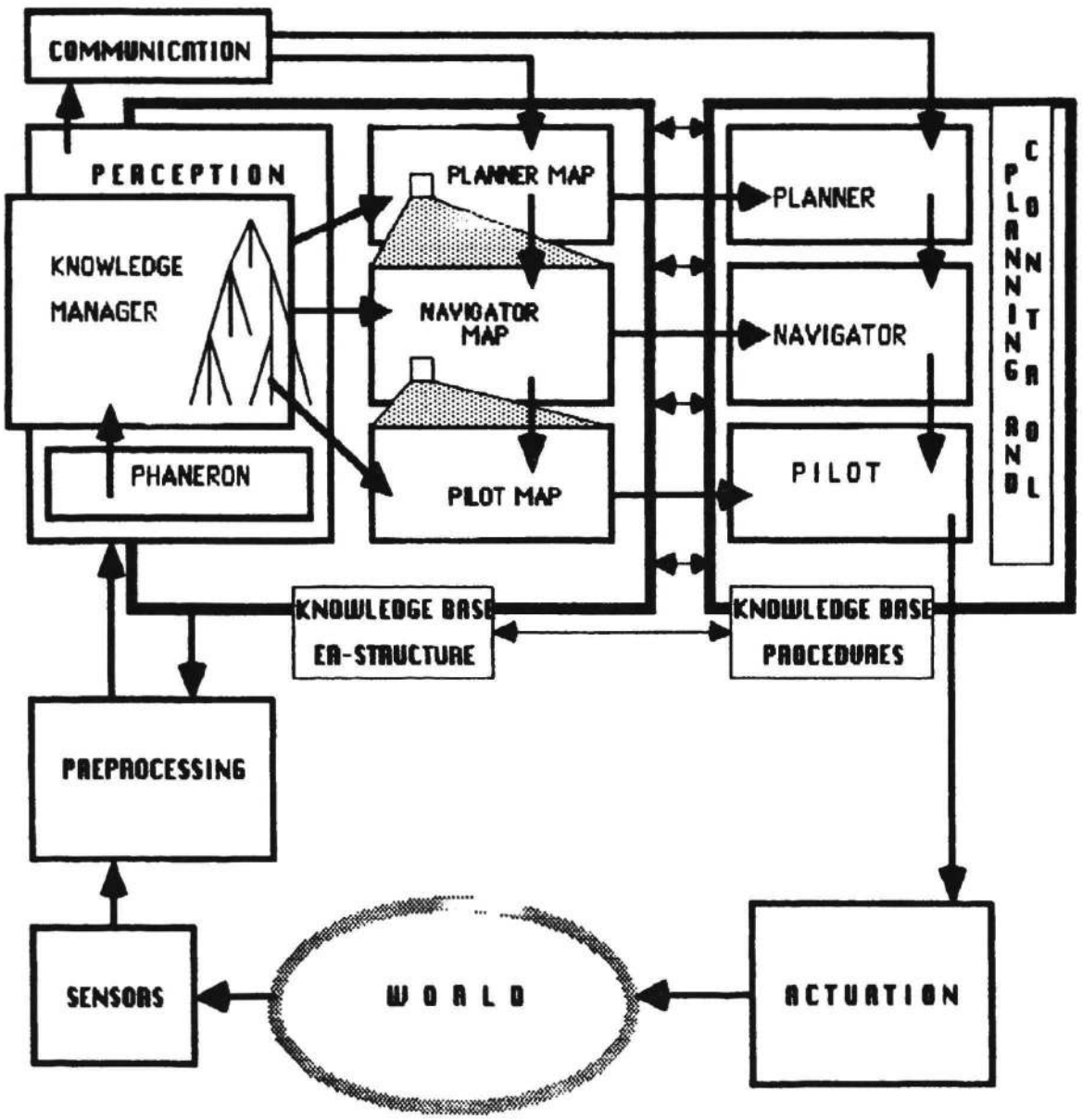


Figure 3. Refined ACS structure: HNC operation is illustrated

Perception Stratified by Resolution

Knowledge acquisition is understood as a **two-step process** consisting of **information acquisition** (which is done by sensors), and **information organization** (which in fact, transforms the raw information into knowledge). First, sensors deliver **phaneron** to the knowledge manager. (The term "phaneron" was introduced by C.S. Pierce for the totality of information which can be called **phenomenal world**). Phaneron is not structured at the moment of arrival, it should be recognized, identified within some ER-structure which might not yet been created. These processes are broadly discussed in literature, and the importance of such phenomena as "**attention**", and "**resolution**" in the process of knowledge acquisition was emphasized many times in literature (R. Bajcsy, M. Levine, A. Hanson, E. Riseman, etc.).

The result of this first step of knowledge acquisition (a **snapshot of the world**), contains information part of which can be different in the next snapshot, and part won't change (e.g. about relations among objects and/or their properties). Thus, the identification can be done only in the context, i.e. in constant interaction with knowledge base. This affects the set of preprocessing procedures which are being separated from the rest of the intelligent module primarily because of successful experience of modular manufacturing of the computer vision systems systems. Simultaneously with the process of finding phaneron structure (or image interpretation) the problem of proper allocation of the information contained within phaneron should be done. Thus, the system of nested hierarchy of Planner-Navigator-Pilot maps is being created.

Knowledge Representation Stratified by Resolution

Separation in levels appeared to be a natural phenomenon linked with the properties of **attention**, and its intrinsical links with the process of **generalization**. In fact, generalization is required to provide the efficiency of computing resources use and allocation, and attention is one of its tools. Thus, the new class labels which are created by the process of generalization, are being considered as new primitives of the upper level of world representation. This rule: **the class labels of the lower level are considered as primitives for the higher level**, is one of the laws of the mechanism of nested hierarchy.

Knowledge represented by ACS contains at least two parts: thesaurus and context. Thesaurus is maintained independently of particular operation to be performed, and it constitutes the ASS "wisdom", "education", and "experience". Context is determined by the task within a domain of thesaurus, and can be considered as a "map" of the world in which the operation must be designated together with the list of rules pertaining to this map. Map of the world is extracted from the series of snapshots.

Knowledge in the form of "planner's map" should be maintained for a long time due to the "slow rhythm" of this level. Changes in the upper level map are not frequent. "Navigator's map" is to be regularly updated but it can be maintained as a part of "planner's map". Pilot may or may not need a map maintained as a part of "Navigator's map". Actually, from our first experience of dealing with ACS we found that intelligent module cannot afford maintenance of the pilot map (the lowest level of world representation), and therefore all processes related to the real time operation have **ephemeral structure** with a number of logical filters determining whether this ephemeral information contains anything to be included in the long term memory.

Maps that have emerged on the surface of the ER-structure as a working representation of phaneron, imply corresponding procedures of motion planning and control.

Planning/Control Stratified by Resolution

Planning is traditionally considered to be a process which is performed separately from the process of control. This is acceptable for the vast multiplicity of systems where planning can be performed off-line, and the process of control can be initiated given a set of highly generalized **units of knowledge** together with a number of **unchangeable goals**. By lowering the level of generalization and keeping the certainty and belief in the required limits of the level **resolution**, we can build in a hierarchy of nested planning processes. In this hierarchy, the desirable trajectory determined at the higher level arrives to the lower level as a **fuzzy stripe** (FS). The new planning is being done within FS at a higher resolution.

This decoupling of the decision-making upper levels (or *off-line stages*) from the lower levels of decision-making and immediate performance (or

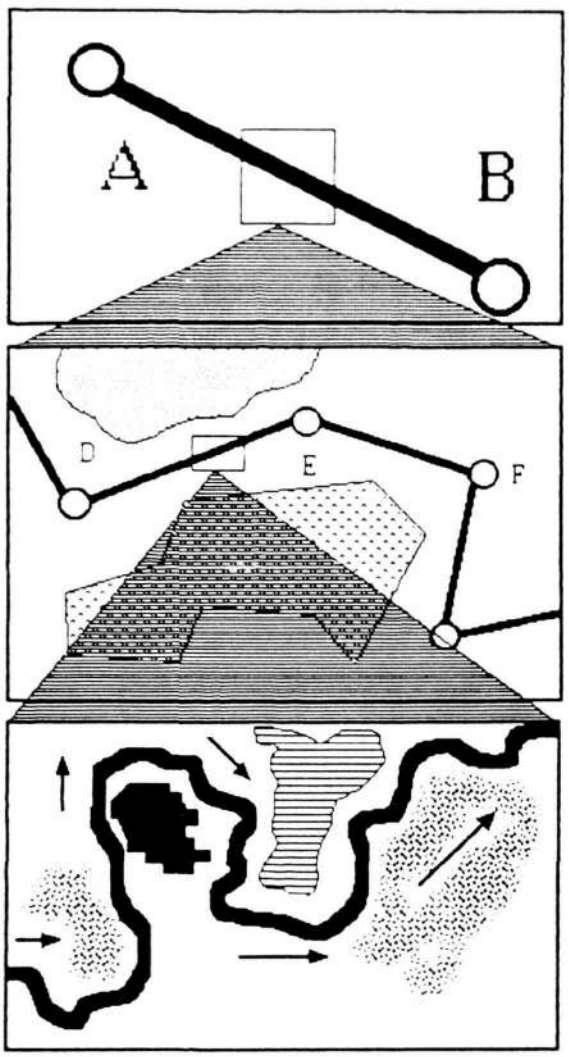


Figure 4. Planning as a nested hierarchical process

on-line stages) is probably the most characteristical property for telling the planning stages from the control stages of operation as well as distinguishing the corresponding subsystems or any device where **constant human involvement** is presumed. This decoupling does not take place in ACS: planning and control are the inseparable parts of the unified HNC. The levels of planning and control are connected together by an intermediate level of decision making dealing with processes which have to use knowledge at a definite level of generalization and yet after processes of updating are completed.

This means that at this intermediate level, the results of the ongoing motion affect the results of generalization (sinse the system of "Perception" initiates processes of information updating). We name planning processes **navigation** *per se* at the level of "Planning-control" subsystem where the results of real-time updating are becoming crucial for the results of planning. In Figure 4, three consecutive nested operations of planning are performed at three resolutional levels of the system. One can see that at the level of the least resolution, the plan is visualized as a straighr line AB. After zooming a segment of this line into higher resolution, new information is obtained, the straight line is being substituted by another plan: DEF. The next zooming discloses even more details about the environment. The actual motian among small obstacles is shown in the third map. So, these three plans show different paths, the direction of motion seems to be different. And yet, all of them are correct plans after being related to the resolution of a particular level.

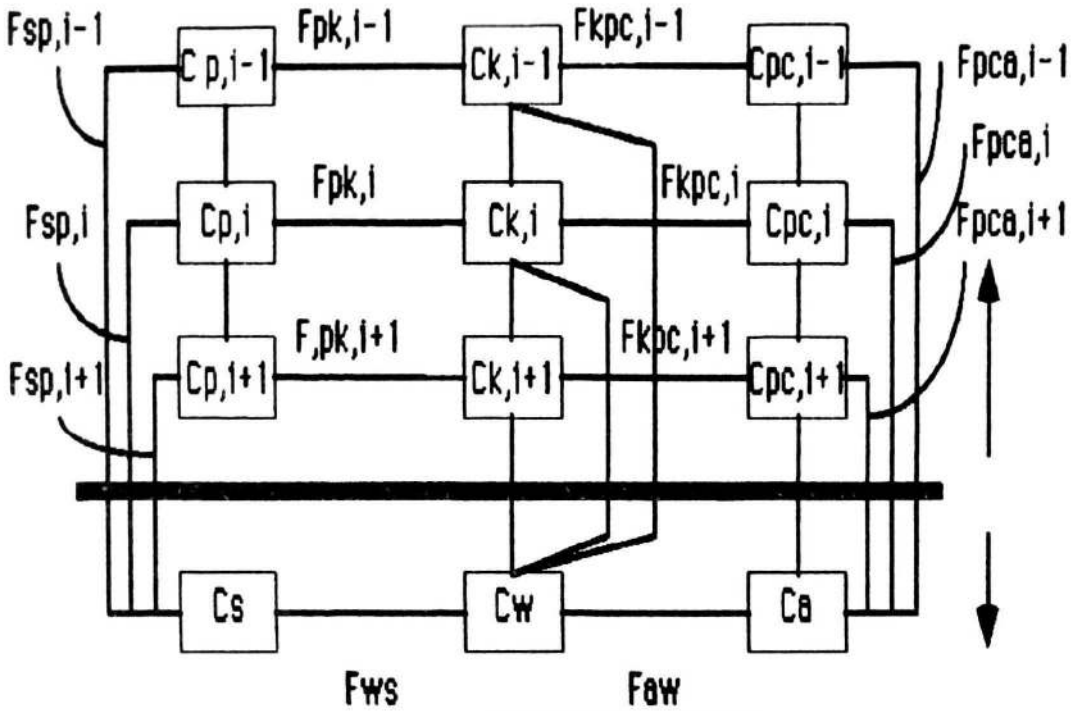


Figure 5 Category-theoretical Representation of ACS

Nested hierarchy of perception does not require having any hierarchy of sensors although does not preclude any acceptable hardware solution. Nested hierarchy at the stage of preprocessing is being viewed as a result of sequential **zooming** operation, or in other words operation of the **focusing of attention.** In ACS zooming must be based upon focusing of attention otherwise, the constraint of the limited computing power would not be satisfied. (One can see that this concept can be interpreted within the framework of existing theories of image organization and interpretation, see A. HANSON, E. RISEMAN, 1978).

Nested Hierarchical Production System

All of the planning-control levels of the mechanism of **knowledge-based navigation** interact vertically via recursion of the **algorithms of sequential production** providing sequential refinement top-down, and correctional replanning bottom-up. Functioning of the hierarchical production systems of perception, and planning-control, is supported by vertical interaction of levels in the "Knowledge Base" via aggregation and decomposition based upon preassigned values of resolution per level. So, the thesaurus as well as context, exist as a result of internal processes of self-organization within the body of knowledge.

On the contrary, the two couple of subsystems: "Perception-Knowledge Base" and "Knowledge Base-Planning/Control" (shown in Figures 1 and 3) are being viewed in our theory as **vertical nested knowledge processing hierarchies with horizontal interaction per level**. Indeed, all new knowledge acquired should be organized, the list of primitives in operation must be verified and updated. This procedure is being done at a horizontal level as well as excercising the algorithms of control. In the latter case, the map of the world as well as the list of rules to deal with this map are becoming an object of heuristic discretization and search.

Category-theoretical description of ACS

Considering subsystems as **categories C**, and the interaction among them as **functors F**, the commutative diagram can be shown as follows (indices mean: s-sensing, p-perception, k-knowledge, pc-planning/control, a-actuation, w-world) as shown in Figure 3. (Feedbacks are not shown: boxes are connected by "functors" which characterize the structure conservation in a set of mappings of interest). The bold horizontal line separates two major different parts of the system: what is below, is a world of real objects, and what is above the bold line, is the world of information processing.

All of the "boxes" in Figure 3 are fuzzy-state automata. They are easily and adequately described in terms of the automata theory, provide consistency of the descriptions, computer representations, control operation, and they are taylored for dealing with knowledge processing. Then, the search can be done by combining A* and dynamic programming, discretization of the space is being determined by the level of resolution, and the rules which are formulated within the given context.

Knowledge-based Optimum Control of ACS

In this area, the methodology of knowledge engineering can give substantial benefits. The problem of motion planning was given attention in the literature on AI and robotics. However, in a pure analytical domain problems of **optimum planning** as well as **optimum control** until now do not have applicable solutions. Motion planning is frequently understood in the context of **"solvability" of the problems of positioning or moving** the object rather than in the context of finding the desirable trajectory of motion. Nevertheless one cannot argue that the real problem of concern is finding the

location and/or trajectory of motion which provides a desired value of some "goodness" measure (e.g. the value of some cost-function). These features of the problem, constitute a good "bridge" for interference of knowledge engineering methodology.

The emphasis of the well known concept of the "configuration space approach", is done upon techniques of constructing the **admissible** swept volume but no **optimality** is considered, and certainly, no dynamics of the motion is discussed. Most of the algorithms based upon the theories mentioned above, are oriented toward **off-line operation**, they require considerable time and constant human involvement. Finally, all of the existing works presume complete knowledge of the environment, and operate in a structured world. No result is known contemplating planning of motion in unstructured situation. In the meantime, this situation is a typical one for a hierarchical system of **knowledge-based autonomous control**.

Considering the problem of motion planning as a pure geometric issue, can be understood given complexity of this problem, and the mathematical elegance of solutions it generates. We would like to express here our appreciation of the results containing the advancements in using configuration space (T. LOZANO-PEREZ, M.A. WESLEY, W. RED, H.V. TRUONG-CAO. A.A. PETROV, T.M. SIROTA), in finding the minimum distance path under geometrical constraints (J.Y.S. LUH, C.S. LIN, L.A. LOEFF, A.H. SONI, S.M. UDUPA, C.E. CAMPBELL), upon the network (G. GIRALT, R. SOBEK, R. CHATILA, V.A. MALYSHEV), using the Voronoi diagrams for motion planning with and with no retraction (R.A. BROOKS, C.K. YAP) as well as introduction and the treatment of such problems as "moving the ladder", "moving the piano", and so on (C.K. YAP). Various methods of minimum path construction have been applied based upon determining the "potential field" surrounding the obstacles (O. KHATIB), global flow analysis using Gauss-Jordan elimination (R.E. TARJAN), applicable when the full knowledge of the world is presumed to be given. An interesting example of using neural networks for optimization of motion (J.J. HOPFIELD, D.W. TANK, 1985), is promising within the aspect of this paper.

The following comment should be taken in account: the above mentioned works reflect a paradigm of **off-line static planning of motion trajectory in a cluttered limited well known space**. Clearly, this is only a part of the whole problem- an important one but just a part. As soon as the **on-line real-time** planning is required, as soon as **dynamics** is involved, as soon as the "plant" is **complex and hierarchical** one, and the world is **not uniform** and **not well known**, finally, as soon as the

computer power turns out to be limited (as happens in all autonomous systems) - then the old premises are not working anymore.

Experience of Simulation and Testing

Based upon this approach, a nested hierarchical intelligent module has been developed for knowledge-based control of an autonomous mobile robot. The module was simulated, and the processes ofknowledge acquisition, organization and knowledge-based planning-control have been analyzed in a variety of situations including a number of terrains including flat and 2 1/2 D ones (D. GAW, A. MEYSTEL, 1986), obstacle strewn environment (A. MEYSTEL, A. GUEZ, G. HILLEL, 1986), and different cost-functions for optimum control. Simulation of nested hierarchical planning is illustrated in Figure 6. One can see that after development of the upper level plan ("**go straight from initial position to goal**"), next level (Navigator) changes the plan upon updating map by obstacles information. Middle-level plan is developed in presumption that there is an exit on the right. The lowest level, Pilot visualizes the obstacle, and the motion is again replanned and corrected.

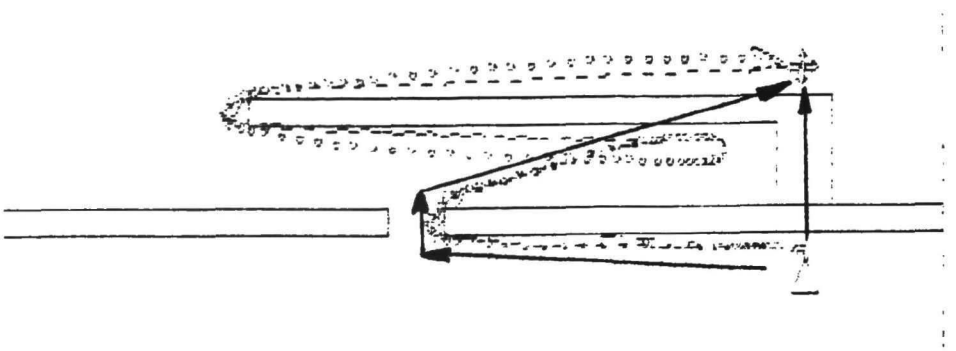


Figure 6. Example of NHC simulation

The processes of dynamic navigation have been analyzed (A. GUEZ, A. MEYSTEL, 1985). The results have confirmed that nested structure is applicable for goal oriented motion refinement of minimum time dynamic system. The software package corrected after computer simulation, is being verified by testing an indoor mobile autonomous robot (with ultrasonic "vision"). Indoor testing has confirmed the NHC analysis. New advanced algorithms of Piloting has been developed.

The outdoor system is being developed for operating upon terrain 5x5 sq. mi. Navigator's focus of attention is 1500x1500 sq.ft. Focus of attention at the Pilot level is 200x200 sq. ft. The principle of Nested Hierarchical Control is

represented consistently in all of the subsystems. In order to provide the outdoor test of the the system, a vehicle is being
outdoor test of the the system, a vehicle is being manufactured with three levels of vision (with a laser-scanner, with a CCD camera, and with ultrasonic sensors). NHC enable development of a new principle of Vision at the Pilot Level using segmentation with no edge detection.

Computer Architectures

Theory of hierarchical nested control not only generates the conceptual knowledge acquisition and processing oriented architectures of cognitive modules for autonomous mobile robots, but also suggests number of preferable computer architectures as well as techniques of dealing with the problem of assemblying the system from existing architectures. These architectures are implicitly described in a part of Figure 3 above the bold line.

References.

J.S. Albus, 1975, "A New Approach to Manipulator Control: The Cerebellar Model Articulation Controller (CMAC)", Transactions of the ASME, Journal of Dynamic Systems, Measurement, and Control, September, p.p. 220-227.

J.S. Albus, 1979, "Mechanisms of Planning and Problem Solving in the Brain", Mathematical Biosciences, v. 45, p.p. 247-293

J.S. Albus, C.R. McLean, A.J. Barbera, M.L. Fitzgerald, 1985, "Hierarchical Control for Robots, and Teleoperators", Proc. of the Workshop on Intelligent Control, Troy, NY.

M.A. Arbib, 1969, Theories of Abstract Automata Englewood Cliffs, NJ, Prentice Hall.

M. A. Arbib, 1972, The Metaphorical Brain, Wiley, New York.

R. Bajcsy, L. Lieberman, 1974, "Computer Description of Real Outdoor Scenes", Proceedings of the 2-nd International Joint Conference on Pattern Recognition, Copenhagen.

R. Bajcsy, D. Rosenthal, 1980, "Visual and Conceptual Focus of Attention", In Structured Computer Vision, eds. by S. Tanimoto and A. Klinger, Academic Press, New York.

M. Braae, D.A. Rutherford, 1979, "Theoretical and linguistic aspects of the fuzzy logic controller", Automatica, Vol. 15, No. 5, September.

R.J. Brachman, 1979, "On the Epistemological Status of Semantic Networks", In Associative Networks: Representation and Use of Knowledge by Computer, ed. by N.V. Findler, Academic Press, New York.

R.A. Brooks, 1982, "Solving the Find-path Problem by Good Representation of Free Space", Proc. of the 2nd AAAI Conference, Pittsburgh.

R. Chavez, A. Meystel, 1984, "Structure of Intelligence For An Autonomous Vehicle", Proc. of the IEEE Int'l Conf. on Robotics and Automation, Atlanta, GA.

D. Gaw, A. Meystel, 1986, "Minimum-Time Navigation of an Unmanned Mobile Robot in a 2 1/2 World with Obstacles", Proc. of IEEE Conf. on Robotic and Automation, San Francisco, CA

G. Giralt, R. Chatila, M. Vaisset, 1983, "An Integrated Navigation and Motion Control System For Autonomous Multisensory Mobile Robots", 1st Int'l Symp. of Robotic Research.

G. Giralt, R. Sobek, R. Chatila, 1979, "A Multilevel Planning and Navigation System For a Mobile Robot: A First Approach to Hilare", Proc. of IJCAI-79, Vol. 1, Tokyo.

A. Guez, A. Meystel, "Time -Optimal Path Planning and Hierarchical Control via Heuristically Enhanced Dydemic Programming: a Preliminary Analysis", 1985, Proc. of the Workshop on Intelligent Control, Troy, NY.

A. Hanson, E. Riseman, 1978, "VISIONS: a Complete System for Interpreting Scenes", In Computer Vision Systems, ed. by A. Hanson, E. Riseman, Academic Press, New York.

P.J. Hayes, 1979, "The Logic of Frames", in Frame Conceptions and Text Understanding ed. by D. Metzing, Walter de Gruiter & Co, Berlin.

J.J. Hopfield, D.W. Tank, 1985, " 'Neural' Computation of Decisions in Optimization Problems", Biological Cybernetics, 52, p.p. 141-152.

Y.C. Ho, 1985, "System Theory, and Operation Research - A New Fusion of Mathematical Modeling and Experimentation", Proc. of the Workshop on Intelligent Control, Troy, NY.

C. Isik, A. Meystel, 1984, "Knowledge-based Pilot for an Intelligent Mobile Autonomous System", Proc. of the First Conference on Artificial Intelligence Applications", Denver, CO.

C. Isik, A. Meystel, 1986, "Structure of a Fuzzy Production System for Autonomous Robot Control", Proc. of SPIE, vol. 135, Applications of Artificial Intelligence III, ed. by J. Gilmore, Orlando, Fl.

M. Julliere, L. Marce, H. Place, 1983, "A Guidance System For a Mobile Robot", Proc. of the 13th Int'l Symposium of Industrail Robots and Robots 7, Vol. 2, "Future Directions", Chicago, IL.

R.E. Kalman, P.L. Falb, M.A. Arbib, 1969, Topics on Mathematical System Theory, McGrow-Hill, New York

O. Khatib, 1985, "Real-Time Obstacle Avoidance For Manipulators and Mobile Robots", (Preprint)

E. Koch, C. Yeh, G. Hillel, A. Meystel, C. Isik, 1985, "Simulation of Path Planning For a System With Vision and Map Updating", Proc. of IEEE Int'l Conf. on Robotics and Automation, St. Louis, MO.

M. Levine, 1980, "Region Analysis Using a Pyramid Data Structure", in Structured Computer Vision, eds. S. Tanimoto and A. Klinger, Academic Press, New York.

L.A. Loeff, A.H. Soni, 1971, "An Algorithm for Computer Guidance of a Manipulator between Obstacles", ASME Transactions, paper 74-DET-89.

T. Lozano-Perez, 1983, "Spatial Planning: A Configuration Space Approach", IEEE Transactions on Computers, Vol. C32, No. 2.

T. Lozano-Perez, M.A. Wesley, 1979, "An Algorithm For Planning Collision-Free Paths Among Polyhedral Obstacles", Communications of ACM, Vol. 22, No. 10.

J.Y.S. Luh, 1981, "A Scheme For Collission Avoidance with Minimum Distance Traveling For Industrial Robots", Journal of Robotic Systems, 1(1).

J.Y.S. Luh, C.E. Campbell, 1982, "Collision-Free Path Planning for Industrial Robots", Proc. of the 21st IEEE Conference on Decision and Control, vol. 1, Orlando, FL.

J.Y.S. Luh, C.S. Lin, 1981, "Optimum Path Planning For Mechanical Manipulators", Trans. of the ASME, J. of Dynamic Systems, Measurement and Control, Vol. 102, June.

V.A. Malyshev, 1981, "Representation of An External Medium and Planning Construction of Program Motions of A Manipulator", Engineering Cybernetics, No. 3.

E.H. Mamdani, S. Assilian, 1975, "An experiment in linguistic synthesis with fuzzy logic controller", Int. J. Man-Machine Studies, Vol. 7, p.p. 1-13

L. Marce, M. Julliere, H. Place, 1984, "An Autonomous Computer-Controlled Vehicle", Proc. of the 1st Int'l Conf. on Automated Guided Vehicles.

A. Meystel, 1983, "Intelligent Control of a Multiactuator System", In IFAC Information Control Problems in Manufacturing Technology 1982, ed. by D.E. Hardt, Pergamon Press, Oxford.

A. Meystel, A. Guez, G. Hillel, 1986, "Minimum Time Path Planning for a Robot", Proc. of the IEEE Conf. on Robotics and Automation, San Francisco, CA.

A. Meystel, 1986, Primer on Autonomous Mobility, Drexel University, Philadelphia, PA.

J. Mylopoulos, H.J. Levesque, 1984, "An Overview of Knowledge Representation", In On Conceptual Modelling ed. by M.L. Brodie, et al, Springer-Verlag, New York.

H.R. van Nauta Lemke, W.J.M. Kickert, 1976, "The application of fuzzy set theory to control a warm water process", Vol. 17, p.p. 8-18.

J.J. Ostergaard, 1976, "Fuzzy logic control of a heat exchanger process", Tech. report No. 7601, Tech. Univ. Denmark, El. Power Eng. Dept.

Y.-H. Pao, 1985, "Some Views on Analytic and Artificial Intelligence Approaches", Proc. of the Workshop on Intelligent Control, Troy, NY.

A.A. Petrov, T.M. Sirota, 1983, "Obstacle Avodance by a Robot Manipulator Under Limited Information About Environment", Automation and Remote Control, No. 4.

W. Red, H.V. Truong-Cao, 1984, "The Configuration Space Approach to Robot Path Planning", Proc. of the 1984 ACC, Vol. 1, San-Diego, CA.

D. Rutherford, G.A. Carter, 1976, "A heuristic adaptive controller for a sinter plant", Proc. 2-nd IFAC Symp. Automation in Mining, Minerel, and Metal Processing, Johennesburg.

G.N. Saridis, 1977, Self-Organizing Control of Stochastic Systems, Marcel-Dekker, New York.

G.N. Saridis, 1983, "Intelligent Robotic Control", IEEE Transections on Automatic Control, Vol. AC-28, No. 5.

G.N. Saridis, J.H. Graham, 1984, "Linguistic Decision Schemata for Intelligent Robots", Automatica, Vol. 20, NO 1.

G.N. Saridis, 1985, "Foundations of the Theory of Intelligent Control", Proc. of the Workshop on Intelligent Control, Troy, NY.

R.E. Tarjan, 1981, "Fast Algorithm for Solving Path Problems", Journal of the ACM, Vol. 28, No. 3.

R.E. Tarjan, 1981, "A Unified Approach to Path Problems", Journal of the ACM, Vol. 28, No. 3.

S. Udupa, 1977, "Collision Detection and Avoidance in Computer Controlled Manipulators", Proc. 5-th IJCAI, Cambridge, MA.

C.K. Yap, 1985, "Algorithmic Motion Planning", in Advances in Robotics v. 1, ed. by J.T. Schvartz, C.K. Yap, Lawrence Erlbaum Publ.

L.A. Zadeh, 1965, "Fuzzy Sets", Information and Control, Vol. 8.

L.A. Zadeh, 1981, "PRUF - a Meaning Representation Language for Natural Languages", in Fuzzy Reasoning and its Applications, ed. by E.H. Mamdani, B.R. Gaines, Academic Press.

L.A. Zadeh, 1984, "Fuzzy Probabilities", Information Processing and Management, vol. 20, NO. 3.

Self-supervised Learning: A Scheme for Discovery of "Natural" Categories by Single Units

Paul Munro
Institute for Cognitive Science C-015
University of California San Diego
La Jolla, CA 92093

ABSTRACT

Several dynamical systems have been previously proposed to give a neural-like (i.e. connectionist) description of category formation. These typically either involve supervised training (as in Sutton & Barto, 1981; Reilly et al., 1982) or identify dense regions ("clusters") in the stimulus distribution as natural categories (Amari & Takeuchi, 1978; Rumelhart & Zipser, 1985). By combining two existing connectionist-type learning procedures, one supervised and one unsupervised, a hybrid "self-supervised learning" (SSL) mechanism for concept and category learning has been developed. Each unit in the network comes to represent some concept of the order of complexity of a single word; the activity of the unit signals the contribution of its associated concept to the current mental state. A crucial assumption of this approach is that every concept unit (C-unit) receives inputs from two or more information streams. The self-supervised learning process is governed by a data-driven dynamical rule which results in a two-stage learning process. In the first stage, a C-unit becomes selectively responsive to a particular pattern s^{trig} from one of the information streams, ignoring all other patterns in that stream. This is followed by an associative stage in which the unit develops graded response properties to stimulus patterns incident from the other information stream(s). The trigger feature thus becomes a kind of prototype for the concept to be formed by the C-unit. Populations of C-units display interesting representational properties; these are seen to have attributes of both local and distributed representations.

THEORETICAL ANALYSIS

Model architecture

The elements described in this model are labelled C-units (**concept** or **category** units). Each C-unit receives input from two or more (n) groups of afferents (Figure 1), or **input banks**; in principle, n need not be the same for every unit.

In general, indices will follow the convention that superscripts denote the bank and subscripts the component within the bank: afferent j of bank i delivers activity s_j^i via a synapse of strength w_j^i such that a partial response r^i is computed over each bank by the unit in a two-step process consisting of a linear summation followed by a nonlinear "squashing" or "compressing" function:

$$x^i(\mathbf{w}^i, \mathbf{s}^i) = \sum_j w_j^i s_j^i \tag{1}$$

$$r^i = \sigma(x^i) \tag{2}$$

where σ is subject to the condition

$$\sigma(0) = 0 \tag{3}$$

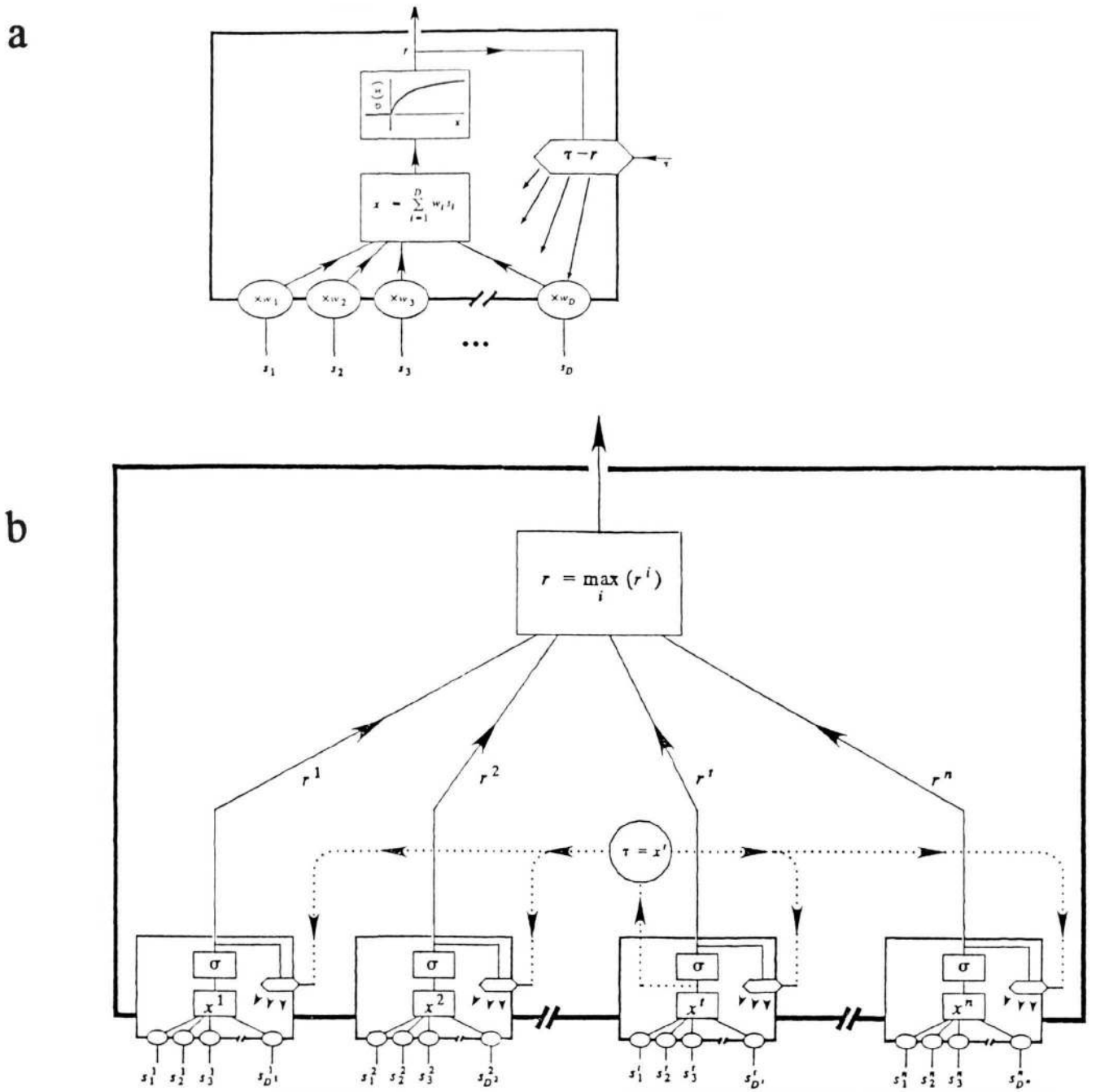


Figure 1. *Information flow in a C-unit.* **a.** An information flow diagram for a single bank, which autonomously follows an algorithm for supervised learning. The stimulus components s_i are weighted by corresponding unit parameters w_i to give a linear activation value x, which is passed to the squashing function σ yielding the unit response r. The value of r is compared with the training signal τ to generate the error value $(\tau - r)$ which is used to adjust the weights according to the rule $\Delta w_i = \alpha(\tau - r)s_i$, where the learning rate α is a small number. **b.** The complete C-unit consists of several input banks (four of these are shown), which each act as a supervised semi-linear unit. Each input bank receives a common training signal τ, but applies the signal to stimulus patterns from different environments. The linear summation stage of one of the banks generates this training signal $(\tau = x^i)$ such that this bank effectively follows the rule for selectivity maximization described in the introduction. The output r of the C-unit is given by a function of the bank responses r^i; in this paper, $r = \max_i(r^i)$.

$$\sigma(x) > x\sigma'(x) \quad \text{for all } x > 0$$

The response r of the unit is a function of the partial sums $r^1 \cdots r^N$. The precise form of this function need not be specified at this point, but it should be nondecreasing in all the r^i; i.e. $\frac{\partial r}{\partial r^i} \geq 0$ for all i. Two cases have been considered – the sum (more generally, an arbitrary linear combination) and the maximum.

The training bank is denoted by the superscript t and is assumed to become selectively responsive to some pattern from its environment E^t. This pattern is the **trigger feature** of the unit and is denoted by s^{trig}. The partial sum and partial response induced by the trigger feature are correspondingly labelled x^{trig} and r^{trig}.

Modification dynamics: the learning rule

The self-supervised learning (SSL) rule is expressed in terms of the time derivatives of the connectivity values w_j^i in terms of the corresponding afferent activity s_j^i, two partial responses (that of the bank to which w_j^i belongs and another that is produced by a special "training bank"), and a variable q that is driven by the training bank's partial response.

$$\Delta w_j^i = \alpha(x^t - q\sigma(x^i))\, s_j^i \tag{4}$$

$$\Delta q = \alpha x^t (x^t - q)$$

where the learning rate α is a small number and the superscript t specifies the *training bank*, such that the partial response x^t "trains" the other partial responses $\{r^i ; i \neq t\}$ to approximate it to the degree that the pattern s^i predicts the pattern to the training bank s^t. The function σ is monotone increasing and satisfies the conditions given by (3).

Final states of the training bank

The SSL equation (4) reduces to the selectivity maximization rule of Bienenstock et al. (1982) along the training bank; the function σ as constrained by Eq. (3) is included to ensure this. For the training bank, equation (4) becomes

$$\begin{aligned} \Delta w_j^t &= \alpha(x^t - q\sigma(x^t))\, s_j^t \\ \Delta q &= \alpha x^t (x^t - q) \end{aligned} \tag{5}$$

The response r^t achieves very high selectivity over the environment E^t. Under the assumption of linear independence within the subenvironments, the training bank attains maximum selectivity; i.e. it responds to exactly one pattern in E^t. Let the chosen pattern, i.e. the trigger stimulus of the training bank, be denoted by s^{trig} and let the corresponding partial sum and partial response [i.e. $\mathbf{w}^t \cdot s^{trig}$] be respectively denoted by x^{trig} and r^{trig}.

Final states of the trained banks

Consider a trained $(i \neq t)$ bank for which the corresponding subenvironment E^i consists of linearly independent patterns. Stable equilibria can then be found by setting the expression (4a) for

Δw_j^i to zero for each pattern s in E^i. If p is the conditional probability that s^{trig} is present on the training bank given s^i at bank i, then for all patterns $s \in E^i$:

$$p(x^{trig} - q\sigma(x^i)) + (1-p)(-q\sigma(x^i)) = 0 \tag{6}$$

If the training bank has reached equilibrium then $x^{trig} = q$ and hence,

$$p = \text{Prob}(s^i = s^{trig} \mid s^i) = \sigma(x^i) \tag{7}$$

REPRESENTATIONS OF STIMULI ACROSS POPULATIONS OF C-UNITS

Up until this point, the description and analysis of SSL has been confined to the single-unit level. While this is appropriate for presentation of the learning mechanism, it is inadequate for understanding certain more global properties, such as those pertaining to the representation of the current state of the world.

If the number of information streams is restricted to just two and all patterns within their subenvironments are equiprobable, then this observation follows concerning the total activity level of the population: *The sum of the unit activities evoked by a given presentation across two information streams decreases with increasing joint probability of the stimulus combination.* That is, the net activity of the population is correlated with the novelty of the stimulus. The relationship between the net activity

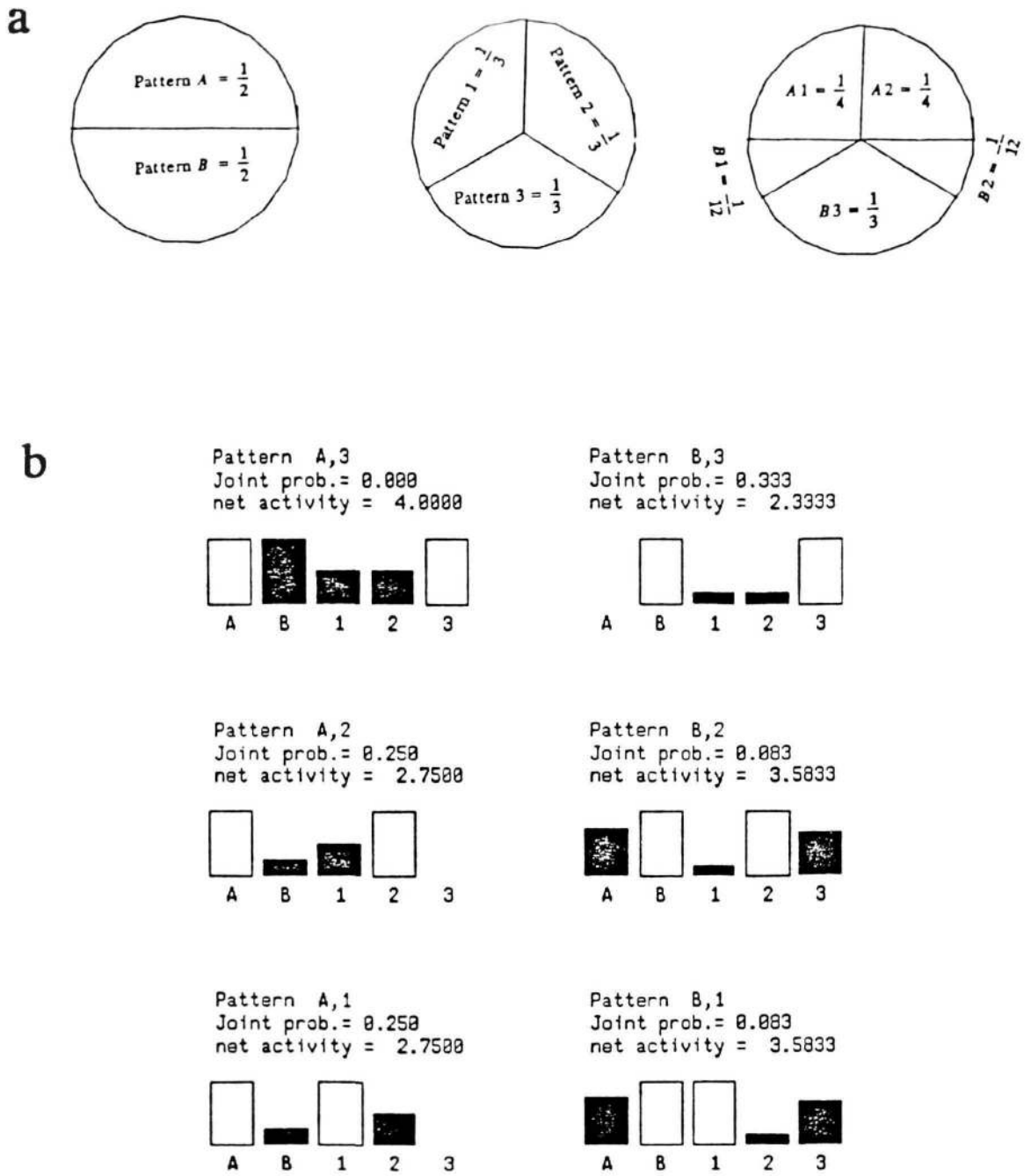


Figure 2. ***A complete representation of a 2x3 environment.*** **a.** The subenvironment E^I consists of the two equiprobable patterns A and B, E^{II} consists of the three equiprobable patterns 1, 2, and 3. Their joint probabilities are shown in the pie chart and the table. **b.** The five subpatterns (A,B,1,2,3) are each the trigger feature for one of the C-units in this minimally complete population. The C-unit representation is shown as an activity pattern over these units for each of the six possible joint patterns, together with the pattern's joint probability and the net activity elicited in the population. For every joint pattern, at least two units correspond to the constituent subpatterns and are thus maximally active. These two components of the activity pattern are shown as unshaded bars, while the "associative" components are shown as shaded bars.

and the joint probability can be shown to be: (by Baye's theorem):

$$A_{TOT}(ij) = 4 - (N_I + N_{II})\, p_{ij} \tag{8}$$

An example system In this example, a population has come to equilibrium with two pattern streams. The respective subenvironments E^I and E^{II} consist respectively of two and three equiprobable patterns: $E^I = \{A,B\}$; $E^{II} = \{1,2,3\}$. Only five $(N_I + N_{II})$ equilibrium states are possible, so for convenience consider a population of just five units, each having converged to a different equilibrium state. Thus the units can be labelled according to the stimulus selected along their training bank. The statistics of the environment and the representations of joint stimuli across the population are described in Figure 2. This simple example illustrates several basic aspects of distributed representations by high-order units. Note that for each pattern, the net activity plus 5 times the joint probability is 4 and hence Eq. (8) is verified.

REFERENCES

Amari, S. & Takeuchi, A. (1978) Mathematical theory on formation of category detecting nerve cells. *Biol. Cybern.* 29:127-136

Bienenstock, E., Cooper, L., and Munro, P. (1982) Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *J. Neurosci.* 2:32-48

Reilly, D., Cooper, L., and Elbaum, C. (1982) A neural model for category learning. *Biol. Cybern.* 45:35-41

Rumelhart, D. & Zipser, D. (1985) Feature discovery by competitive learning. *Cognitive Science* 9:75-112

Sutton, R. & Barto, A. (1981) Toward a modern theory of adaptive networks: expectation and prediction. *Psych. Rev.* 88:135-170

ANATOMISING LEXICAL DECISION IN PHRASAL CONTEXTS: WHEN DOES TRUCK NOT PRIME CAR?

Padraig G. O'Seaghdha
University of Rochester

ABSTRACT

Context effects on lexical decisions were anatomised by manipulating lexical relatedness in syntactic and asyntactic sequences. In a Syntactic condition, related or unrelated word-pairs were embedded in simple sequences (e.g., a truck or a CAR/FLOOR). In a Scrambled condition, two inapposite function words were substituted between the related and unrelated nouns (e.g., the truck that before CAR/FLOOR). The phrases were presented serially and subjects made lexical decisions to their terminal elements. Substantial relatedness effects were found only in syntactic sequences, whether presentation rate was slow or whether it exceeded the rate of normal reading. The syntactic relatedness effect was shown to consist, in equal proportions, of facilitation of related words and inhibition of unrelated words. These results argue against a role for intralexical priming in on-line reading. They point up the roles of syntactic connectedness and of the current interpretation even in very rudimentary contexts.

INTRODUCTION

The phenomenon of lexical priming -- facilitation of the perception of one word in the context of another -- has been extensively studied in the past 15 years (e.g., Fischler & Goodman, 1978; Meyer & Schvaneveldt, 1971; Neely, 1977). More recently, researchers have examined how responses in lexical tasks, primarily lexical decision and naming, are modulated in sentence contexts (e.g., She cleaned the dirt from her SHOES/HANDS/TERMS). These studies show facilitation of congruent completions (SHOES) and inhibition of incongruent completions (TERMS), relative to appropriate but unlikely completions (HANDS) or other control conditions. Larger inhibition effects have usually been observed in the lexical decision than in the naming task (see Fischler & Bloom, 1979, 1980; Stanovich & West, 1979, 1983).

The main theoretical interest of these studies lies in the facilitation effects. Do they represent a genuine influence on lexical access of the kind that is thought to occur in single-word contexts (e.g., truck-CAR)? One recent view, motivated by a conception of the lexicon as an encapsulated system or module (Forster, 1979; Stanovich & West, 1983), is that genuine facilitation effects depend on the presence of lexically related words in the contexts. For example, mailed and letter might prime STAMP in the sentence He mailed the letter without a STAMP.

TABLE 1
EXAMPLES OF COUNTERBALANCED MATERIALS

Context Words				Related/Unrelated Target
1	2	3	4	
Syntactic (All Experiments)				
the	author/owner	of	this	BOOK/FLOOR
the	ceiling/circle	and	the	FLOOR/BOOK
Scrambled (Experiments 1 and 2)				
the	author	the	and	BOOK/FLOOR
the	ceiling	this	of	FLOOR/BOOK

On this view, activation in the lexicon occurs independently of concurrent interpretive processes. That is, facilitation effects resulting from such activation are intralexical. Such a process would not account for facilitation in contexts that lack lexical associates (e.g., She cleaned the dirt from her SHOES). This kind of facilitation is ascribed by Forster (1979, 1981) and others to a positive response bias following the occurrence of a predictable target. Thus two sources of facilitation are identified -- priming in a lexical store, and a response bias determined by predictability.

Because semantically related words are rarely adjacent in connected text, and because the process of reading is probably too fast to allow the development of predictions, some authors (e.g., Mitchell, 1982) have concluded that contextual facilitation plays little or no role in normal skilled reading. This conclusion may be premature, however. First, the proximity of related words may not be critical. Foss (1982) showed that phoneme monitoring latencies were facilitated in contexts containing related words, but not when the same related words were presented in randomised texts. Further, when the contexts were coherent, facilitation was not affected by the distance separating the priming and target words. He interpreted this as evidence that latencies were not determined by intralexical priming but by the facility with which new material was integrated in a text-level representation.

Secondly, although explicit expectancies would take time to develop, it is not clear that facilitation in sentence contexts requires such specificity. Although some researchers have argued that it does not occur at rapid rates, Fischler and Bloom (1980) observed facilitation when contexts were read at rates well in excess of normal reading speed. Also, Fischler and Bloom (1979) found no cost to less likely completions in highly constraining contexts. Such costs should be incurred if the benefits to the most likely completions depended on explicit prediction.

However, if facilitation reflects the more efficient integration of new material into a sentence or text-level representation, the cost-free outcome makes sense.

Control of materials is a problem in assessing lexical and nonlexical contributions to facilitation effects in sentence frame contexts. In particular, some sentences contain lexical relatives of the targets, but others do not. In the experiments reported here, this problem was addressed by defining relatedness in terms of a single content word in each context (e.g., a truck or a CAR; see other examples in Table 1). Thus, an equivalent level of control to the simple one-word context case was achieved. With these materials it was possible first to address the intralexical issue, and then to reassess the nature of the facilitation effect.

EXPERIMENTS 1 AND 2: ARE THERE INTRALEXICAL EFFECTS?

A series of lexical decision experiments where lexical relatedness and syntactic connectedness were independently manipulated was conducted. That is, related and unrelated noun pairs were embedded in syntactic or scrambled sequences. The scrambling consisted in substituting anomalous function words between the nouns (see examples in Table 1).

The sequences were presented serially, word by word, at a fixed location in the center of a CRT. Subjects made lexical decisions to the last word in each sequence which was printed in uppercase letters. There were eighty word and forty nonword filler trials in an experiment.

Experiments 1 and 2 were designed to assess the extent of relatedness effects in the syntactic and scrambled sequences. In Experiment 1, the contexts were presented at a slow, 800 msec per word, rate. Under these conditions, deliberate prediction and exploitation of lexical relatedness might occur in both kinds of context. In Experiment 2, the 200 msec rate exceeded that of normal reading, thus giving intralexical activation a better chance to show, and at the same time addressing the concern with the relevance of lexical priming at reading speed.

The results of both experiments are summarised in Table 2. In the case of Experiment 1, the responses of forty-eight subjects to ten targets are represented in each estimate. (These subjects also served in a second block involving an additional manipulation not reported here. See O'Seaghdha (1986) for details). In Experiment 2, twenty-four subjects responded to twenty targets in each subcondition.

Both experiments show a clear relatedness effect in the syntactic contexts, but little or no effect in the scrambled condition. There is a nonsignificant 13 msec effect in the scrambled condition at the 200 msec rate, but the contrast between this effect and that in the syntactic condition leads to the conclusion that the relatedness effects in the latter cannot be attributed to intralexical priming among related words.

Of course the relatedness effect in the syntactic condition

TABLE 2
EXPERIMENTS 1 AND 2: MEANS OF BIWEIGHT ESTIMATED REACTION TIMES WITH STANDARD ERRORS AND PERCENTAGE ERRORS AS A FUNCTION OF CONTEXT CONDITION AND RATE

Experiment 1: 800 msec rate

	Syntactic		Scrambled	
	Related	Unrelated	Related	Unrelated
	512	561	532	534
	9.1	12.2	10.5	8.3
Error%	1.7	4.6	1.3	2.9

Experiment 2: 200 msec rate

	Syntactic		Scrambled	
	Related	Unrelated	Related	Unrelated
	495	537	512	525
	11.8	16.6	12.3	11.6
Error%	1.5	4.2	2.0	3.1

must be partly inhibitory. Substantial inhibition of anomalous targets is reliably obtained in lexical decision experiments. An informal comparison of latencies in the related and unrelated syntactic contexts to latencies in the nonpriming scrambled condition indicates that the overall effect divides equally into facilitation and inhibition components. A more direct evaluation of these components is provided in Experiment 3.

EXPERIMENT 3: FACILITATION AND INHIBITION

In Experiment 3, the relatedness effect in syntactic sequences was partitioned into facilitation and inhibition components. For this purpose, a neutral condition containing unrelated context nouns was developed (see Table 1). In this condition, the phrases made sense (e.g., the owner of the BOOK), but there was no lexical relation between the contextual content word (owner) and the target (BOOK). The related and unrelated conditions were the same as in Experiments 1 and 2, and the procedure was identical. Three presentation rates -- 200, 400, and 800 msec -- were tested, and sixteen subjects were run at each rate. The questions in Experiment 3 were whether facilitation occurred, relative to the neutral context, and whether the division into facilitation and inhibition varied with rate of presentation.

The evidence shown in Table 3 confirms that the relatedness effect consists, in approximately equal proportions, of facilitation and inhibition. The inhibition of anomalous targets

TABLE 3
EXPERIMENT 6: MEANS OF BIWEIGHT ESTIMATED REACTION TIMES WITH STANDARD ERRORS AND PERCENTAGE ERRORS AS A FUNCTION OF CONTEXT CONDITION AND RATE

		Context Condition		
Rate		Related	Neutral	Unrelated
200 msec		495	512	534
		10.7	10.4	6.9
	Error%	0.9	1.9	5.6
400 msec		491	519	530
		15.9	13.2	17.6
	Error%	0.9	2.5	3.1
800 msec		476	503	526
		15.3	17.4	20.2
	Error%	1.6	1.2	3.8

replicates previous findings in the literature, and was expected. However, contrary to what previous readings of the literature (e.g., Mitchell, 1982) would suggest, facilitation did not vary substantially with rate of presentation. If there were little facilitation at the 200 msec rate, but increasing facilitation at slower rates, the interpretation of this experiment would be straightforward. Facilitation would then be attributable to a postlexical response bias. Because this outcome was not obtained, it appears that the emphasis in the literature on slow-acting predictive processes as the source of facilitation may be misplaced.

DISCUSSION

Neither intralexical priming (e.g., Stanovich & West, 1983), nor a response bias favouring predictable targets (Forster, 1979, 1981), accounts for the facilitation effects in these experiments. An alternative account of facilitation effects in phrasal or sentence contexts is therefore required.

One candidate is Foss's (1982) proposal that facilitatation in his phoneme monitoring experiments reflected the speed of integration of new material with a discourse level representation of the texts. This idea seems compatible with the present data. In addition, it is consonant with the notion of a positive response bias with appropriate sentence completions in a lexical decision task (see Forster, 1981), but this bias is characterised as a byproduct of integration, not as a consequence of explicit

prediction of targets. Further, the invariance of performance with rate of presentation suggests that it reflects normal processes of comprehension and not merely the special requirements of the lexical decision task.

The main point demonstrated in these experiments is that syntactic connectedness is prerequisite for the occurrence of relatedness effects. It might be urged that lexical priming is somehow "transmitted" in syntactic sequences. This must be true, at least in a metaphorical sense. However, such a view would still require a revision of the view of facilitation as the product of a blind process of spreading activation. Perhaps the most important implication of the research is that, if relatedness effects are as sensitive to syntactic coherence as these data indicate, the interdependence of parsing and interpretive operations requires further detailed analysis. We know one circumstance where truck does not prime car. The adequacy of our understanding of the cases where priming does occur is therefore brought into question.

ACKNOWLEDGMENTS

The research reported here is extracted from the author's Doctoral Dissertation. I thank Colin MacLeod, Paul Muter, and Gary Dell for comments.

REFERENCES

Fischler, I. & Bloom, P. A. (1979). Automatic and attentional processes in the effects of sentence contexts on word recognition. Journal of Verbal Learning and Verbal Behavior, 18, 1-20.

Fischler, I. & Bloom, P. A. (1980). Rapid processing of the meaning of sentences. Memory & Cognition, 8, 216-225.

Fischler, I. & Goodman, G. O. (1978). Latency of associative activation in memory. Journal of Experimental Psychology: Human Perception and Performance, 4, 455-470.

Forster, K. I. (1979). Levels of processing and the structure of the language processor. In W. E. Cooper & E. Walker (Eds.), Sentence processing: Psycholinguistic studies presented to Merrill Garrett (pp. 27-85). Hillsdale, NJ: Erlbaum.

Forster, K. I. (1981). Priming and the effects of sentence and lexical contexts on naming time: Evidence for autonomous lexical processing. Quarterly Journal of Experimental Psychology, 33A, 465-495.

Foss, D. J. (1982). A discourse on priming. Cognitive Psychology, 14, 590-607.

Meyer, D. E. & Schvaneveldt, R. W. (1971). Facilitation in recognising pairs of words: Evidence of a dependence between retrieval operations. Journal of Experimental Psychology, 90, 227-234.

Mitchell, D. C. (1982). The process of reading. London: Wiley.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited capacity attention. Journal of Experimental Psychology: General, 106, 226-254.

O'Seaghdha, P. (1986). The dependence of lexical relatedness effects on syntactic connectedness. Unpublished Doctoral Dissertation, University of Toronto.

Stanovich, K. E. & West, R. F. (1979). Mechanisms of sentence context effects in reading: Tests of a two-process theory. Memory & Cognition, 7, 77-85.

Stanovich, K. E. & West, R. F. (1983). On priming by a sentence context. Journal of Experimental Psychology: General, 112, 1-36.

STRATEGY SELECTION AND DEGREE OF EXPERTISE IN MEDICAL REASONING

Vimla L. Patel

Jose F. Arocha

Guy J. Groen

McGill University

The psychological theory of expert problem solving has been established in relatively clearcut domains, such as physics, with problems that do not involve the processing of complex verbal information. There is no apriori reason why the results should extend to more complex domains. One of the more important notions of the theory is that experts tend to use a strategy of forward chaining based on a highly elaborate knowledge base. The purpose of this paper is to consider the extent to which this is true in a more complex and "messy" domain.

It is widely held in the area of medical problem solving that experts do not use forward reasoning of this type. Instead, they are presumed to use a process of generating and testing hypotheses that involves a large component of backward reasoning (e.g., Kassirer, 1984) even though, as Groen and Patel (1985) have pointed out, this is largely unsupported by empirical evidence with the exception of some results reported by Elstein, Shulman and Sprafka (1978). In contrast, Patel and Groen (1986) have found evidence of forward chaining in a task involving an explanation of the reasoning involved in diagnosing a clinical case. In general, the subjects (specialists in cardiology) who yielded accurate diagnoses used a process of forward reasoning, from the clinical facts to the final diagnosis. The use of backward reasoning was associated with inaccurate diagnoses.

There are a number of reasons why this might be the case. Elstein et al's method of "stimulated recall" is retrospective by nature and may induce backward reasoning as a necessary strategy. A more important possibility, which has the advantage of being consistent with the intuition of physicians regarding the widespread use of backward reasoning, is that it is used when there is uncertainty about the diagnosis. Part of its importance stems from the fact that it seems reasonable to assume that it is a general phenomenon not restricted to medicine. It is well-known (Bhaskar & Simon, 1977; Hayes, 1985) that problem solvers use different strategies depending on the matching of the task to expertise. Glaser (1985) suggests that when a task is difficult, expert problem solvers use general methods working backward from the goal rather than domain-specific methods working forward from the data given. Clearly, an attempt to solve a difficult problem is usually associated with a feeling of uncertainty. It is possible to assume that a feeling of uncertainty may give rise to the strategies associated with solving difficult problems. This kind of uncertainty is not necessarily a vague subjective state. Expert systems for medical diagnosis such as MYCIN and INTERNIST generate and manipulate uncertainty factors. Physicians may use them in an analogous fashion.

In this paper, we consider two factors that might be expected to result in greater uncertainty regarding aspects of a clinical case. The first is the difficulty of the case. The second is the match between the physician's expertise and the knowledge domain underlying the clinical case. We consider two aspects of this latter factor: the physician's specialty and research orientation (i.e., whether primarily a clinician or a basic researcher). Our main interest is in testing the generality of the findings of Patel and Groen (1986) by determining the extent to which these variations perturb the pattern of forward reasoning found in that study. Apart from these variations in choice of subjects and clinical cases, the procedure and method of analysis are identical to the previous study. The experiment reported in the paper by Joseph and Patel elsewhere in this volume studies the same variations but with a completely different empirical procedure. However, it tends to

induce backward chaining, because only partial information is available to the subject. Hence it is unsuitable for examining the issues of concern in the present paper.

EXPERIMENTAL METHOD

Subjects

Four expert physicians associated with the Faculty of Medicine volunteered for the study. The four physicians were two Endocrinologists (1 practitioner and 1 researcher) and two Cardiologists (1 practitioner and 1 researcher).

Clinical Problems

Two texts were constructed describing the history, physical examination, and laboratory tests of two patients. The first text which is shown in Table 1 describes a case of a 63 year old woman who suffers from an endocrine disorder called Hashimoto's thyroiditis, a form of hypothyroidism of autoimmune origin. It can be decomposed into three diagnostic components. The most general and prototypical component is hypothyroidism. This is indicated by the textual cues suggesting fluid accumulation and decreased thyroid function. The second component, myxedema, indicates that the patient is in an advanced state of hypothyroidism. The clinical cues that comprise the third component are suggestive of a very specific origin for the disease process which is an autoimmune process known as Hashimoto's thyroiditis.

The second clinical text which is shown in Table 2 describes a case of a 62 year old man who was diagnosed as having cardiac tamponade with pleural effussion. This case is more difficult than the endocrinology case because there is more overlap of knowledge among causal patterns leading to alternative diagnoses. In order to diagnose the case, the physician must decide whether the problem is left or right sided failure and then identify the presence of pericardial effusion and cardiac tamponade. Determining the actual causal process (right heart failure) is a

TABLE 1

ENDOCRINOLOGY CASE

A 63 year old woman with a one-week history of increasing drowsiness and shortness of breath was brought to the emergency room by her daughter. The patient had not been well for over a year. She complained of feeling tired all the time, had a loss of appetite, a 30 lb. weight gain and constipation. A month later she had been diagnosed as having " chronic laryngitis" and was prescribed a potassium iodide mixture as an expectorant.

Physical examination revealed a pale, drowsy, obese lady with marked periorbital edema. She had difficulty speaking, and when she did speak her voice was noted to be slow and hoarse. There were patches of vitiligo over both her legs. Her skin felt rough and scaly. Her body temperature was 36 deg. C. Pulse was 60/minute and regular. B. P. was 160/95. Examination of her neck revealed no jugular venous distention. The thyroid gland was enlarged to approximately twice the normal size. It felt firm and irregular. There was grade 1 galactorrhea. The apex beat could not be palpated. Chest examination showed decreased movements bilaterally and dullness to percussion. There was no splenomegaly. Neurological testing revealed symmetrical and normal tendon reflexes but, with a delayed relaxation phase. Urinalysis was normal. Chest X-ray showed large pleural effusions bilaterally. ECG revealed sinus bradycardia, low voltage complexes and non-specific T-wave flattening. Routine biochemistry (SMA=16) showed Na=125, K=3.8, BUN=8 mg/100ml. Arterial blood gases P02=50 mm Hg, PCO2=60 mm Hg. The patient was admitted to the intensive care unit for further management.

TABLE 2

CARDIOLOGY CASE

This 62 year-old retired Air Force mechanic was apparently well until about 5 months before presenting to the hospital. He then noted he was "winded" after walking about 40 feet. He was increasingly breathless lying down, tried using 4 pillows to sleep and most recently is sleeping sitting up. He has occassionally awoken extremely short of breath. He has a mild non-productive cough and agrees that his voice is a little hoarse. During this time his legs have been swelling. His appetite has decreased yet his abdomen has increased and he has gained weight. He says "no food tastes good" and he has constant mild nausea but has not vomited. He has had no chest or abdominal pain. He does not smoke, drinks alcohol socially but less lately. His only admission to the hospital was for a heart attack 12 years ago. He recovered completely and was walking 6 miles a day a year ago. He is taking no medication.

On examination: H.R. 80/min. and regular. B.P. 120/98 mm Hg. Pulsus paradoxicus 12 mm Hg. No cyanosis. Pronounced peripheral edema of legs and presacrum. Some edema over abdominal wall and scrotum. Abdomen was large with shifting dullness and a fluid wave was demonstrated. Liver edge was smooth, 3 cm. below the right costal margin. Spleen was not palpated. No masses. Jugular veins distended to the angle of the jaw at 45 deg.; apex not palpable, heart sounds faint, no S3, no S4, no murmurs. Some dullness to percussion at right lung base. Breath sounds diminished at both lung bases with decreased chest expansion. Fine end inspiratory crepitutions noted. Remainder examination was normal.

Hb=13.5gm % (percent), WBC=5,500 with a normal differential. Prothrombin time 12.5 (control 11.8), P.T.T. 34 (control 34), T4=7.5 (normal 4.5-10.5). Urinalysis was normal except urobilinogen 4.0 (normal 0.1-1.0); SMAC 16 normal except: Albumin 3.5 (N= 3.7-4.9), total bilirubin 1.7 (N 0.2-1.0); alkaline phosphotase 169 (N 30-105),. Chest X-ray: "Enlarged cardiac silhouette; no evidence of pulmonary edema, right pleural effusion, partial atalectasis in right lower lobe. " ECG: remote inferior myocardial infarction. Diffuse ST sagging with T-wave inversion. Generally low voltage QR's with voltage fluctuation.

This patient has been referred from an outlying hospital for definitive management.

difficult task because many of the features the patient presents with are common to different diagnostic possibilities. There are, however, a few cues which serve to rule out alternative diagnoses.

Procedure

The following empirical paradigm was used:

1) Present the subject with a description of the case and ask him/her to read it for 3.5 minutes;
2) Obtain a free recall protocol (a written summary of the case);
3) Ask the subject to describe the underlying pathophysiology of the case;
4) Ask for a diagnosis.

The subjects were tested individually. Each protocol was analyzed according to the techniques of propositional and frame analysis developed by Frederiksen (1985). The outcome was a structural representation containing nodes and links expressing mostly conditional and causal relationships between propositions. The reader is referred to Patel and Groen (1986) for a detailed account of these analytical procedures.

RESULTS AND DISCUSSION

All subjects were able to summarize the information in the text presenting the clinical case irrespective of their specific areas of expertise. Both Cardiologists and Endocrinologists were able to select the relevant cues and provide a summary of the most important features of the text even when an inaccurate diagnosis was ultimately given. They repeated a high percentage of the summary information in the pathophysiology when solving the patient problems in their own specialty but did not do so when solving the case outside their area of specialization. Practitioners, however, showed the opposite effect, where fewer cues were used from the summary to explain the pathophysiology of the problem in their own domain but used more cues to explain the problem outside their domain of expertise.

In this paper, we concentrate on the pathophysiological explanations contained in four protocols which indicate clearly the pattern of forward and backward chaining that emerged. Figure 1 shows the structural representation of the protocol of the endocrinology practitioner in the endocrinology problem. The square boxes indicate information in the text that presented teh clnical case. The labelled arrows indicate causal and conditional rules, as defined in Patel and Groen (1986). The order in which the propositions appear in the figure, from the top to the bottom, reflects the order in which they appear in the protocol.

First, the subject identifies the disease in a top level rule which provides a diagnosis without generating explicit intermediate mechanisms. Then, he describes some general rules which explain the mechanisms common to hypothroid patients, (i.e., hypoventilation, pleural effusion and hyponatremia and the effect of iodide administration on the precipitaton of the problem). The pathophysiology presents very little text information and it is constructed to justify a diagnosis. The order of reasoning is completely forward except possibly at the very end of the protocol, where a condition related to "hypometabolic state" is explained in terms of a possible observable outcome ("respiratory failure").

Figure 2 shows the structural representation of the protocol of the endocrinology reseacher in the same problem. This also shows a pattern of forward chaining except once again at the end of the protocol where a number of loose ends are explained. However, the protocol as a whole is different in that it contains a great amount of detailed information from the basic sciences clustered around affected body systems. Unlike the clinicians, the top level clinical rules are not used in the explanations.

Figure 3 shows the structural representation of the protocol of the cardiology practitioner on the cardiology problem. For various reasons, as already indicated, this was a more difficult case and nobody obtained a completely correct diagnosis. This was primarily due to the fact that the protocol did not specify the results of a critical test which would have unequivocally indicated the presence of the actual disorder. Both cardiologists indicated in their explanations that this

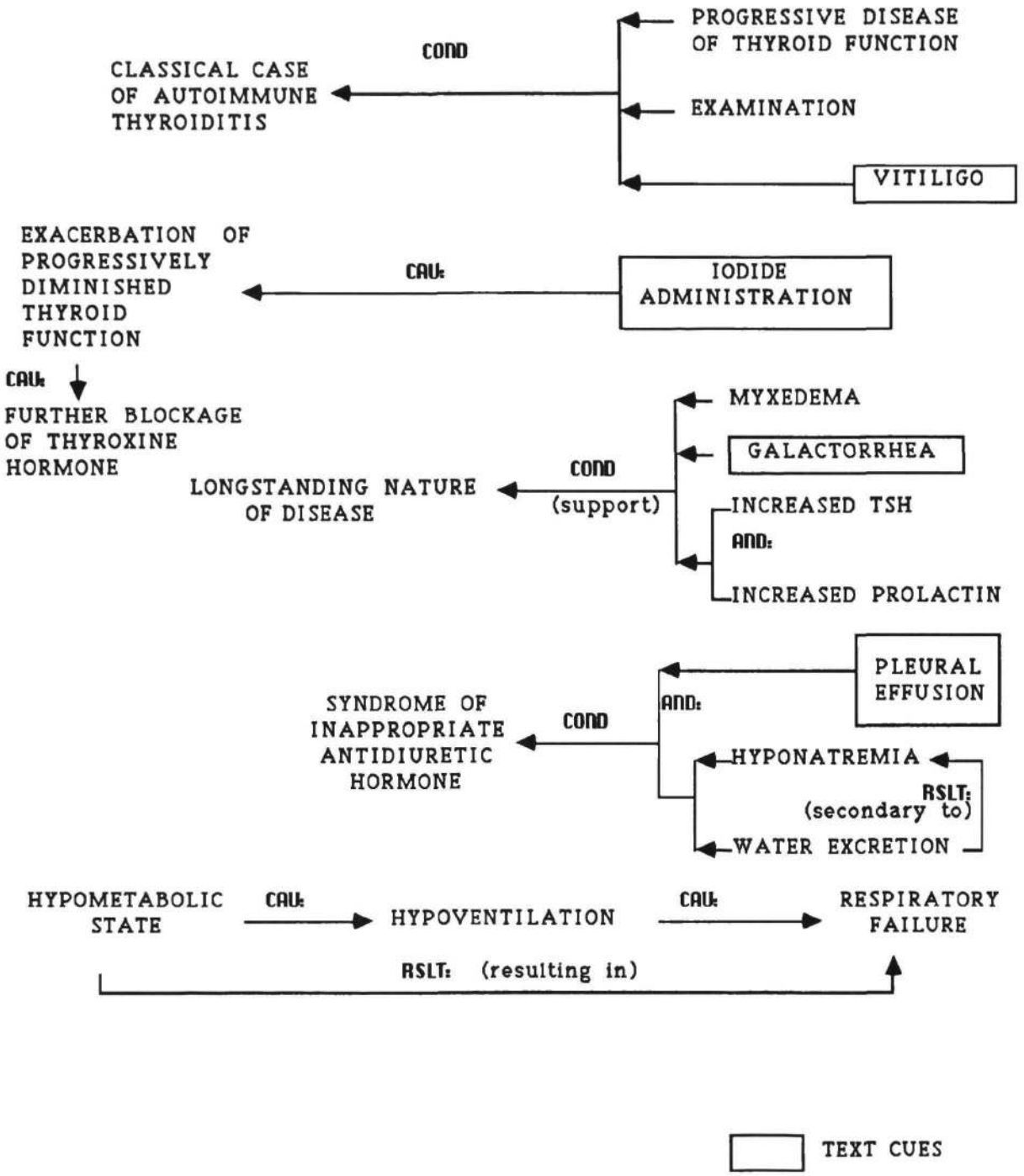


FIGURE 1

STRUCTURAL REPRESENTATION OF ENDOCRINOLOGY PROBLEM BY ENDOCRINOLOGIST PRACTITIONER

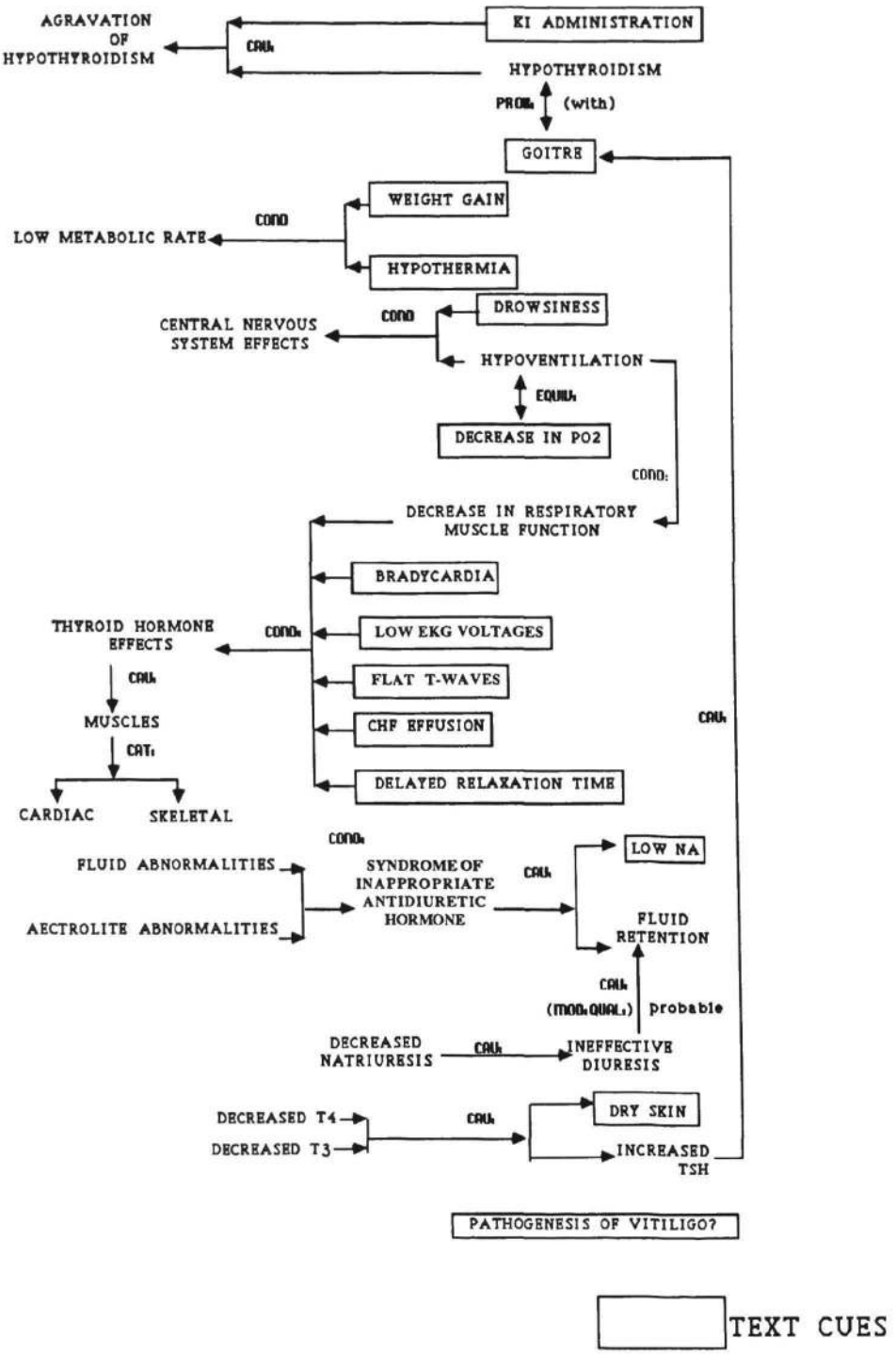


FIGURE 2

STRUCTURAL REPRESENTATION OF ENDOCRINOLOGY PORBLEM
BY ENDOCRINOLOGY RESEARCHER

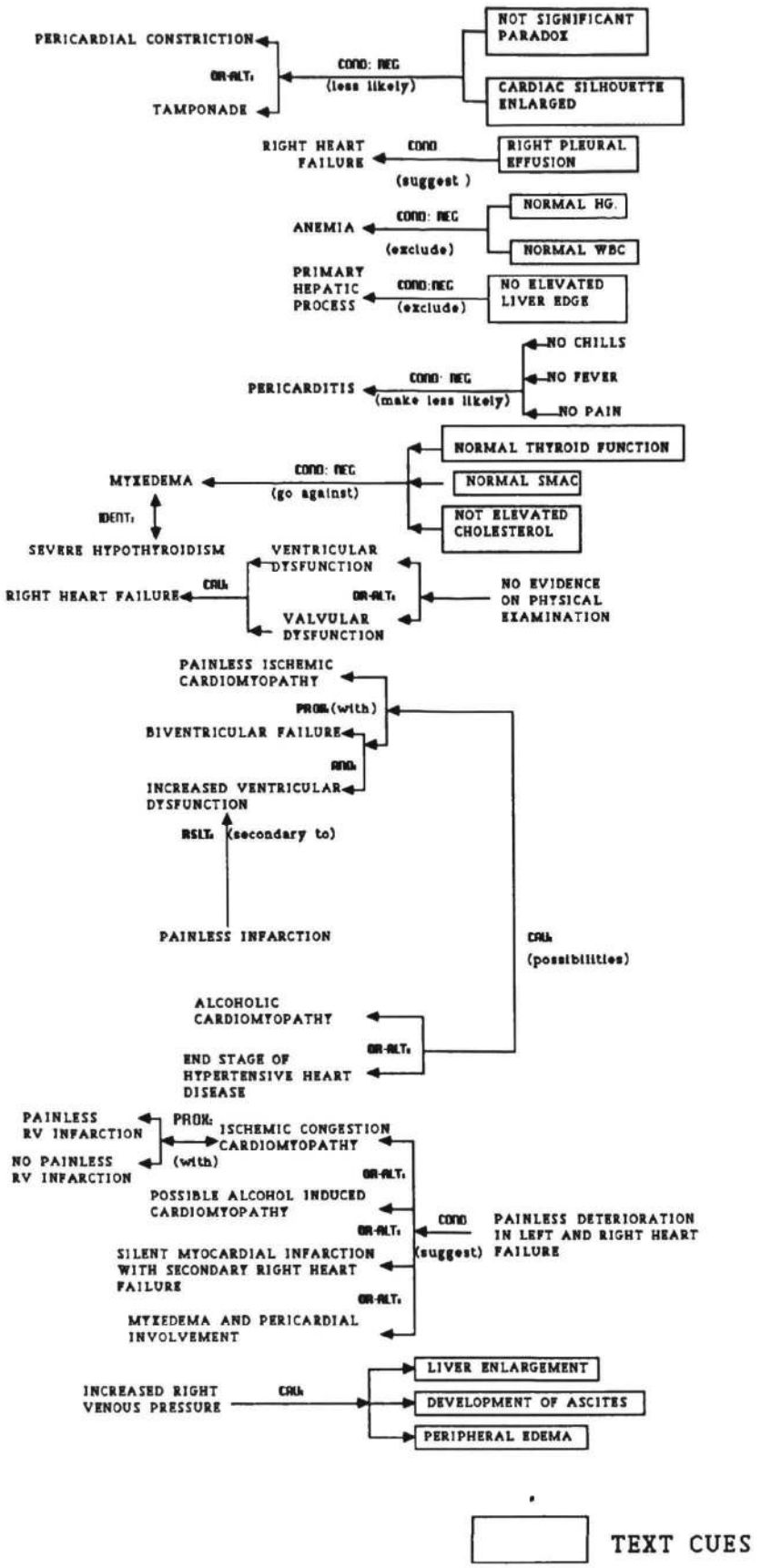


FIGURE 3

STRUCTURAL REPRESENTATION OF CARDIOLOGY PROBLEM BY CARDIOLOGIST PRACTITIONER

information was missing. The reasoning follows the pattern in the first two figures (Figures 1 & 2). The subject uses purely forward reasoning except to account for factors not critical for the diagnosis. However, most of the rules are not used to generate a causal structure but to rule out possible disorders. This is accomplished before proceeding to the explanation of the actual disease process.

Figure 4 is included to give an example of almost pure backward reasoning. This is the structural representation of the protocol of the cardiologist researcher on the endocrinology problem. It shows a pattern that begins with a general endocrine disorder and attempts to proceed from there to specific symptoms. However, most of these symptoms are stated in general physiological terms. Only two specific symptoms described in the text are used in the explanation. The diagnosis was inaccurate.

Of the protocols not shown here, all contained more of a mixture of backward and forward reasoning and hence are difficult to describe in a concise fashion. They were all associated with inaccurate or incomplete diagnoses. As was the case with the protocols we discussed above, the researchers tended consistently to make more use of causal patterns deriving from basic science issues.

In general, these results are consistent with those obtained by Patel and Groen (1985) and also serve to extend them. Clinical practitioners use forward reasoning in their own area of specialization when they feel certain about their diagnoses. In the case of Figure 2, we have seen that it can occur even when only a partial diagnosis is possible. More insight regarding the use of partial information is given by the Joseph and Patel study elsewhere in this volume.

This can be taken to imply that case difficulty alone is not a predictor of the selection of forward and backward reasoning strategies. Because in this study we also obtain a strong relationship between accuracy of diagnosis and degree of backward chaining (when the issue of the missing test in the cardiology case is factored out), it seems that the degree of uncertainty is a more important factor. The main difference between researchers and clinicians does not seem to lie as

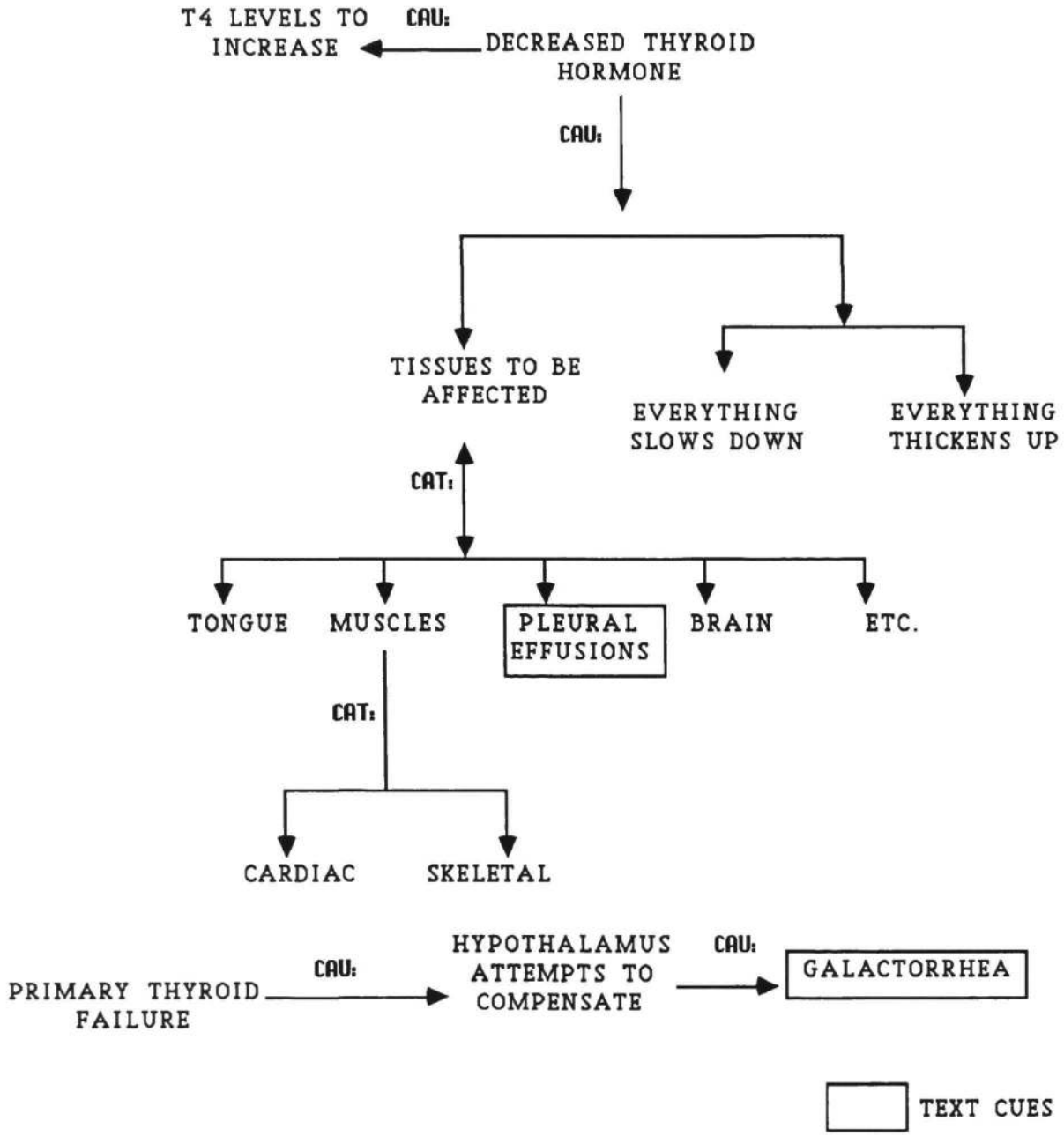


FIGURE 4

STRUCTURAL REPRESENTATION OF ENDOCRINOLOGY PROBLEM
BY CARDIOLOGIST RESEARCHER

much in their direction of reasoning as in their greater interest in issues stemming from their involvement in the biomedical sciences.

These results may have some implications for a more adequate theory of general problem solving. Currently this has two well developed components. The first is a theory of heuristic problem solving in puzzle-like domains as developed, for example by Newell and Simon (1972) and the second is theory of expert problem solving. What is lacking is a theory that accounts for the problem solving behavior of novices and other less expert subjects in knowledge based domains. The research strategy we have pursued in this paper has allowed us to make use of a well developed theory of experts, yet by placing experts in situations where they cannot use all their knowledge, we are able to examine their performance at a sub-expert level. This may provide a means of arriving at a first approximation to a more adequate theory of sub-expert knowledge based problem solving.

ACKNOWLEDGEMENT

Work on this project was supported in part by a grant from Josiah Macy Jr. Foundation (Grant no. B8520002). The authors would like to acknowledge the physicians, Hugh Scott and Yogesh Patel, for their contributions and thank the Endocrinologists and Cardiologists who took part in the study. Finally, we would also like to acknowledge Guy-Marie Joseph for data collection.

REFERENCES

Bhaskar, R. & Simon, H.A. (1977) Problem solving in semantically rich domains. Cognitive Science, **1:** 193-215

Elstein, A.S., Shulman, L.S., & Sprafka, S.A. (1978) Medical Problem Solving. Harvard University Press: Cambridge, MASS.

Frederiksen, C.H. (1986) Cognitive Models and Discourse Analysis. In C.R. Cooper and S.Greenbaum (Eds.) Studying Writing: Linguistic Approaches. Sage Publications, Beverly Hills, CA.

Glaser, R. (1985) Thoughts on Expertise. Technical Report # 8, Learning Research and Development Center, University of Pittsburgh: Pittsburgh, PA.

Groen, G.J. and Patel, V.L. (1985) Medical problem solving: some questionable assumptions. Medical Education, **19:** 95-100.

Hayes, J.R. (1985) Teaching Thinking Skills. In S.F. Chipman, J.W. Segal and R. Glaser (Eds.), Thinking and Learning Skills, Vol. 2. Lawrence Erlbaum Associates, Hillsdale, N.J.

Kassirer, J.P. (1984) Teaching clinical medicine by iterative hypothesis testing: Let's preach what we practice. New England Journal of Medicine, **309:** 921-923.

Newell, A. and Simon, H.A. (1972) Human Problem Solving. Englewood Cliffs: McGraw-Hill.

Patel, V.L. and Groen G.J. (1986) Knowledge based solution strategies in medical reasoning. Cognitive Science, **10:** 91-116.

A REAL-TIME ATTENTIONAL-ASSOCIATIVE NETWORK FOR CLASSICAL CONDITIONING OF THE RABBIT'S NMR

Nestor A. Schmajuk
Center for Adaptive Systems
Department of Mathematics
Boston University
and
John W. Moore
Department of Psychology
University of Massachusetts

INTRODUCTION

Schmajuk and Moore (1986) described two real-time attentional models of classical conditioning. In their present form, both models are capable of real-time descriptions of many classical conditioning paradigms. However, the models do not encompass higher-order conditioning paradigms or sensory preconditioning. They also lack performance rules that permit realistic descriptions of conditioned responding in real time. These considerations prompted the development of a new class of models that incorporate higher-order conditioning, sensory preconditioning, and performance rules.

Because of the large amount of data on classical conditioning of the rabbit's nictitating membrane (NM), this preparation is particularly attractive for a formal treatment. Therefore, the present paper contrasts experimental results regarding classical conditioning using the NM preparation, with computer simulations.

SECOND-ORDER ATTENTIONAL-ASSOCIATIVE NETWORKS

This section describes a class of attentional-associative network that can be applied to CS-CS as well as CS-US paradigms.

Consider the case of one CS, CS_i, that predicts event k. Net associative value, $\dot{V}_i^k$, represents the first-order prediction of event k by CS_i. Consider now the case of two CSs, CS_i and CS_r, that predict event k. It is assumed that CS_i predicts k directly by $\dot{V}_i^k$ and indirectly by predicting CS_r, by $\dot{V}_i^r$. In turn CS_r predicts k by $\dot{V}_r^k$. The second order prediction of k by CS_i, is expressed as the product $\dot{V}_i^r \dot{V}_r^k$.

B_i^k, the first- and second-order prediction of event k by CS_i, is

$$B_i^k = (\dot{V}_i^k + \Sigma_r w_i^r \dot{V}_i^r \dot{V}_r^k) \tau_i . \qquad [1]$$

$\dot{V}_i^k$ is the net associative value of CS_i with event k. The sum over the index r involves all CSs with index $r \neq k$. $\dot{V}_i^r$ is the net associative value of CS_i with all CSs with index $r \neq k$. $\dot{V}_i^r$ is the net associative value of all CS with event k. τ_i is the trace of CS_i. The mathematical expression for τ_i is given below. Coefficient w_i^r serves to adjust the relative weights of first- and second- order predictions in paradigms such as conditioned inhibition. In order to avoid redundant CS_i-US and CS_i-CS_i- US associations, $w_i^r = 0$ when $i = r$, and $w_i^r > 0$ when $i \neq r$. B^k, the aggregate prediction of event k made upon all CSs (including the context) with $\tau > 0$ at a given moment, is

$$B^k = \Sigma_i B_i^k . \qquad [2]$$

As described below, variable B^k participates in the computation of $\dot{V}_i{}^k$. In addition, through adequate performance rules, B^{US} determines the topography of the NM response.

THE M-S-S NETWORK

This section describes an attentional-associative network that incorporates variable B^k. Net associative values, $\dot{V}_i{}^k$, are computed with the rules proposed by Moore and Stickney (1980, 1982, 1985; see also Schmajuk and Moore, 1986).

Changes in associative values.

When the CS_i is accompanied or followed by event k, the associative value between CS_i and event k, $V_i{}^k$, increases by

$$\Delta V_i{}^k = \theta \alpha_i \tau_i (1 - V_i{}^k), \qquad [3]$$

and the antiassociative value, $N_i{}^k$, decreases by

$$\Delta N_i{}^k = \theta' \alpha_i \tau_i (0 - N_i{}^k), \qquad [3']$$

When event k does not occur, $V_i{}^k$ decreases by

$$\Delta V_i{}^k = \theta' \alpha_i \tau_i (0 - V_i{}^k) B^k, \quad [4]$$

and $N_i{}^k$ increases by

$$\Delta N_i{}^k = \theta' \alpha_i \tau_i (1 - N_i{}^k) B^k, \quad [4']$$

where α_i is CS's associability, θ ($0 < \theta \leq 1$) is the rate of change in $V_i{}^k$, θ' ($0 < \theta' < \theta$) is the rate of change in $N_i{}^k$, τ_i is the trace of CS_i , and B^k is defined by Equation

2.

The net associative value of CS_i and event k is

$$\dot{V}_{ik} = V_{ik} - N_{ik}. \qquad [5]$$

Changes in associability

The associability of CS_i , α_i , may increase, decrease, or remain unchanged depending on the associative value of CS_i with event k and the associative value of another CS, CS_j , with the same event k.

When CS_i , CS_j , and event k are presented together, and provided that $\dot{V}_{ik} > \dot{V}_{jk}$

$$\Delta \alpha_{ik} = c (1 - \alpha_i) (\dot{V}_{ik} - \dot{V}_{jk}), \qquad [6]$$

where $\dot{V}_{jk}$ is the second highest net associative value with respect to event k of all the CSs present with the CS_i , including the context.

When $\dot{V}_{ik} \geq \dot{V}_{jk}$

$$\Delta \alpha_{ik} = c (0 - \alpha_i) (\dot{V}_{jk} - \dot{V}_{ik}), \qquad [7]$$

where $\dot{V}_{jk}$ is the highest net associative value with respect to k of all the CSs present with the CS_i . Parameter c in Equations 6 and 7 is a constant set $0 < c < 1$.

When all the components of $\Delta \alpha_{ik}$ related to a given CS_i or the US have been computed, they are combined in the expression

$$\Delta \alpha_i = \Sigma_k \Phi_k \alpha_{ik} / \Sigma_h \Phi_h . \qquad [8]$$

The sum over the index k in the numerator involves all the

events present with the CS_i . The sum over the index h in the denominator involves all the events the subject has encountered in previous experiences in the same context, even though they may not be present at the time $\Delta \alpha_i$ is computed. The weighting factors, Φ_k , are selected such that $\Phi_{US} > \Phi_{CS} > \Phi_X$, because the US is presumed to be biologically more significant than the CSs and the context (X).

NM RESPONSE CONDITIONING

During acquisition of the NM conditioned response (CR), percentage of CRs generated in each session increases, CR latency decreases, and CR amplitude increases. (see Gormezano, Kehoe, & Marshall, 1983)

The trace hypothesis

Conditioning is typically more efficacious when the CS precedes the US than when the two are presented together. Hull (1943) proposed that stimuli give rise to traces in the central nervous system that somehow impinge simultaneously on critical loci of learning, despite the non-simultaneous arrangement as observed in the periphery.

It is assumed that a CS_i generates a trace, τ_i , that increases over time to a maximum, stays at this level for a period of time independent of the CS duration, and then gradually decays back to zero.

Formally, trace τ is defined for t <= 200 msec by

$$\tau(t) = CS_{max} (1 - e^{-(k1\ t)}), \qquad [9]$$

where CS_{max} is the maximum intensity of the CS and k1 is a constant, 0 < k1 < 1. Parameter k1 is selected so that the ISI for optimal conditioning is 200 msec.

$\tau(t)$ remains equal to CS_{max} as long as the CS does not

decay. If the CS = 0 and t > 200 msec, τ (t) decays by

$$\tau(t) = CS_{max} (\exp -(k_1 t)), \quad [10]$$

If CS_i is not present 200 msec after its onset, the trace decays to zero by Equation 10.

Performance Rules

Performance rules were selected to relate variable B^{US} to the topography of NM responses.

Time of CR onset is the earliest time t such that

$$\Sigma^{t}_{t'=t_i} \Sigma_j B_j^{US}(t') >= L1 , \quad [11]$$

where t_i denotes the time step at which CS_i onset occurs. The sum over the index j involves B_j^{US} of all CSs with τ_j > 0 , excluding the context. Sum over index t involves all time steps for which τ_j > 0 , starting at the time step when the amplitude of the NM response as defined by Equation 11 equals zero. L1 is a threshold greater than zero. Equation 11 implies that as B_j^{US} increases over trials, CR onset moves progressively to an asymptote determined by L1.

During the CS period, for time steps t > t_i , the amplitude of the NM response, NMR(t), is changed by

$$\triangle NMR (t) = k2 (B^{US}(t) - NMR(t)), \quad [12]$$

where k2 is a constant (0 < k2 < 1).

During the US period, while $B^{US}(t) < \lambda^{US}(t)$, is given by Equation 12. However, when $B^{US}(t) > \lambda^{US}(t)$, NMR (t) increases by

$$\triangle NMR (t) = k2 (\lambda^{US}(t) - NMR(t)), \quad [13]$$

When $B^{US}(t)$ and $\lambda^{US}(t)$ equal zero, NMR(t) decays to baseline by

$$\triangle \text{NMR}(t) = - k2\ \text{NMR}(t). \qquad [\ 14\ ]$$

COMPUTER SIMULATIONS

In the simulations, continuous time was converted to discrete time steps or bins of 10 msec in duration. Each trial consisted of 60 bins. Otherwise specified, the simulations assumed 200 msec CSs, the last 50 msec of which overlaps the US.

Parameters values for variations of associative values were : $\theta = 0.1$, and $\theta' = 0.001$. For antiassociative value were: $\theta = 0.005$ and $\theta' = 0.1$, with exception of the inhibitory conditioning cases for which $\theta = 0.05$. For variations in associability : $\Phi_{US} = 1$, $\Phi_A = 0.16$, $\Phi_B = 0.16$, $\Phi_X = 0.01$, and $c = 0.6$. Initial values of Vs and Ns were zero for all i's. Initial values of associability were always selected $\alpha X = 0.1$ and $\alpha_A = \alpha_B = 0.5$. For computations of B_i^k : $w_i^k = 0.4$ when $i \neq r$; and $w_i^k = 0$ when $i = r$. For computations of the NM CR : L1 = 2. For computation of the trace: k1 = 0.1 , and for the NM response topography : k2 = 0.5.

Simulation results.

Acquisition. Figure 1 shows simulations of a delay conditioning paradigm. As CR acquisition proceeds, CR onset latency decreases, and CR amplitude increases. Maximal response amplitude (CR peak) is located at the time of the US occurrence. Context associability decreases and CS associability increases over trials. Simulation results agree with data on delay conditioning (see Gormezano, et al.,

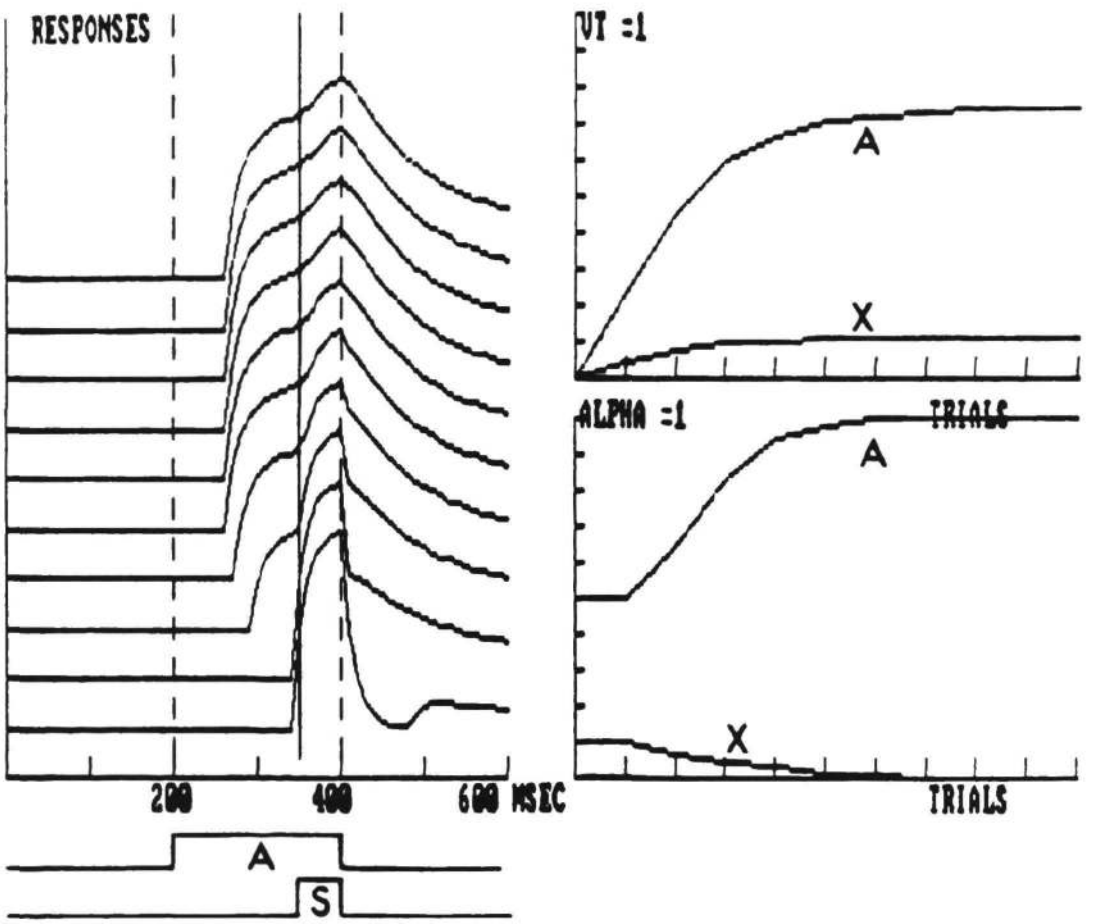


Figure 1. Delay conditioning. A : CS(A). X : Context. Left Panels: NM response topography in 10 reinforced trials. Upper-Right Panels: Net associative values (VT) at the end of each trial, as a function of trials. Lower-Right Panels: Associability (ALPHA) at 350 msec, as a function of trials.

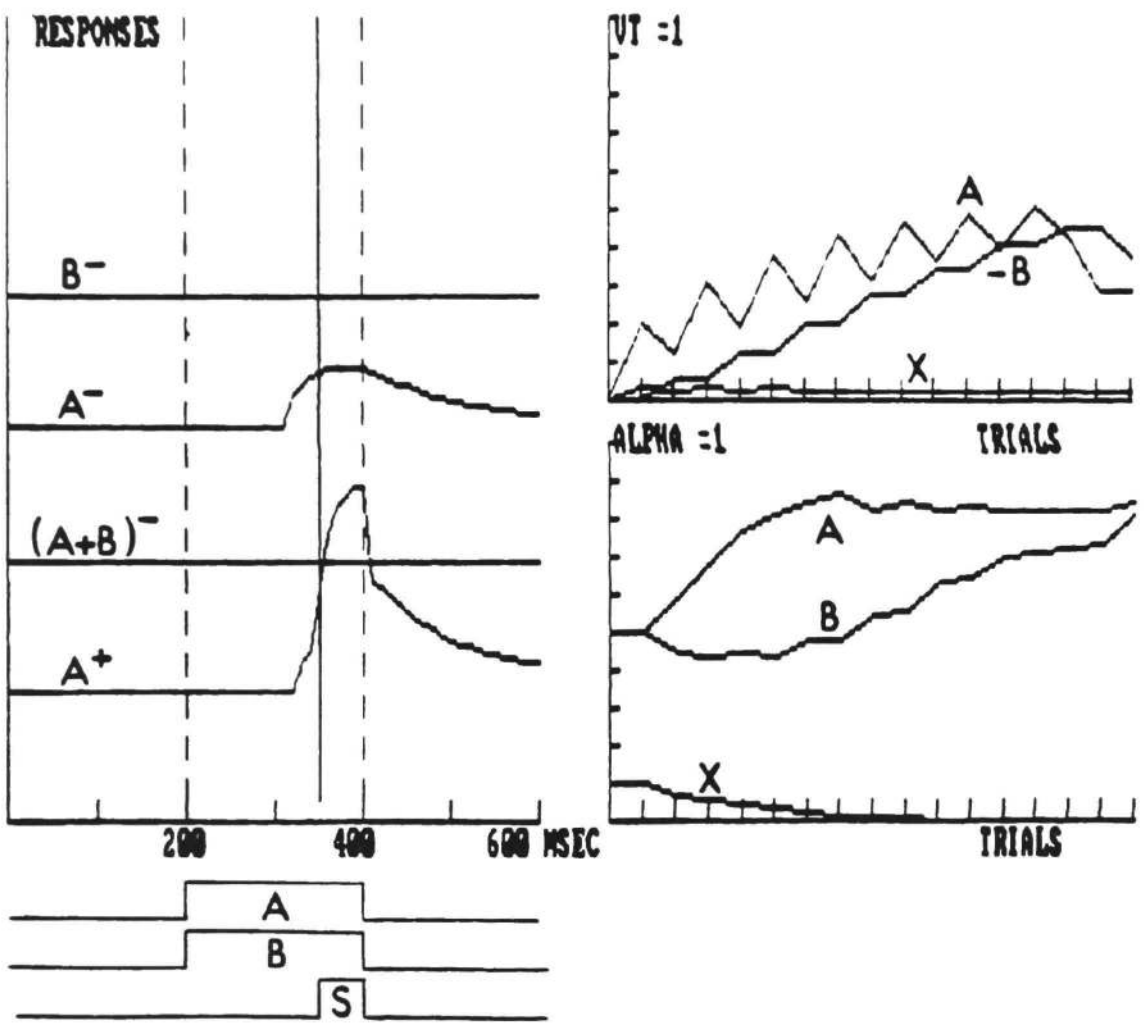


Figure 2. Conditioned Inhibition. A : CS(A). B : CS(B). X : Context. Left Panels: NM response topography in A+, (A+B)-, A-, and B-trials, after 10 alternated A+ and (A+B)- trials. Upper-Right Panels: Net associative values (VT) at the end of each trial, as a function of trials. Lower-Right Panels: Associability (ALPHA) at 350 msec, as a function of trials.

1983)

Conditioned Inhibition. Figure 2 shows simulations of a conditioned inhibition paradigm. During conditioned inhibition two types of trials were alternated: reinforced trials consisted of a single reinforced CS (A), and nonreinforced trials consisted of a compound CS (A and B). Stimulus B was the conditioned inhibitor. After 10 simulated trials, the CR elicited by A and B together was smaller than that elicited when A was presented alone because B has acquired inhibitory associative value. Associabilities of both CSs increased and context associability decreased over trials. Simulation results agree with data on conditioned inhibition reported by Marchant, Miss, and Moore (1972).

Blocking. Figure 3 shows simulations of a blocking paradigm. Experimentals received 5 trials with one CS (blocker) paired with the US followed by 5 trials with the same CS and a second (blocked CS) paired with the US. Controls received 5 two-CS trials in which both CSs were presented together and paired with the US. Controls were subject to mutual overshadowing between the two component CSs. The network showed simulated blocking because CR for the designated blocked CS was smaller than CR for the blocker CS, both after 5 training trials. The results agree with blocking data in the rabbit NM response preparation as reported by Marchant and Moore (1973).

Sensory preconditioning. Figure 4 shows simulations of a sensory preconditioning paradigm. In the first phase, 5 nonreinforced trials with a compound CS(A and B). During the second phase, one of the nonreinforced CSs (A) was reinforced for 5 trials. A test trial assessed the CR to CS(B) never paired with the US. Simulations showed that context associability decreases during preconditioning. In the

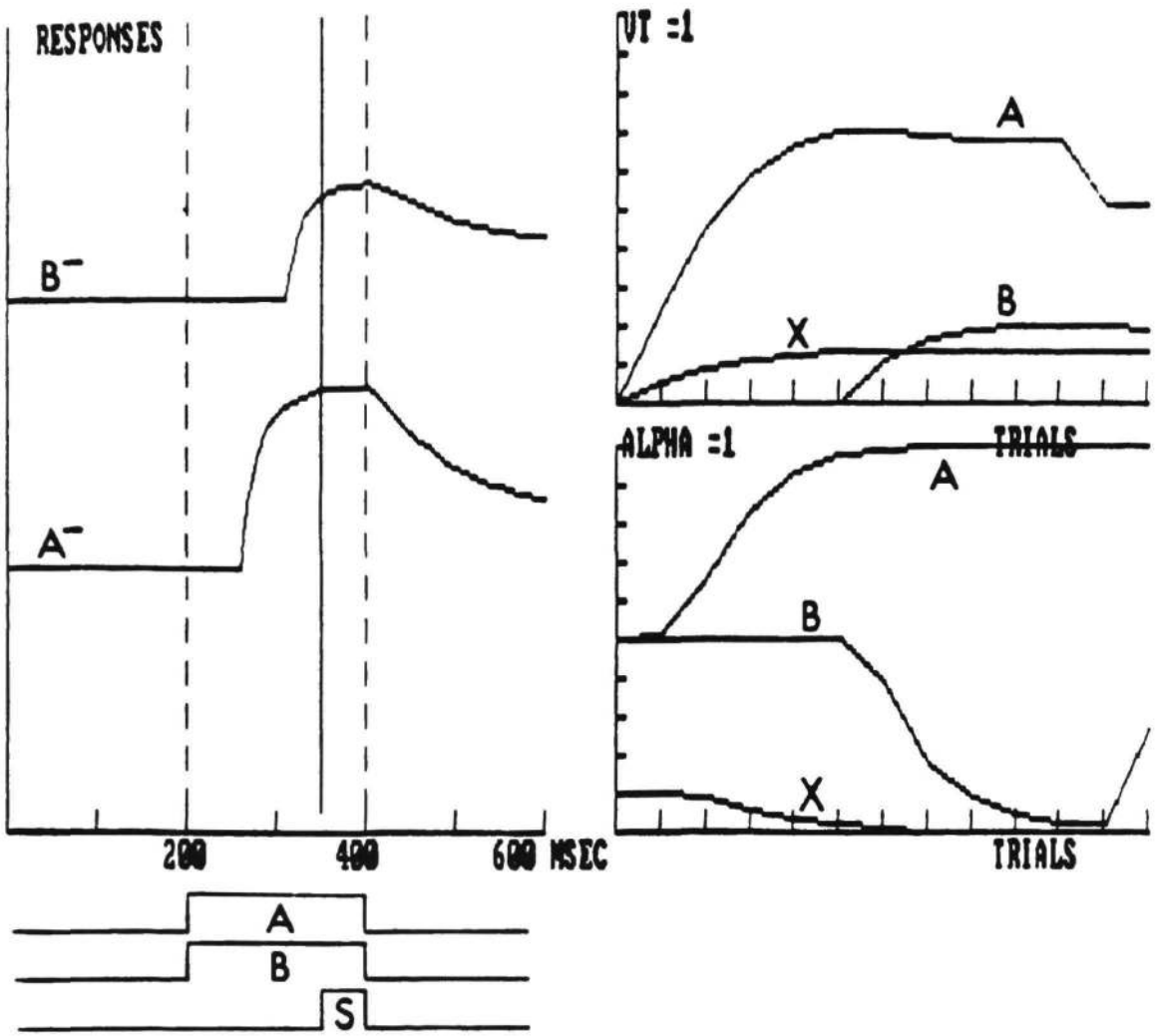


Figure 3. Blocking. A : CS(A). B : CS(B). X : Context. Left Panels: NM response topography in A- and B- test trials, after 5 CS(A) reinforced trials and 5 CS(A) and CS(B) reinforced trials. Upper-Right Panels: Net associative values (VT) at the end of each trial, as a function of trials. Lower-Right Panels: Associability (ALPHA) at 350 msec, as a function of trials.

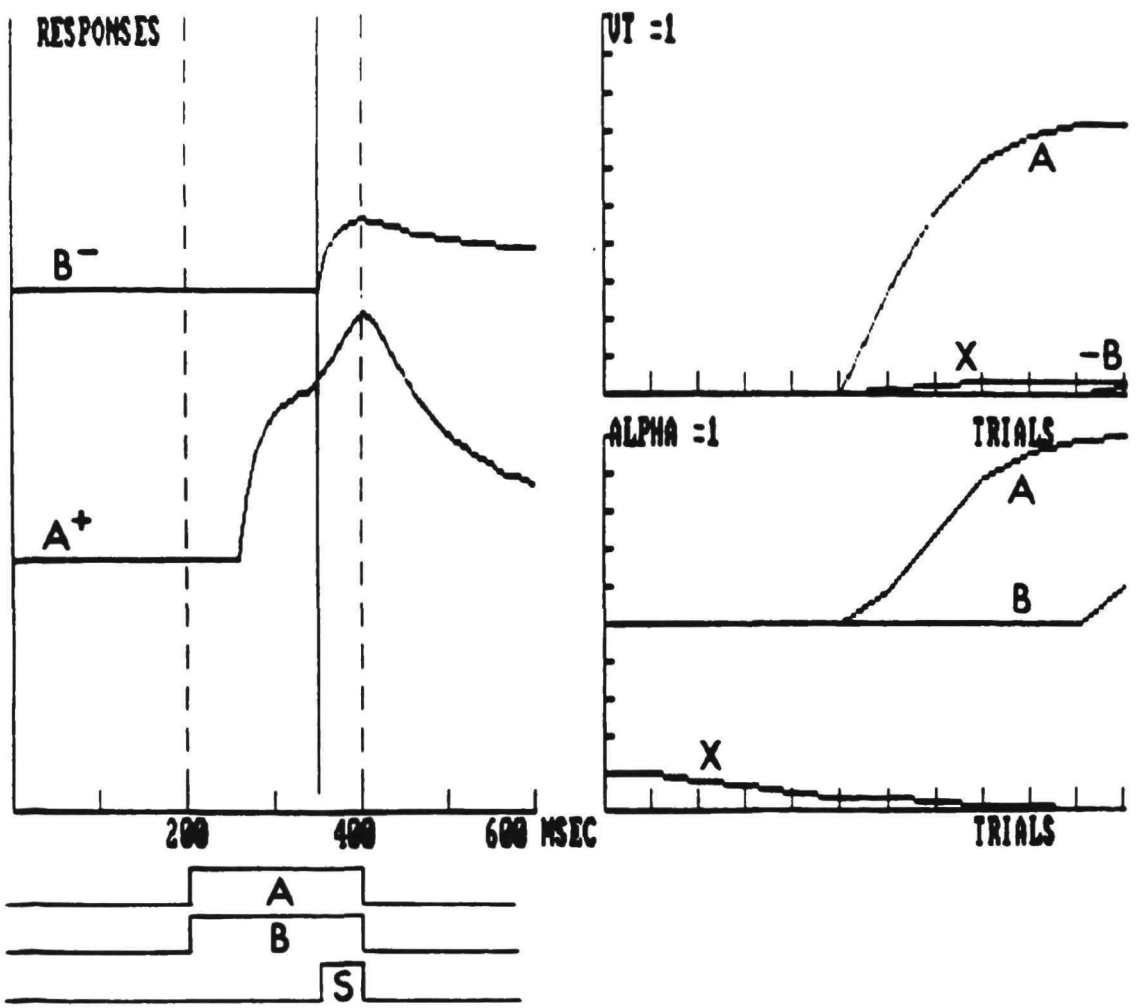


Figure 4. Sensory preconditioning. A : CS(A). B : CS(B). X : Context. Left Panels: NM response topography in A+ and B-trials, after 5 CS(A) and CS(B) nonreinforced trials and 5 CS(A) reinforced trials. Upper-Right Panels: Net associative values (VT) at the end of each trial, as a function of trials. Lower-Right Panels: Associability (ALPHA) at 350 msec, as a function of trials.

nonreinforced test trial CS(B) acquired inhibitory associative value because it was presented in a context with excitatory associative value. CS(B) generated a CR. Simulation results are in agreement with data reported by Port and Patterson (1984).

DISCUSSION

The present paper illustrates how the M-S-S network, a second-order attentional-associative architecture, describes several classical conditioning paradigms in real-time. The network incorporates performance rules that convert net associative values into strength and timing of the rabbit's NM response.

REFERENCES

Gormezano, I., Kehoe, E.J., & Marshall, B.S. (1983). Twenty years of classical conditioning research with the rabbit. **Progress in Psychobiology and Physiological Psychology, 10,** 197- 275.

Hull, C.L. (1943). **Principles of Behavior.** New York: Appleton-Century-Crofts.

Marchant, H.G., Miss, F.W., & Moore, J.W. (1972). Conditioned inhibition of the rabbit's nictitating membrane response. **Journal of Experimental Psychology, 95,** 408-411.

Marchant, H.G., & Moore, J.W. (1973). Blocking of the rabbit's conditioned nictitating membrane response in Kamin's two-stage paradigm. **Journal of Experimental Psychology, 101,** 155-158.

Moore, J.W., & Stickney, K.J. (1980). Formation of attentional-associative networks in real time: Role of the hippocampus and implications for conditioning. **Physiological Psychology,** 8, 207-217,

Moore, J.W., & Stickney, K.J. (1982). Goal tracking in attentional-associative networks: Spatial learning and the hippocampus. **Physiological Psychology, 10,** 202-208.

Moore, J.W., & Stickney, K.J. (1985). Antiassociations: Conditioned inhibition in attentional-associative networks. In R.R. Miller and N.E. Spear (Eds.), **Information processing in animals: Conditioned Inhibition.** Hillsdale: Lawrence Erlbaum.

Port, R.L., & Patterson, M.M. (1984). Fimbrial lesions and sensory preconditioning. **Behavioral Neuroscience, 98,** 584-589.

Schmajuk, N.A., & Moore, J.W. (1986). Real-time attentional models for classical conditioning and the hippocampus. **Physiological Psychology**, 13, 278-290.

Using Mental Schemata: An Experimental Analysis of Computer Skill Acquisition

Marc M. Sebrechts
Charlotte T. Furstenberg

Department of Psychology
Wesleyan University
Middletown, CT 06457

Abstract

Although mental schemata are central to many contemporary learning theories, the precise relationship between such schemata and specific types of learning remains vague. This paper describes an analysis of subjects learning basic computer skills when presented with four different kinds of elaborations that should influence subjects' schemata: (1) a simple description with no model; (2) a redundant elaborated text also with no explicit model; (3) a functional model; (4) a descriptive analogy. Subjects were tested on procedures, general command concepts, and system questions. Models and analogies were shown to improve initial performance on all types of questions. However only system questions showed this advantage after a delay. It is argued that the utility of building a mental model through elaboration depends on the specific tasks that are analyzed.

Although there is consensus concerning the general utility of mental models, there is relatively little agreement concerning the ways in which these models are formed and used in specific tasks. These more detailed relationships are necessary both for a general theory of learning and for designing improved instruction. This paper provides one instance of specific model-task analysis, by investigating the use of models during the acquisition of basic computer (operating system and editor) skills. There are three questions that this research is designed to address: (1) Do explicit attempts to modify the subjects' schemata by providing elaborations influence performance? (2) Are the effects of such manipulations specific to particular kinds of tasks or do they generalize to a range of tasks? (3) What underlying mechanisms can account for these effects?

In order to evaluate the effectiveness of explicit models, we manipulated the kinds of information (models) presented to subjects. Specifically, the study investigated functional models of the underlying

---

This research was supported in part by NSF Grant IST82-17572. We wish to thank Richard L. Marsh for collecting the data.

system operations, analogies describing system performance, and specific examples of commands. The question of the influences of models on different tasks was addressed by a series of tests assessing knowledge of specific procedures, knowledge of the general principles behind the procedures, and knowledge about the general functioning of the system.

Method

Subjects and Materials

Eight students who had some computer experience but no prior experience with UNIX served as subjects.

Instructions. The experimental materials consisted of a set of four booklets, each describing a subset of UNIX commands. Each booklet presented a general description of a particular aspect of the system followed by a description of the procedure for using four UNIX commands.

Eight versions were written for each of the booklets, reflecting four different types of elaboration combined orthogonally with two formats of command examples. The four elaboration conditions consisted of the following: (1) The Simple version stated the relevant concepts, but provided no elaboration of their meaning. (2) The Elaborated version contained an extended statement of the simple concepts, while avoiding explanatory material insofar as possible. (3) The Model version provided a functional model describing certain properties of the system. (4) The Analogy version provided an analogy from a familiar domain.

For each elaboration condition, there were two types of example format. The abstract format included a general description of the procedure together with an abstract statement of the form of the command (e.g. "mv NAME1 NAME2<cr>"). The concrete example format described a specific instance of command use and indicated the way in which that particular command would be written (e.g. "mv george harry<cr>").

Questions. For each set of concepts and commands (each booklet), a set of related questions was developed. These questions were divided into three groups: procedural questions testing knowledge of specific procedures: "true-false" declarative questions, testing general knowledge of command usage; and system questions, testing knowledge of overall system structure or function.

Procedure

Each subject was run individually on three consecutive days. On each of the first two days, subjects were presented with two of the four instruction booklets, each followed by a related set of questions. On the third day they were presented with all of the questions from the first two days in quasi-random order. For each booklet of instructions and each set of questions, the subject was

asked to read aloud the material and to say whatever came to mind while reading or while answering questions.

Results

Elaboration Condition. The effect of elaboration condition for the three types of questions immediately after training (initial performance) is shown in Figure 1. In general, performance was better in the model and analogy conditions than in the simple and elaborated conditions. Procedural and declarative questions showed reliable effects of elaboration condition ($F(3,9) = 9.20$, $p < .01$). Models and analogies were best for procedural questions, whereas the elaborated condition was worst. For the declarative questions, analogies were better than the other three elaborations. The greatest effects of type of elaboration were evidenced on the system questions ($F(3,9) = 4.03$, $p < .05$). Responses in the model and analogy conditions indicate 76% correct, whereas the responses in the simple and elaborated conditions were at 50%.

Subjects were better with the editor questions (procedural 62%; declarative: 76%) than with the other command subsets (procedural: 39%; declarative 63%). This is probably due to subjects having had more previous experience with editors than with other types of commands.

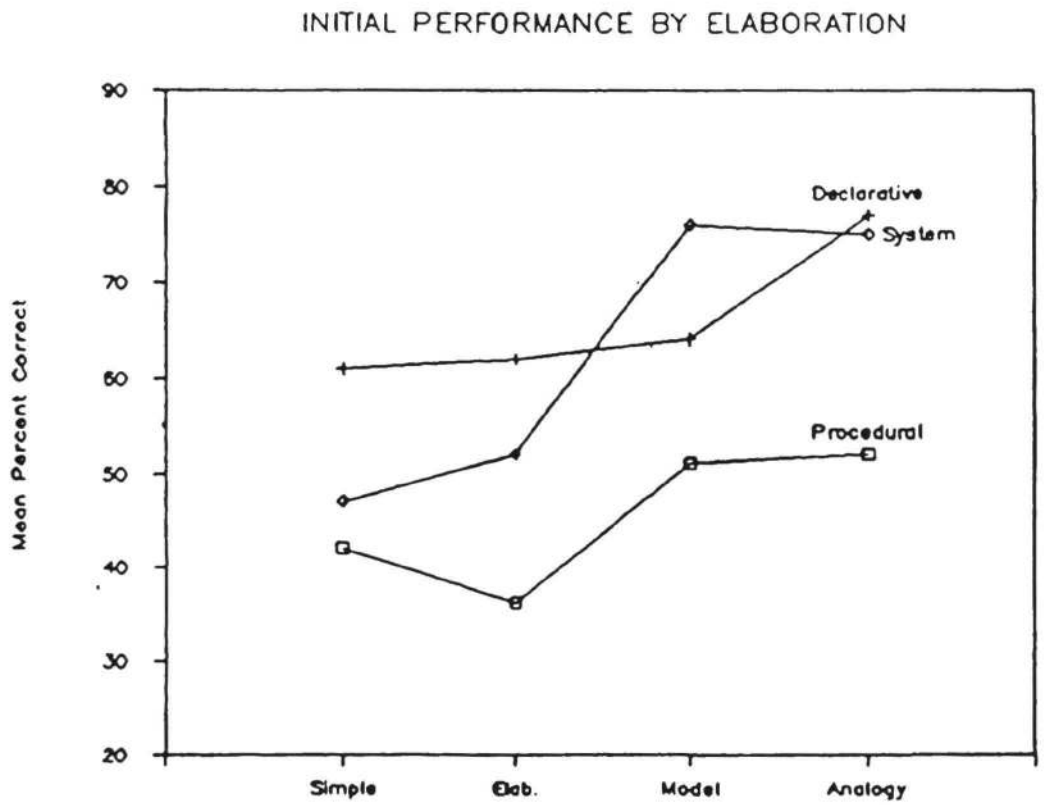


Figure 1

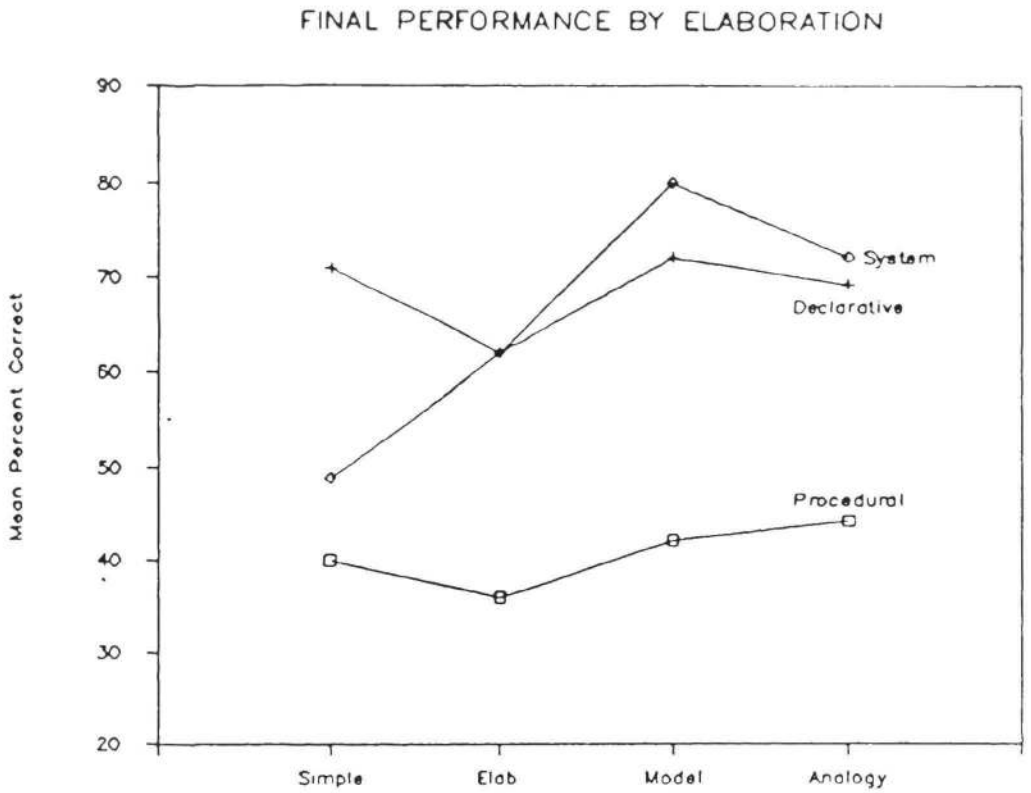


Figure 2

Changes in the effects of elaboration over time were measured by a final test after the completion of all training. The results for this performance are shown in Figure 2. The model and analogy conditions continued to show a performance advantage for system questions which required general knowledge of concepts rather than individual procedures. The rank order effects for procedural questions remained the same, but differences among elaboration conditions were no longer reliable. Presumably the models and analogies provided relatively little help in retaining specific aspects of individual commands.

Examples. As can be seen in Figure 3, the primary effect of examples was in the procedural case. Those using concrete examples got 51% of the procedures correct, whereas those using an abstract format got only 39% correct. Presumably the specific examples provide a concrete instance that is easier to assimilate than the general statements. These specific syntactic components, however, did not seem to have general utility for either the more general command information in the declarative questions, or for the system questions. At the same time, it is important to note that abstract formats did not improve responses to general command (declarative) questions. A single concrete example appears to be as useful in this case as the more abstract statement.

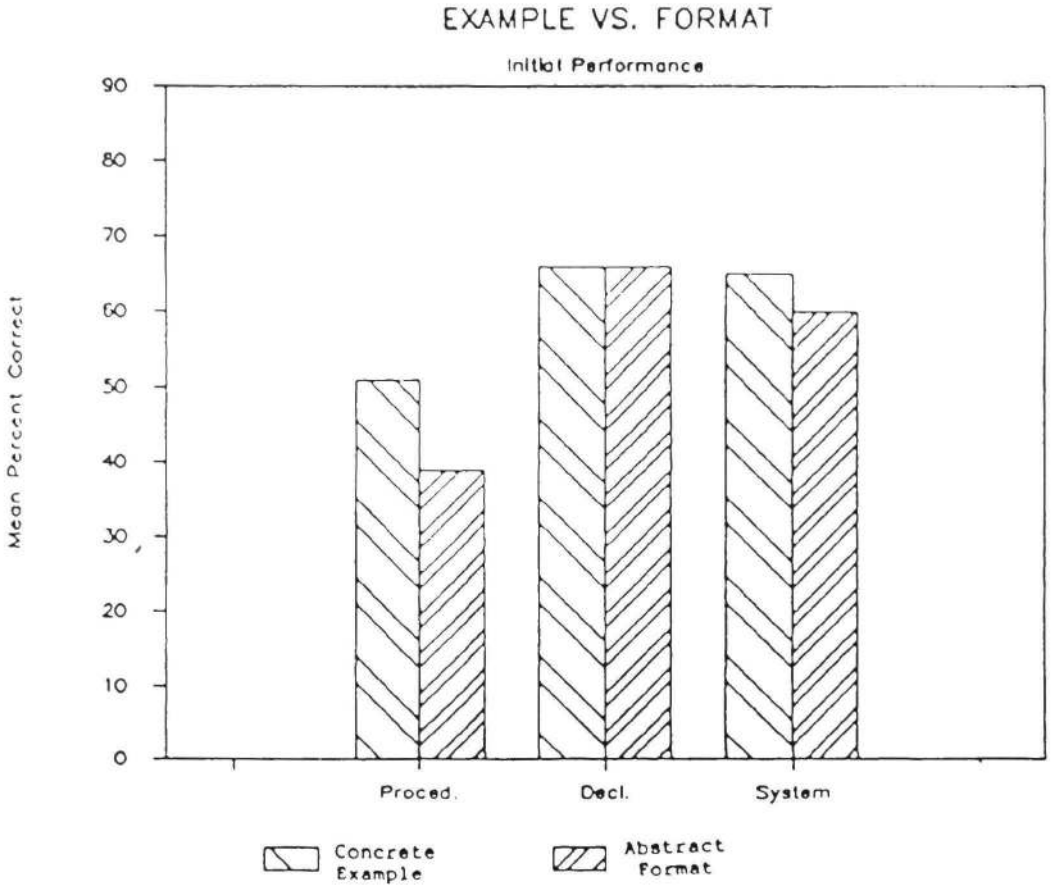


Figure 3

The effect of examples is largely eliminated during final testing. Unlike the model information, examples appeared to be encoded as separate instances. Individual retrieval of these items required in initial performance is fairly straightforward; however, such individual instances apparently do not provide the kind of schema that is necessary to distinguish among a set of commands as required during final performance.

Learning and Forgetting. Changes in schemata over time were assessed by analyzing response changes between initial and final testing. Learning was assessed by determining the percentage of wrong answers on initial performance that were correct on final performance. Forgetting was measured by determining the percentage of questions that were initially correct, but which were wrong on the final performance test. The results of this analysis are shown in Figure 4. In general, there is more forgetting than learning on the procedural questions, whereas there is more learning than forgetting on both the declarative and system questions.

These results are consistent with the notion that subjects are actively integrating information during initial testing as well as during study whenever it is available. The procedural questions require specific sequences of steps, and there is relatively little information about those sequences outside of the specific commands. The declarative and system questions, on the other hand, do not require specific syntactic knowledge. Information from other commands can be informative about the kinds of constraints imposed or about possible generalizations. As a consequence, to the extent that the material from the different booklets is actively integrated, subjects may actually be learning or improving their understanding of new concepts.

This kind of integrative learning, however, can also hurt performance. If subjects are actively assimilating information they may take false statements as true. In fact, we found that our subjects tended to change almost twice as many of their responses on declarative questions from false to true (32%) as from true to false (17%).

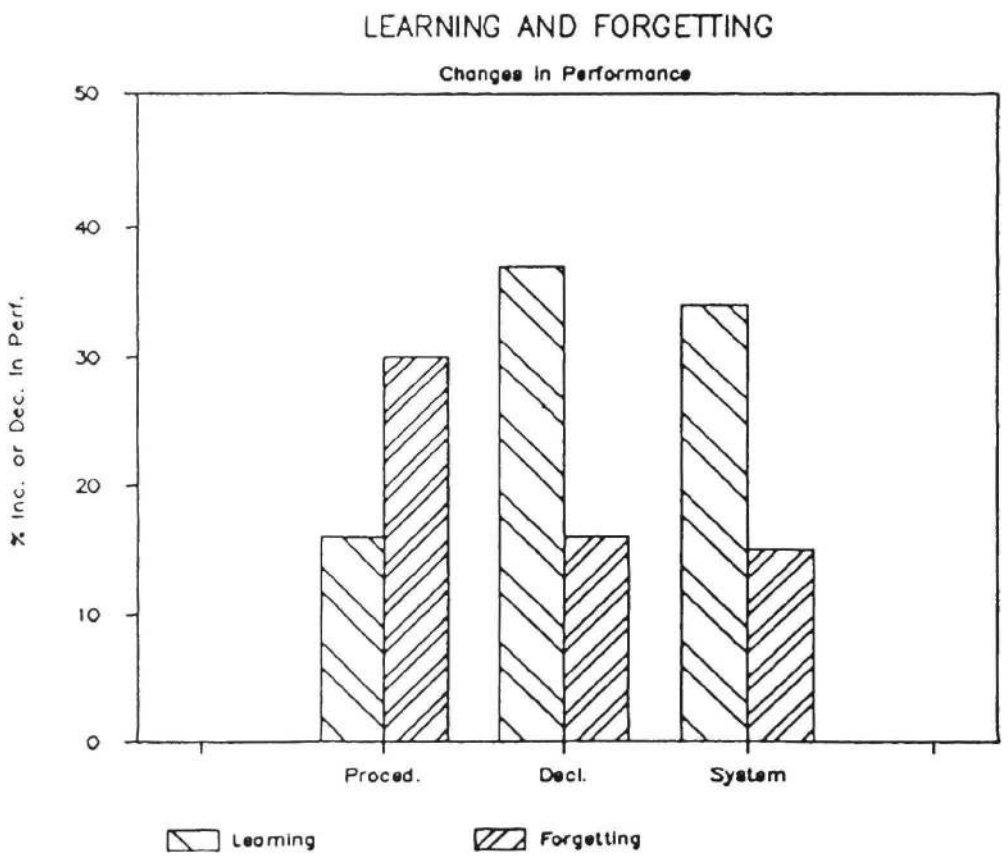


Figure 4

Discussion

The results reported here indicate that certain types of elaboration can in fact improve performance. The most dramatic effects of models and analogies were found on system questions which tapped general system knowledge. These results suggest that a model or analogy can provide a framework to which new information is assimilated. Improved performance was evident during both initial and final tests, suggesting that the elaborations provided a useful schema for retention. These results are consistent with those reported for basic text editing (Bott, 1979) and simple programming tasks (Mayer, 1975).

Models and analogies also showed improvement on procedural and declarative questions, suggesting that elaborations can influence different types of information. An analysis of errors indicated that improved performance was not restricted to particular questions. It is likely that the additional information provided by models and analogies improved performance through multiple access routes activated in memory (Anderson & Reder, 1979; Reder, 1979). These effects, however, are largely eliminated during final performance. This seems to be due to two factors. First, the model information is not specifically linked to individual procedures. Second, the information about command use is being integrated across the various commands. This amounts to a type of self elaboration in which subjects organize individual commands according to their own models, thus minimizing differences among the elaboration conditions.

Although there is some evidence of assimilation, our data suggest that much of the information is stored as individual facts or small clusters of knowledge rather than as general schemata. This type of encoding is perhaps most evident in the use of specific procedures and examples. Concrete examples did improve performance on procedures, even in the absence of general models. By storing a specific instance, subjects could apparently retrieve that individual instance at the time of test with a related example. The importance of this effect of individual instances was noted by Ross (1984) in a study on reminding. He found that during the acquisition of text-editing skills, subjects retrieved procedures by being reminded of superficial resemblances of the examples. Thus subjects would often use specific instances rather than any general model.

This approach seemed useful to our subjects when the number of commands was very restricted as in initial testing. However, the advantage of examples was not present during final testing when subjects were required to remember numerous instances. Presumably the examples are useful only as long as they provide sufficient information to discriminate among different procedures.

Changes in performance suggested that subjects were continuing to integrate information during the course of the study. For experts in any domain, assimilation of new information is often sufficient since they already have highly elaborated models of a domain. In the present study, some of our subjects appeared to perform like experts with respect to editing information, a domain which

was quite familiar. In general, however, our subjects schemata were either missing or distorted. The data suggest that providing explicit models can help in the task of restructuring schemata (Rumelhart & Norman, 1978) or in constructing new ones. That process, however, is complex; although providing specific types of elaborations can improve performance by providing an explicit framework for integration of knowledge, subjects also introduce their own strategies for developing a representation of the domain.

References

Anderson, J.R. & Reder, L.M. (1979). An elaborative processing explanation of depth of processing. In L.S. Cermak and F.I.M. Craik (Eds.), **Levels of processing in human memory**. Hillsdale, NJ: Erlbaum.

Bott, R. (1979). **A study of computer learning: Theory and methodologies**. Center for Human Information Processing Technical Report No. 82. University of California, San Diego.

Mayer, R.E. (1975). Different problem-solving competencies established in learning computer programming with and without meaningful models. **Journal of Educational Psychology**. *67*, 725-734.

Reder, L.M. (1979). The role of elaboration in memory for prose. **Cognitive Psychology**. *11*, 221-234.

Ross, B. (1984). Remindings and their effects in learning a cognitive skill. **Cognitive Psychology**. *16*, 371-416.

Rumelhart, D.E. and Norman, D.A. (1978). Accretion, tuning and restructuring: Three modes of learning. In J.W. Cotton and R. Klatzky (Eds.). **Semantic factors in cognition**. Hillsdale, NJ: Lawrence Erlbaum Associates.

A MODEL OF ATTENTION FOCUSSING DURING PROBLEM SOLVING *

Jude W. Shavlik†
Gerald F. DeJong

Coordinated Science Laboratory
University of Illinois at Urbana-Champaign

ABSTRACT

We propose that three qualitatively different strategies help focus attention during problem solving. The first strategy is to apply operators that will lead to definite progress toward the goal. Attention will be focussed by this strategy as long as some operator in this class is applicable, When clear progress can not be achieved, the problem solver must decide how best to proceed. It then invokes the second strategy to select operators that preserve important characteristics of the current problem. These operators are likely to keep the problem solver from diverging sharply from the goal while possibly enabling the application of operators by the first strategy. When the problem solver can follow neither of the first two strategies, it invokes the third strategy of arbitrarily applying legal operators. We see the second strategy as an essential difference between novice and expert problem solvers. It is easy to recognize definite progress toward a goal and it is easy to recall which operators can be legally applied. Expertise involves knowing which characteristics of a situation should be preserved (or created) when no way to definitely progress toward the goal is known. This three-strategy theory has been implemented and tested in a system that performs mathematical calculations in the course of solving physics problems. We describe a number of mathematical calculation operators used under each strategy.

INTRODUCTION

A novice problem solver attacks a problem in one of two ways. He may immediately notice a way to progress toward the solution. Alternatively, he may flounder around performing legal, but aimless, operations in an attempt to transform the problem into a familiar form. An expert, on the other hand, can perform in a qualitatively different manner. If the solution is not immediately apparent, he can focus his efforts in a much more guided way. Instead of simply thrashing around, he has an appreciation of what kinds of transformations are likely to change the current problem into a soluble problem.

Consider the problem of evaluating the expression

$$mass_1\ velocity_1 + mass_2\ velocity_2$$

when $mass_1$ and $mass_2$ are known but $velocity_1$ and $velocity_2$ are not. This expression cannot be evaluated directly. A valid approach might be to substitute equivalent expressions for the unknowns. A novice might perform the unappealing substitutions of figure 1. While these transformations are valid, they are unlikely to yield a solution. An expert will appreciate this, and be more likely to perform the pleasing problem transformations of figure 1. In this example, there is something about the parallel structure of the problem that makes a parallel substitution more appealing. Yet parallel substitutions are not always aesthetically appealing. For example, consider the substitutions of figure 2. In this example, parallel substitution using Newton's third law (every action has an equal and opposite reaction) misses a useful variable cancellation. If only one instantiation of this formula is used, the two forces can be eliminated from the calculation.

* This research was partially supported by the National Science Foundation under grant NSF IST 85-11542.

† University of Illinois Cognitive Science/Artificial Intelligence Fellow.

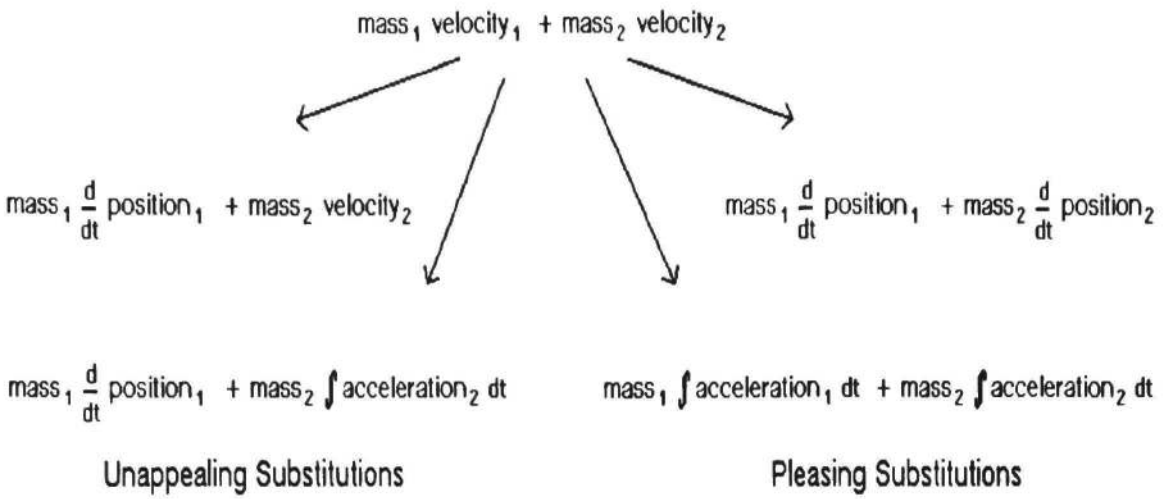


Figure 1. Sample Mathematical Substitutions

We propose three qualitatively different strategies for human problem solving. Attention is selectively focussed according to one of these strategies. *Strategy 1* is hill climbing. Attention is focussed on this strategy as long as some operator moves the problem solver closer to its goal. Occasionally the problem solver will reach a local maximum; there will be no way to move closer to the goal. At these times, the problem solver must diverge from his goal, in the hopes of transforming the current situation into one where hill climbing can occur. We propose there are two qualitatively different ways by which a problem solver re-focusses his attention during this divergent phase. A novice problem solver merely selects an arbitrary legal operator. This can lead to aimless floundering, due to the large number of possible combinations of operator sequences. An expert can often wisely choose how to transform the current situation. Characteristics that are believed to be of general or domain-specific problem-solving importance (e.g., symmetry) are to be maintained or introduced; introduction of troublesome characteristics is to be avoided. Such motivated diversions compose *strategy 2*. However, when an expert exhausts his expertise he also resorts to any legal operator. This unmotivated application of operators is termed *strategy 3*.

AN APPLICATION OF THE MODEL IN MATHEMATICAL CALCULATION

Our three-strategy theory has been implemented and tested in a system called **Physics 101** [1]. This system performs mathematical calculations in the course of solving physics problems. We have identified a number of mathematical calculation operators used by the three strategies.

Our model is a *state-space* model. The problem solver is provided an initial state and a goal description. The role of the problem solver is to successively apply legal transformations to the current state until a state satisfying the goal description is reached. This process in the example domain of mathematical calculation is illustrated is figure 3. Notice that more than one

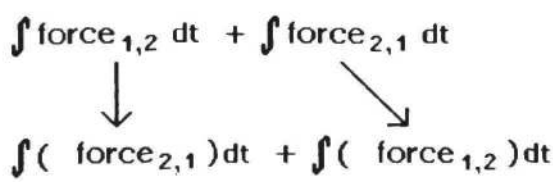


Figure 2. Inappropriate Parallel Substitutions

substitution is allowed during the transformation from one state to the next.

Strategy 1 - Definite Progress.

In our **Physics 101** system, often the goal is to produce an expression that only contains variables whose values are known. (Once in this state, the expression can easily be evaluated.) The hill-climbing measure is the number of variables in the expression whose value is not known (these variables will be called *unknowns* from now on).

Physics 101 contains two basic techniques for reducing the number of unknowns. One, unknowns can be replaced using known formulae if these formulae introduce known-valued terms or lead to the cancellation of unknowns. An example of this technique is shown in figure 4 (discussed below). Two, values of variables can used if doing so eliminates unknowns. If our current expression is $A * B * C$, and A's value is zero, both B and C can be cancelled by replacing A with its numerical value. Similarly, given the expression $A * B * C - B * C$, where A equals one, a numerical replacement can lead to a reduction in unknowns.

A problem-solving operator applied by strategy 1 is illustrated in figure 4. The goal is to evaluate the top expression, but the values of the two inter-object forces ($force_{1,2}$ and $force_{2,1}$) are not known. A "substitute-to-cancel-unknowns" operator can detect that the two inter-object forces cancel due to Newton's third law. This operator is a complex operator. We allow operators to comprise a sequence of problem-solving steps that achieve some goal. (Operators comprising other operators have variously been called macro-operators [2], frames [3], scripts [4], and schemata [5].) These "super" operators possess the desirable property that once one is selected, several problem-solving steps can be carried out without the need for intervening search. The "substitute-to-cancel-unknowns" operator first applies Newton's third law to replace one of the inter-object forces. It then brings these potential cancelling terms into a position where cancellation can take place. This requires that the two integrals be combined. Finally, the two troublesome variables can be eliminated.

Bundy's meta-level solution methods [6] follow strategy 1. He considers solving complicated equations containing a single unknown variable (there may be multiple occurrences of the unknown, however). His *attraction*, *collection*, and *isolation* methods always bring one closer to the goal of having the only occurrence of the unknown isolated on the left-hand side of an equation.

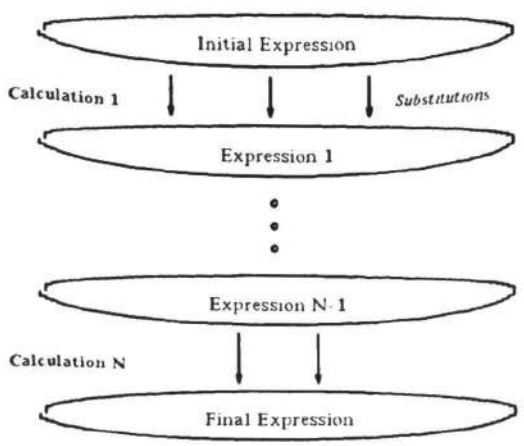


Figure 3. The Structure of a Calculation Sequence

Consider the Expression

$$\int \text{force}_{1,2}\, dt + \int \text{force}_{2,1}\, dt$$

Choose a Variable-Cancelling Substitution

$$\int (\ \text{force}_{2,1}\ )dt + \int \text{force}_{2,1}\, dt$$

Bring Cancellers Together

$$\int (\ \text{force}_{2,1} + \text{force}_{2,1}\ )dt$$

Cancel Variables

$$\int 0 \text{kg m} / \text{s}^2\, dt$$

Figure 4. Example of an Operator Applied by Strategy 1

Strategy 2 - Motivated Diversions.

There are several techniques used in **Physics 101** that follow the second strategy. The major ones are presented below, in the order they are used by the system.

(1) *Elimination of variables whose values are known.* Even when is not possible to reduce the number of unknowns in an expression, it is a good idea to cancel terms, even those whose values are known. Eliminating these terms may allow productive cancellations that had been prevented by the presence of known-valued terms. For example, suppose we have the following formulae, and the values of A and C are known.

$$A = B / C \quad \text{and} \quad B = -D$$

If our current expression is $A * C + D$, we cannot reduce the number of unknowns. However if we use the first of the above formulae, we can cancel some terms, while momentarily increasing the number of unknowns. Fortunately, in the next step both of the unknowns can be cancelled.

(2) *Elimination of integrals and derivatives.* Performing calculus is harder that performing algebra. When it is not possible to eliminate terms, it is often a good idea to eliminate calculus structure. An example where **Physics 101** removes calculus from an expression is shown in figure 5. Here the program detects that it can eliminate the derivative because it knows the derivatives of all the terms being differentiated. Applying this operator adds four steps to the calculation sequence. After this operator is applied, direct progress toward the goal state can be made. For instance, object 1's mass times its acceleration equals the net force on it (by Newton's second law). Hence, two variables can be replaced by one. (Continued calculation leads to this expression becoming zero, a consequence of the principle of conservation of energy.)

(3) *Preservation of expression type.* Assume we have the following two equations.

$$\text{(i)}\ A = (D * E) \qquad \text{(ii)}\ A = (F + G)$$

If our current expression is $A * B * C$, equation (i) would be preferred as this would maintain the property that the expression is a product of terms. Conversely, given $A + B + C$, the second equation is preferred, because now the expression continues to be a sum of terms. There is a strong reason for preserving expression type, one involving more than aesthetics. In the first example we can produce

$$\frac{d}{dt}\left((1/2)\,mass_1\,velocity_1^{\,2} + mass_1\,g\,position_1\right)$$

SeparateCalculus $$= \frac{d}{dt}\left((1/2)\,mass_1\,velocity_1^{\,2}\right) + \frac{d}{dt}\left(mass_1\,g\,position_1\right)$$

ConstantsOutOfCalculus $$= (1/2)\,mass_1\,\frac{d}{dt}\,velocity_1^{\,2} + mass_1\,g\,\frac{d}{dt}\,position_1$$

SolveCalculus $$= (1/2) * 2\,mass_1\,velocity_1\,\frac{d}{dt}\,velocity_1 + mass_1\,g\,\frac{d}{dt}\,position_1$$

SubstCalculus $$= (1/2) * 2\,mass_1\,velocity_1\,acceleration_1 + mass_1\,g\,velocity_1$$

Figure 5. An Application of the Substitute-Calculus Operator

$$D * E * B * C \quad \text{or} \quad (F + G) * B * C.$$

In the result on the left, all the terms are equally accessible. Future substitutions involving B, for example, can cancel D or E. (Recall that cancellation of variables is the mechanism that leads to the goal state.) The right-hand result requires that a replacement for B cancel F and G together.

(4) *Preservation of structural symmetry.* When similar additive or multiplicative structure is present, the same general rule should be used repeatedly whenever possible. For example, given the following expression, substitutions involving all three of the A's or all three of the B's would be favored.

$$A_1 B_1 + A_2 B_2 + A_3 B_3$$

This accounts for the unpleasing quality of the first transformations of figure 1. It would be better to replace both of the velocities either by the derivatives of position or by the integrals of acceleration. Mixing the two does not seem right.

The mathematical problem-solving methods learned by Silver's **LP** program [7] follow strategy 2. **LP** acquires information that constrains the choice of applicable operators. The learned operators are not guaranteed to bring the problem solver closer to a final solution.

Strategy 3 - Floundering Around.

In **Physics 101** strategy 3 looks for the first legal substitution and applies it. Only one substitution is made, in order to minimize this undirected perturbation of the calculation.

The **LEX** system of Mitchell [8] acquires heuristics that estimate when it is wise to apply an integration operator. It learns how operators previously used only under strategy 3 can be applied by strategy 2. The later **LEX2** system [9] learns under what conditions a specific integration operator will lead to a solution. This can be viewed as learning how to apply, under strategy 1, an operator previously used only under strategy 3.

CONCLUSION

We propose that three qualitatively different strategies help focus attention during problem solving. The first strategy is to apply operators that will lead to definite progress toward the goal. Attention will be focussed by this strategy as long as some operator in this class is applicable. When clear progress can not be achieved, the problem solver must decide how best to proceed. It then invokes the second strategy to select operators that preserve important characteristics of the current problem. These operators are likely to keep the problem solver from diverging sharply from the goal while possibly enabling the application of operators by the first strategy. When the problem solver can follow neither of the first two strategies, it arbitrarily applies operators. This three-strategy theory has been implemented and tested in a system that performs mathematical

calculations in the course of solving physics problems. We have identified a number of mathematical calculation operators used by the three strategies.

When several operators are seen as being equally viable, a problem solver must choose which to apply. Under strategy 3 the choice is made arbitrarily. In the other two cases, the choice is made by using the second strategy. For example, if there are a number of ways to cancel two unknowns, a way that preserves the symmetry of the situation is preferred.

These three strategies can be viewed in terms of simulated annealing problem-solving models [10, 11]. Progress using strategy 1 involves the movement toward a solution space minima. Strategy 2 is analogous to slightly increasing the system "temperature" when stuck at a local minima that is not a goal state. Here it is hoped that the problem solver does not drift too far in the problem space. Strategy 3 potentially involves much greater increases in system temperature, and, hence, much greater jumps in the problem space.

We see the second strategy as an essential difference between novice and expert problem solvers [12-14]. It is easy to recognize definite progress toward a goal and it is easy to recall which operators can be legally applied. Expertise involves knowing which characteristics of a situation should be preserved (or created) when no way to definitely progress toward the goal is known.

REFERENCES

1. J. W. Shavlik and G. F. DeJong, "Building a Computer Model of Learning Classical Mechanics," *Proceedings of the Seventh Annual Conference of the Cognitive Science Society*, Irvine, CA, August 1985, pp. 351-355.
2. R. E. Fikes, P. E. Hart and N. J. Nilsson, "Learning and Executing Generalized Robot Plans," *Artificial Intelligence 3*, (1972), pp. 251-288.
3. M. L. Minsky, "A Framework for Representing Knowledge," in *The Psychology of Computer Vision*, P. H. Winston (ed.), McGraw-Hill, New York, NY, 1975, pp. 211-277.
4. R. C. Schank and R. P. Abelson, *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures*, Lawrence Erlbaum and Associates, Hillsdale, NJ, 1977.
5. W. Chafe, "Some Thoughts on Schemata," *Theoretical Issues in Natural Language Processing 1*, Cambridge, MA, 1975, pp. 89-91.
6. A. Bundy and B. Welham, "Using Meta-level Inference for Selective Application of Multiple Rewrite Rules in Algebraic Manipulation," *Artificial Intelligence 16*, 2 (1981), pp. 189-212.
7. B. Silver, "Using Meta-level Inference to Constrain Search and to Learn Strategies in Equation Solving," Ph.D. Thesis, Department of Artificial Intelligence, University of Edinburgh, 1984.
8. T. M. Mitchell, P. E. Utgoff and R. Banerji, "Learning by Experimentation: Acquiring and Refining Problem-solving Heuristics," in *Machine Learning: An Artificial Intelligence Approach*, R. S. Michalski, J. G. Carbonell, T. M. Mitchell (ed.), Tioga Publishing Company, Palo Alto, CA , 1983, pp. 163-190.
9. T. M. Mitchell, "Learning and Problem Solving," *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, Karlsruhe, West Germany, August 1983, pp. 1139-1151.
10. S. Kirkpatrick, C. D. Gelatt and M. P. Vecchi, "Optimization by Simulated Annealing," *Science 220*, (1983), pp. 671-680.
11. G. F. Hinton, T. J. Sejnowski and D. A. Ackley, "Boltzmann Machines: Constraint Satisfaction Networks that Learn," Technical Report CMU-CS-84-119, Carnegie-Mellon University, Pittsburgh, PA, May 1984.
12. J. H. Larkin, J. McDermott, D. P. Simon and H. A. Simon, "Models of Competence in Solving Physics Problems," *Cognitive Science 4*, 4 (1980), pp. 317-345.
13. M. T. Chi, P. J. Feltovich and R. Glaser, "Categorization and Representation of Physics Problems by Experts and Novices," *Cognitive Science 5*, 2 (1981), pp. 121-152.
14. M. Chi, R. Glaser and E. Rees, "Expertise in Problem Solving," in *Advances in the Psychology of Human Intelligence, Volume 1*, R. Sternberg (ed.), Lawrence Erlbaum and Associates, Hillsdale, NJ, 1982, pp. 7-75.

TWO PROBLEMS WITH BACKPROPAGATION AND OTHER STEEPEST-DESCENT LEARNING PROCEDURES FOR NETWORKS

Richard S. Sutton *
GTE Laboratories Incorporated
Waltham, MA 02254

ABSTRACT

This article contributes to the theory of network learning procedures by identifying and analyzing two problems with the backpropagation procedure of Rumelhart, Hinton, and Williams (1985) that may slow its learning. Both problems are due to backpropagation's being a gradient- or steepest-descent method in the weight space of the network. The first problem is that steepest descent is a particularly poor descent procedure for surfaces containing *ravines*—places which curve more sharply in some directions than others—and such ravines are common and pronounced in performance surfaces arising from networks. The second problem is that steepest descent results in a high level of interference between learning with different patterns, because those units that have so far been found most useful are also those most likely to be changed to handle new patterns. The same problems probably also arise with the Boltzmann machine learning procedure (Ackley, Hinton and Sejnowski, 1985) and with reinforcement learning procedures (Barto and Anderson, 1985), as these are also steepest-descent procedures. Finally, some directions in which to look for improvements to backpropagation based on alternative descent procedures are briefly considered.

Recent years have seen the development of the first effective learning procedures for connectionist networks that have interior or "hidden" units not directly associated with input or output: the Boltzmann machine learning procedure (Ackley, Hinton and Sejnowski, 1985), the backpropagation learning procedure (Rumelhart, Hinton and Williams, 1985), and the A_{R-P} reinforcement learning procedure (Barto and Anderson, 1985; Williams, 1986). The theory behind these new learning procedures is that of gradient or steepest descent in the space of "weights"—the modifiable memory parameters weighting the efficacy of each connection of the network. At each time step, a step is taken in weight space in the direction in which performance improves most rapidly. Letting $J(w)$ denote the performance measure to be minimized, where w denotes the vector of weights, the steepest-descent strategy can be written

$$\Delta w = -\rho \nabla J(w), \tag{1}$$

where Δw is the change in the weight vector, ρ is a positive learning-rate parameter, and

* I have received help and ideas contributing to this article from a large number of people. I wish to particularly acknowledge Steve Epstein, Andy Barto, John Aspinall, Martha Steenstrup, Ron Williams, Glenn Iba, and Oliver Selfridge.

the gradient $\nabla J(w)$ is the vector of first partial derivatives

$$\nabla J(w) = \left(\frac{\partial J(w)}{\partial w_1}, \cdots, \frac{\partial J(w)}{\partial w_N} \right),$$

where N is the number of weights. Most single-unit and single-layer learning procedures are also steepest-descent procedures, including the perceptron (Rosenblatt, 1962) and the Widrow-Hoff rule (Widrow and Hoff, 1960).

Steepest-descent procedures make their largest changes to those weights where the first partial derivatives are greatest.* This may seem a good strategy, but there are at least two reasons for doing exactly the opposite. First, a small derivative may indicate a shallow, wide, gently-curving part of the surface, where large steps need to be made, whereas a large derivative may indicate a very steep and sharply curving part, where the step size must be reduced to prevent instability. Second, the derivative can only be large for weights and units to the extent that they affect performance. It follows that those with large derivatives will be those that already play a role in the behavior of the network, for example, those that have already formed features of use to the rest of the network. When the network has need to adapt and create new features for a new situation, it should do so while minimizing interference with the existing useful features, meaning that these large-derivative weights should be changed least.

The rest of this paper elaborates on these two problems as they arise in Rumelhart et al.'s backpropagation learning procedure. We concentrate on backpropagation because most results obtained so far suggest that it is significantly more efficient than the other network learning procedures (Anderson, in prep.; Hinton, personal communication), and its capabilities have been impressively demonstrated (e.g., Sejnowski and Rosenburg, 1986). The backpropagation procedure also illustrates the problems we wish to point out particularly clearly. As steepest-descent methods, the other network learning procedures are also subject to the same problems to various degrees.

Rumelhart et al.'s backpropagation procedure is the application of steepest descent to acyclic networks receiving signed multi-dimensional errors as their teaching signals (see Figure 1). In acyclic or feedforward networks, information flows in one direction only: if there is a connection from unit A to unit B, then there can be no connection or series of connections from B back to A. At each time step, a set of input units take on the values of an input pattern, activity is propagated through interior or "hidden" units to a set of output units, and then the net is told what each output unit's activity should have been. The acyclic interconnection means that the propagation of activity can occur in one sweep, updating the activity of each unit only after the activity of all its inputs have already been updated.

---

* Here, and throughout, by greatest we mean greatest in the unsigned sense, that is, greatest in absolute value.

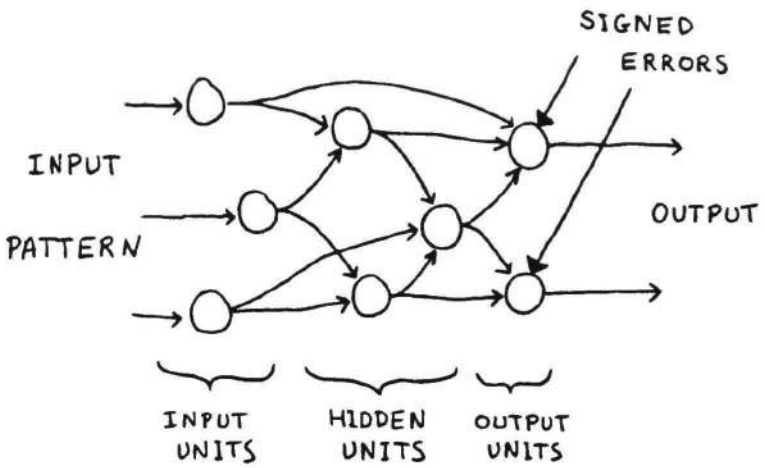


Figure 1. An acyclic network with hidden units and signed-error teaching signals.

In backpropagation, a squared-error performance measure, summed over input patterns, is typically used for $J(w)$. In order for its gradient with respect to the weights to exist, each unit's output activity must be a continuously-differentiable function of its input activity. Typically, each input to a unit is modulated by a separate weight, and the weighted sum of all input is passed through an S-shaped function from $\Re$ to $[0, 1]$. Finally, the name "backpropagation" refers to the way information is propagated in a single sweep in the backward direction, the reverse that of the propagation of activity, to exactly compute the gradient of performance on a step with respect to each weight in the network. This is then averaged or summed over steps to approximate $\nabla J(w)$. We will not need to consider the details of the gradient computation.

STEEPEST DESCENT AND RAVINES

Consider the surface whose contour map is shown in Figure 2a. In region A the surface slopes gently, whereas in region B it is steep. To find one's way from A through B to the optimum in the minimum number of steps, one would clearly want to make larger steps in A than in B; a flat, shallow surface suggests the optimum is far away, and thus that large steps be taken. In this way it is inherently a part of the idea of descent that larger steps should be taken where the gradient is *smallest*. Equation 1, however, results in the opposite, in a step size proportional to the size of the gradient. To acheive the desired step size, the learning-rate parameter ρ must be made much smaller in large-gradient regions such as B than it is in small-gradient regions such as A.

The problem is more serious when the gentle and steep slopes occur simultaneously along different dimensions, as in the surface shown in Figure 2b. Such places, in which the

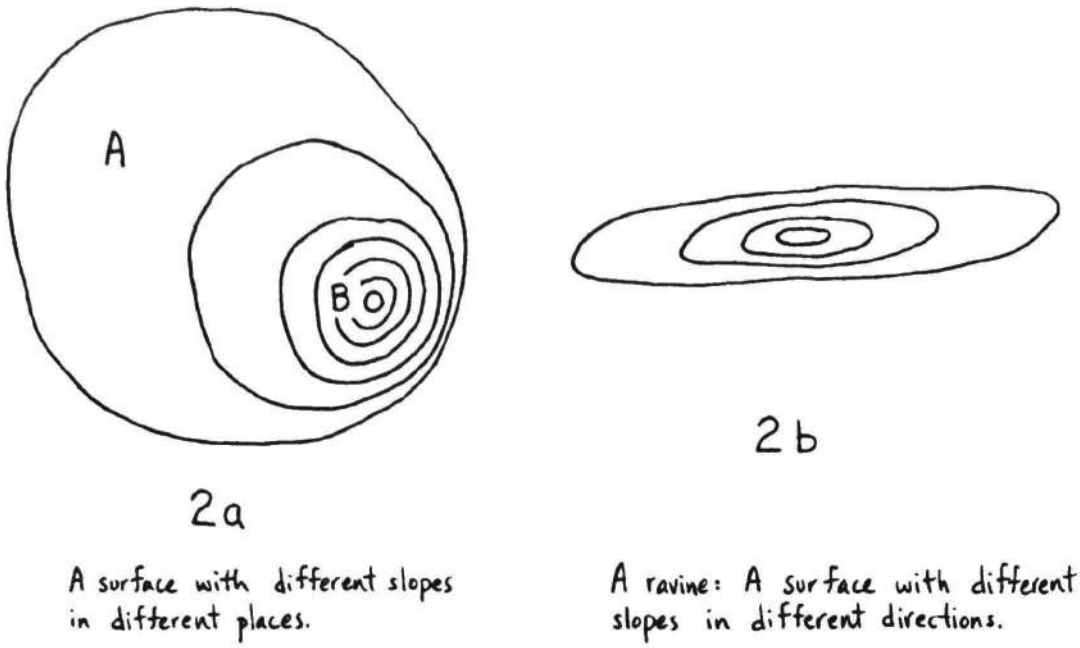


Figure 2.

surface curves much more steeply in one direction than in another, are called *ravines*. As before, we would like to take large steps where the surface is gently sloping—here, along the ravine—and small steps where the surface is steep—across the ravine. Also as before, the magnitude of the gradient is just the opposite of what is desired; it is large across and small along the ravine. Here, however, we cannot solve the problem by varying the learning rate ρ over time. In effect, we need the learning rate to be different in different directions. Since this would alter the direction of the step, it is precisely what is ruled out by steepest-descent procedures, which by definition step directly in the direction of the gradient (i.e., perpendicular to the contour lines; see Figure 2b). If the learning rate is the same in all directions, then it will have to be small enough to prevent instability in any direction, and this means that it will have to be much smaller than optimal in almost all directions, and learning will be very slow and inefficient.

Figure 3 illustrates how ravines arise naturally from the structure of networks. The output unit O receives input from three hidden units A, B, and C, across connections with weights of $w_A = 0.1$, $w_B = 1.0$, and $w_C = 10.0$. The output unit O can effect performance directly, the other three units only through affecting O. Other things being equal, then, changes in C's (input) weights will have ten times the effect on performance as changes in B's weights, and 100 times the effect of changing A's weights. The corresponding ravines will curve approximately 10 and 100 times faster along the dimensions of C's weights than they will along the dimensions of B's and A's weights. Thus, simple variations in the magnitude of weights produce many deep ravines, even more so in deeper and more layered networks. As the signals produced by some units are found to be useful

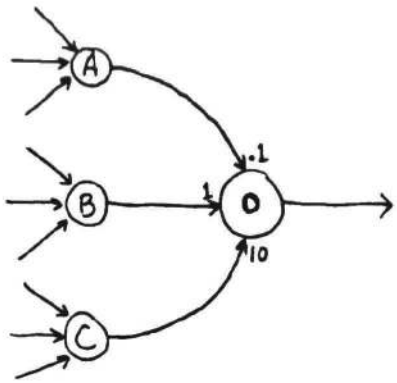


Figure 3. A network fragment. Numbers indicate weight values.

while those produced by others are found not to be, such weight variations undoubtedly will occur.

A possible saving grace is that ravines created in this way always have their principal axes parallel to the principal dimensions of the space (those corresponding to the weights). In two dimensions, for example, such ravines are either horizontal or vertical. Diagonally-oriented ravines occur only in cases of strong interaction between the changes made at different weights. For example, a 45° ravine would mean performance remained good as long as the two weights went up and down together, but worsened quickly if they stepped in different directions. Such interactions certainly can occur in the problems we would like networks to solve, but are not naturally created by network structures the way the weight-axis-aligned ravines are. This observation is important in looking for solutions to the ravine problem; as long as the ravines are aligned with the weight axes, it may be possible to alter or eliminate their effect simply by having different learning rates for each weight, so that larger steps are made along some dimensions, perhaps those of small derivative, than along others, perhaps those of large derivative. Some such possibilities are briefly discussed in a later section.

The ravine problem is appreciated by those who have been designing network learning procedures (e.g., see Derthick, 1984). For example, it is thought that the reason the "momentum" technique* (Rumelhart et al., 1985) improves performance is that it ameliorates the ravine problem. What has not been appreciated, however, is that *steepest* descent

---

* In this modification to the backpropagation algorithm, the weights are changed partly according to the current gradient and partly according to recent past gradients, giving weight motions "momentum". Strictly speaking, this is a departure from steepest descent, but the problems identified here should still be present. The momentum technique increases the rate of learning, but it is still thought to be much too slow (Hinton, personal communication).

is only one of many descent procedures, and one which is known to be inefficient in the presence of ravines (e.g., Tsypkin, 1971; Duda and Hart, 1973; Gill et al., 1981). The prominence of deep ravines in surfaces generated by networks suggests looking beyond steepest-descent procedures.

CROSS-PATTERN INTERFERENCE

Our second problem with steepest-descent network learning procedures has to do with how they handle interference among the various patterns presented to the network. If the network develops a nice set of features for classifying one set of patterns correctly, and then we ask it in addition to classify new patterns, we would like it to do so with minimal interference with the features crafted to classify the first set. If new features are developed to classify the new patterns, they should preferentially be formed from as-yet-unused units rather than by making those already in use serve double-duty.

Unfortunately, the steepest-descent procedure again produces the opposite of the desired behavior. In figure 3, output unit O has learned to listen most strongly to unit C; apparently C has formed a feature of use in solving the problem. Now suppose new patterns are presented, and new features are needed. As discussed previously, C's weights will have much larger derivatives than B's and A's, and so under steepest descent they will change much more dramatically. Alternatively, once C was found to no longer be useful, its incoming weights could have been left unchanged, while its outgoing weight onto O was reduced. Then, if ever the feature provided by C was again needed, its effect could quickly be resurrected rather than its function painstakingly recreated. But steepest descent does not do this. Steepest-descent procedures preferentially change existing, already-useful features rather than make new ones from unused units.

The desired logic here is that of generate and test: Responding to current gradients is the generation process; it is supposed to create any needed new features. The test of the feature provided by a unit is whether it plays a useful role in the network, which will be correlated with its weights having large derivatives. In generate and test we make changes, generate a variety of alternatives, until we find something that passes the test, which we then keep and insulate in some way from further changes. In steepest descent, on the other hand, we make greater and greater changes to a unit's weights the more it is found to be useful and given control over network output. Units with no effect and zero derivatives could experiment arbitrarily without degrading performance. Under steepest descent, however, they will not participate at all in the attempt to find good new features for new situations.

ALTERNATIVE DESCENT PROCEDURES

This article does not propose any specific alternatives or improvements to steepest descent and backpropagation. Here, however, we mention several possible directions in

which to look, and report our experiences with them so far.

The alternative to steepest descent is to still descend, but not directly in the direction of the gradient. A convenient way to think of this is as a distortion of the surface: By stretching the surface perpendicular to ravines, the eliptical contours of a ravine can be converted into circular ones, upon which steepest descent is very effective. Such a distortion of the space is equivalent to distorting individual steps analogously, lengthening them along ravines, shortening them across.

In general, such a distortion involves multiplying the gradient times an $N \times N$ matrix, where N is the number of weights. We will consider this impractical in that it calls for every weight to communicate with every other weight. Such communication is unnecessary if we assume, as discussed earlier, that all ravines are oriented parallel to the weight axes. In this case inter-weight communication is unnecessary; instead of a full matrix multiplication, each weight need only have an individual step-size or learning-rate scale factor. Below we briefly consider three different strategies for setting individual scale factors for each weight.

Squared-error performance measures often result in quadratic or approximately-quadratic surfaces. For such surfaces the direction and distance of the optimum can be accurately estimated from the local first and second partial derivatives. The classic descent procedure that does this is *Newton's method.** It is a matrix method, using the inverse of the *Hessian* matrix D of second partial derivatives:

$$\Delta w = -\rho D^{-1} \nabla J.$$

This method normalizes the first derivative according to how fast it itself is changing; if that rate of change is constant, and $\rho = 1$, then the method brings the weight vector exactly to where $\nabla J = 0$, i.e., to the optimum, in one step. One way to approximate this using only a single scale factor per weight is simply to assume all non-diagonal terms of D are 0, yielding

$$\Delta w_i = -\rho \frac{\partial J}{\partial w_i} \Big/ \frac{\partial^2 J}{\partial w_i^2}.$$

We note in passing that the second partial derivative $\frac{\partial^2 J}{\partial w_i^2}$ can be computed by a backpropagation process entirely analogous to that proposed by Rumelhart et al. for computing for the first derivative.

Newton's method is an entirely analytic method—based on an exactly computed second derivative matrix. Another possibility is to measure empirically the extent to which each weight undergoes changes, and adjust each weight's scale factor so that all weights change by the same amount. This would prevent units with small outgoing connections from

* Newton's method was originally devised to find the zeros of a function. As a descent procedure it is used to find the zeros of the derivative of a function, and thereby the function's extrema.

making only tiny changes and thus being wasted, and would prevent highly useful units from undergoing excessively large changes because of their large derivatives.

Finally, other empirical methods can be taken from the literature on acceleration of convergence of stochastic approximation methods (e.g., Kesten, 1958; see Fu, 1968; Tsypkin, 1971, p. 59). For example, Kesten's method is based on changes in the sign of the individual steps, in this case of the Δw_i; if the sign of the step keeps changing, oscillation is suggested, and the method reduces the step size. Similarly, repeated steps of the same sign suggest that the step size should be increased, but little work has been done pursuing this half of the idea. One particularly interesting possibility is that of using the steepest-descent concept at a second "meta" level to derive procedures for altering each weight's learning rate (as in Barto and Sutton, 1981, Appendix C).

CONCLUSION

Modern network learning procedures such as backpropagation are a significant adavance over previous connectionist learning techniques. This work is particularly exciting because the learning procedures can be directly related to their basis in the theories of stochastic approximation and gradient descent. The intent in this article has been to encourage the widening of the scope of this advance. Gradient or steepest descent is one of the simplest descent procedures, but it is neither the only nor the best one. It is unduly slow in the presence of ravines, which appear ubiquitous in the network domain, and it encourages the destruction of previously useful features upon task switches. Much is already known about the problems of steepest descent and various alternatives to it in a general setting. Just as steepest-descent theory has been successfully carried over to the network domain to produce backpropagation and the other new network learning algorithms, perhaps this other knowledge can be carried over to significantly improve the speed with which they learn.

REFERENCES

Ackley, D.H., Hinton, G.H., & Sejnowski, T.J. (1985) A learning algorithm for Boltzmann machines. *Cognitive Science* 9, 147–169.

Anderson, C.W. (in preparation) Learning new terms in connectionist systems. University of Massachusetts Ph.D. Dissertation.

Barto, A.G. & Anderson, C.W. (1985) Structural learning in connectionist systems. *Proceedings of the Seventh Annual Conference of the Cognitive Science Society*, 43–53.

Barto, A.G. & Sutton, R.S. (1981) Goal seeking components for adaptive intelligence: An initial assessment. Air Force Wright Aeronautical Laboratories/Avionics Laboratory Technical Report AFWAL-TR-81-1070, Wright-Patterson AFB, Ohio.

Derthick, M. (1984) Variations on the Boltzmann machine learning algorithm. CMU Tech Report CMU-CS-84-120.

Duda, R.O. & Hart, P.E. (1973) *Pattern Classification and Scene Analysis.* New York: Wiley.

Fu, K.S. (1968) *Sequential Methods in Pattern Recognition and Machine Learning.* New York: Academic Press.

Gill P.E., Murray W., & Wright, M.H. (1981) *Practical Optimization.* New York: Academic Press.

Kesten, H. (1958) Accelerated stochastic approximation. *Annals of Mathematical Statistics 29*, 41–59.

Rosenblatt, F. (1962) *Principles of Neurodynamics.* New York: Spartan Books.

Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1985) Learning internal representations by error propagation. Institute for Cognitive Science Technical Report 8506, UCSD, La Jolla, CA 92093.

Sejnowski, T.E. & Rosenburg, C.R. (1986) NETtalk: A parallel network that learns to read aloud. Johns Hopkins University Electrical Engineering and Computer Science Technical Report JHU/EECS-86/01.

Tsypkin, Y.Z. (1971) *Adaptation and Learning in Automatic Systems.* New York: Academic Press.

Widrow B. & Hoff, M.E. (1960) Adaptive switching circuits. *1960 WESCON Convention Record Part IV*, 96–104.

Williams, R.J. (1986) Reinforcement learning in connectionist networks: A mathematical analysis, Institute for Cognitive Science Technical Report 8605, UCSD, La Jolla, CA 92093.

INTEREST WHEN NOTHING HAPPENS; A NOTE ON NARRATIVE RETARDATION

Ed S.H. Tan
Department of Theatre Studies
University of Amsterdam

INTRODUCTION

Alfred Hitchcock is famous, among many, many other things, for his mastership in throwing out red-herrings, christened 'MacGuffins' by one of his co-workers. A MacGuffin is a part of a film that has no other function than to keep the story going, being in itself immaterial to the plot.[1] Hitchcock is, of course, not the inventor of the device. It is only one instance of a larger class of narrative material serving the purpose of withholding information that really matters, a purpose basic to the function of stories.[2] Retarding material can assume various forms, for instance, like the MacGuffin, it may not be initially discernable. Alternatively, the recipient may be aware of the fact that he or she is witheld from a desired outcome, sometimes almost painfully, as in high suspense. Furthermore, it can interact in several ways with the delayed actions and events, e.g. act as a false cue as to these or distract attention from them completely. Finally the material itself may consist of various elements like actions, descriptions, comments, etc. And there are more aspects than those just mentioned in which particular retardations differ from each other. (See for an overview Sternberg, 1978 and for applications to film narrative Bordwell, 1985.)

The best known function of delaying action or its outcome is creating suspense. (Cf. Chatman, 1978; Sternberg, 1978) A suspense discourse structure is according to Brewer and Lichtenstein (1982), characterised by an initiating event raising our concern for a protagonist, followed by an outcome or resolution. 'Typically additional discourse material is placed between the initiating event and the outcome event, to encourage the build up of the suspense' (p. 481). In a written narrative this could be achieved by a blank filled with periods, in a film one could resort to presenting a black image for some time. 'Nothing happens' literally. Such a device is generally not very satisfactory to the recipient, especially when the delay is considerable. The means will be judged to be completely subordinated to the end and interest will soon be lost. What then serves as adequate retarding material?

In this paper we shall discuss some additional functions of retarding material by presenting an analysis of an example film narrative. In connection with the additional functions, we will briefly outline some affective consequences of retardation. Finally, in analyzing our example story we will meet with some problems posed by retarding material to current story representation systems.

AN EXAMPLE: RETARDATION IN ACTION

The Film

A short (11.5 min) Dutch fiction film called 'Punishment' ('Straf', Madsen, 1974) was selected for analysis. Anticipating the structural analysis, its story can be summarized as follows:

> Marjan, an approximately eight-year-old girl, demolishes her father's violin. She then joins her little brother and her mother in the living room. Some time later the father returns home from work. They have dinner. After dinner the father goes to his room, discovers the remnants of the violin and punishes Marjan by tearing her dolls to pieces and hanging them on a rope.

Structural Analysis

Identification of elementary events

Our analysis consisted of two steps. In a norming study (Tan & De Wied, in preparation) subjects reported on what they saw happening. Following a method borrowed from Lichtenstein & Brewer(1980) a list of 181 events was obtained which was presented to another group of subjects who were asked to select clusters of events belonging together. Twenty-four groups of events resulted, from now on referred to as 'Events'. They are listed in Table 1.

Table 1. List of Events

1. Marjan goes to father's room and damages his violin.
2. Marjan is called downstairs by mother.
3. Marjan goes downstairs to the living room.
4. Marjan and Robbie play the piano.
5. Mother sits and waits.
6. Mother looks forward to father.
7. Mother announces father's arrival.
8. Marjan, Robbie and mother tidy up the room.
9. Father enters the house.
10. Marjan does not react.
11. Father performs his homecoming ritual.
12. Mother urges Marjan to lay the table.
13. Marjan lays the table.
14. Father ritually ends reading his paper.
15. Father demonstrates Marjan how to lay the table.
16. Dinner starts.
17. Father forbids Robbie to play with his food.
18. Father finds fault with everyone and everything.
19. Robbie tries to tell a story but is interrupted by father.
20. Marjan is humiliated by father.
21. All rise from the table and father goes to his room.
22. Marjan, in expectation, goes to the kitchen.
23. Marjan is fetched by father and sees the damaged dolls.
24. Marjan and father exchange looks; Marjan brings out the hidden scissors.

Identification of plot units: the main action

In order to expose the story structure, Lehnert's plot unit analysis (Lehnert, 1981) was applied to the event units. This analysis is characterized by a bottom-up search for gradually more embracing plot units. The primitive element is the affect state. An event can be positive (+), negative (-) or neutral (M) to the fate of a character. Neutral events are mental acts or states, such as wants, intentions, speech acts and the like. Affect units relate to each other by four possible links: motivation (m), actualization (a), termination (t) and equivalence (e). Pairs of affect units thus linked form simple plot units. These units combine into larger, complex ones, which may involve more than one character. For example, RETALIATION is composed as follows (A and B are characters):

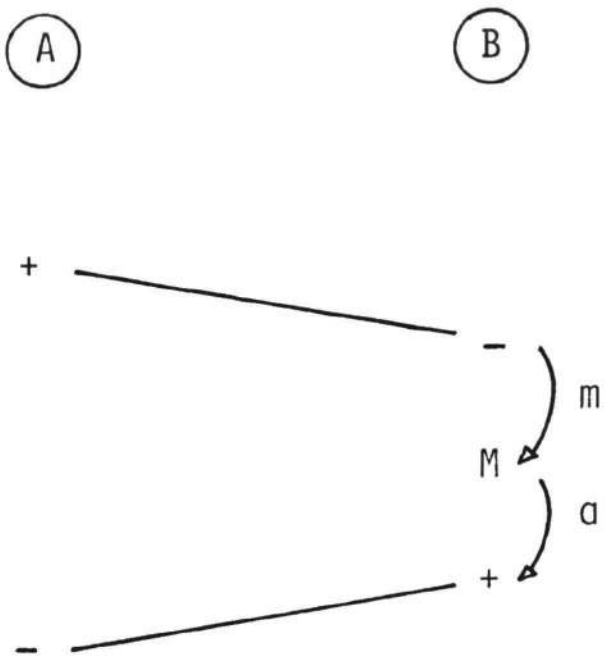


Figure 1 Complex plot unit RETALIATION including from top to bottom Mixed Event, Problem, Success born of Adversity and Mixed Event. From Lehnert (1981)

A succeeds (affect state +) somehow in harming B (-), which motivates B to think of how to harm A (M). The plan works out well (+), resulting in A being harmed. This complex plot unit is composed of overlapping smaller ones, a Mixed Event (+——-), a Problem (-⟶M), a Success Born of Adversity (-⟶M⟶+) and another Mixed Event (-——+).

Figure 2 shows the pivotal plot unit within the largest cluster, a Retaliation. The first element, the positive affect state, corresponds to Marjan's damaging the violin (Event 1). This constitutes a Mixed Event, since it means a loss to father.

The Mixed Event links Event 1 with the series of Events 22 and 23 in which father contrives a loss equivalent to his and shows the outcome to Marjan (Success Born of Adversity).

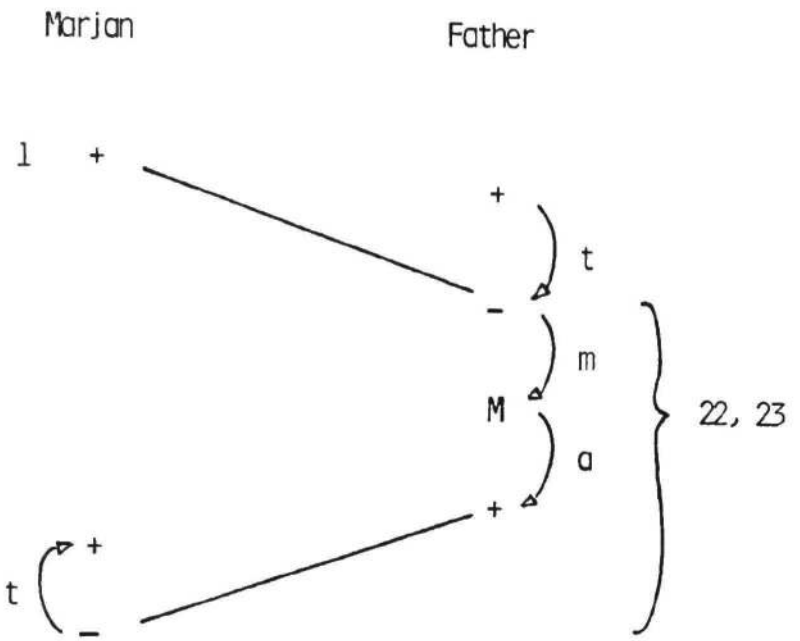


Figure 2 RETALIATION in example film. Event 22 corresponds to discovery of the smashed violin and damaging of the dolls by father.

The validity of this interpretation was tested rather informally. Twenty-three Fresh subjects were asked to name the two most important events of the film. All of them mentioned the demolition of the violin. Eighteen subjects mentioned Marjan's confrontation with the destroyed dolls in addition. Subjects were also asked to describe the reason for the second event. Sixteen out of eighteen mentioned some equivalent of Retaliation in relation to the demolition of the violin: 'revenge' (6), 'punishment' (4) 'retaliation' (3), 'striking back' (3).

The Retarding Events

Determination of the setting function

The retaliation plot unit contains both the initiating event and the resolution of the suspense discourse structure. Events 2 to 21 then, can be regarded as additional discourse material, serving as a delay in the revelation of the outcome, lasting no less than seven minutes and forty seconds in our case. So it seems that almost the whole narrative acts as a delay. That nothing happens seems impossible, then. Let us therefore consider the structural functions of these events in some detail.

If Events 2-21 do not belong to the kernel of the story, there are two possible functions within the structure left. First they can act as unimportant side lines. In this case one should try to incorporate them in the plot unit analysis. It can be easily seen however, that a number of smaller and at best loosely connected clusters will emerge as points of view determining affect states, that is, intentionally acting characters shift frequently. (Marjan, father, Robbie and mother all contribute to these events). Second, they can serve as elements of the setting of the story. It is not unusual practise to treat events which are irrelevant to a plot as elements of setting or exposition. (See Gee & Grosjean, 1984. These authors also applied Lehnert's analysis).

To test whether the latter of these functions does prevail, a new group of subjects were asked to generate a summary. It was expected that very few if any of the Events 2-21 would emerge in the summaries. This prediction was based on Lehnert's model of summarization behavior; summaries are built on pivotal units of the largest clusters.[3] If summaries must be short, side line events will be omitted. Further, it was expected that setting statements would show up that can be linked to Events 2-21. 31 new subjects were asked to generate a summary.

The average length of the summaries was about six lines. Results as to the main action corresponded very closely to these obtained in the answers concerning the most important moments. Of Events 2-21 only father's homecoming was mentioned by 6 subjects. Four subjects mentioned the word dinner. Neither mother nor Robbie were mentioned one single time. It seems reasonable therefore, not to treat Events 2-21 in the same way as the main action and to leave them out of the plot unit structure.

The summaries, short as they were, contained a great deal of setting information, distinguishable from any of the Events. Paramount were characterization statements, mostly concerning father, who was depicted as an unfeeling authority or tyrant by sixteen out of thirty-one subjects. Furthermore, eleven subjects characterized the relation between Marjan and father. ('hatred', 'power', 'competition' and the like). Three subjects characterized the family and one Marjan. Space limitations prevent us from discussing the derivation of such setting information. Apart from film stylistics, Events 11, 14 , 15 and 17 to 20 play an important part, as will easily be seen.

Relations between retarding Events and the main action

First, there are two Events that are directly related to the main action. In 9 father comes home and in 21 he goes to his room. They do not only function as retardation that builds up suspense, but more directly heighten it, by making a feared negative outcome more probable. (Comisky & Bryant (1982) have shown that suspense is maximal if a protagonist's odds are as unfavourable as possible.)

Secondly, the retarding material may be said to be indirectly linked to the main action through the setting. It is difficult to express this two-step relation within Lehnert's plot unit system. That is, setting states and events cannot be collected under some special label. This is probably due to the fact that in a bottum-up representation system instantiation of the smallest primitive is sought first. And in Lehnert's system, as in others, this is a structure having to do with the realization or non-realization of a goal.

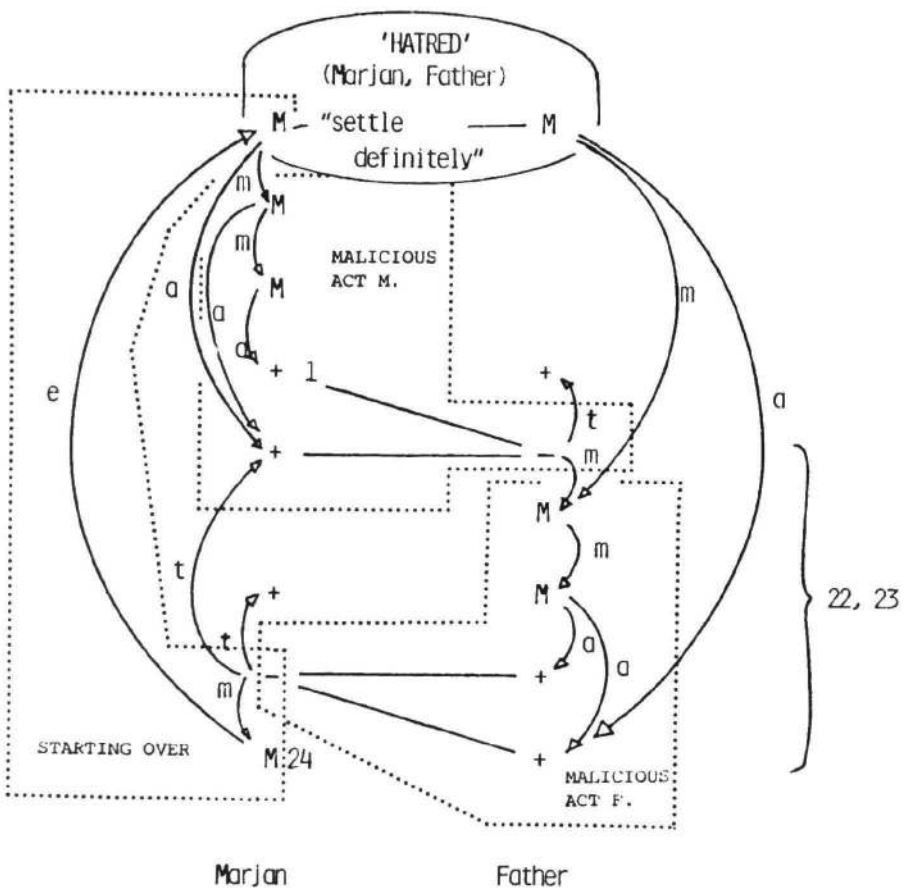


Figure 3 Extended plot unit configuration incorporating setting information. See text.

A primitive affect state reflects satisfaction or non-satisfaction of the needs of a character. However, it is possible to extend any given plot unit with antecedent and consequent units, which in a top-down story grammar driven analysis would fill the slots of a separate structural unit. Figure 3 shows the extensions which are supported by setting information extracted from the summary data.

To begin with, the destruction of the dolls is a Malicious Act, since the setting makes clear that the actor, father, is a mean person. Furthermore the setting gives us further reasons behind the two main acts. If the relation between Marjan and father is characterized by 'Hatred', 'Power' or 'Competition', then Marjan's act is also part of a Malicious Act, a contrived harming of the other producing some form of enjoyment. The father's character and the relation with his daugther enable us furthermore to consider the possibility that the struggle is to continue, that we have witnessed just one Retaliation episode. This means that the Malicious Acts are Nested Subgoals subordinate to some other goal, for instance 'settle definitely with the other'. And that this higher level goal is satisfied by a successful Malicious Act, but is activated again as soon as retaliation has taken place; this plot unit configuration is called Starting Over. It is further supported by the last Event in which Marjan pensively looks at the scissors she used in destroying the violin.

RETARDATION AND BI-DIRECTIONAL PROCESSING

The extended plot unit configuration is gradually built up in the subject during the processing of the film. Now that we have sketched the more or less complete representation of the story after the whole film has been processed, we can take a closer look at the dynamics of retardation and its affective consequences during the actual processing of the film. The first Event comes as a surprise as no expository information has been given. Curiosity is aroused as to the reason for damaging the violin. The representation is not altered very much by Events 2 to 6. Events 7, 8 and especially 9 prepare for a Retaliation and a Malicious Act by father. The latter expectation is weak, however. Simultaneously the ground is prepared for understanding the reason behind Event 1. This 'backward inference' is about as weak as the 'forward inference' of a future Malicious Act by father. After Events 10 and 11 both Malicious Acts have grown in probability and the interpersonal theme Hatred has come up. Events 12 and 13 do not change much in the probability values of forward and backward inferences. Fourteen and 15 heighten probabilities of the inferences and the interpersonal theme again. By now it will be 'certain' for most recipients that Marjan committed a Malicious

Act and that the reason for it is Hatred. Here then a first basis is laid for the Starting Over plot unit. Events 17 to 20 increase the probability of a Malicious Act by father as a retaliation. Events 21 and 22 increase the expectation of a Retaliation to a maximum. The last part of 23, where Marjan is confronted with the puppets completes most plot units: father's Malicious Act, (mutual) Hatred, father's supergoal to settle definitely with Marjan, Retaliation. Starting Over may be weakly activated. The last event completes this high-level plot unit.

This sketchy account of the process of plot unit formation may at best serve as a preliminary version of a theoretical subject protocol. Sophisticated experimentation is needed to obtain a 'timetable' representing the state of completion of the various plot units and the momentaneous strengths of the inferences. It may show, however, what the function of various retarding events can be. Some of them seem to perform no other function than to bring the progress of the story towards a resolution to a halt: Events 2 to 6, 12 and 13 act like this. They contribute to suspense in that they furnish the 'additional discourse material'. Events 2 and 3 can be 'postdicted' (see Kintsch, 1980) because a change of location is, as turns out soon, required by the main action.This insight, however, is not realized at the moment of their presentation. Events 21 and 22 are also merely retarding, but the level of suspense is so high at that point, that they are probably not experienced as 'unmotivated cheap tricks'. Moreover, their function is to raise suspense still further, by increasing 'expectation strength'.[4] The remaining retarding Events seem to help in 'bi-directional processing'(Sternberg, 1978): They establish forward expectations which increase suspense and at the same time resolve curiosity due to gaps in the representation of the story left at an earlier point. The recipient is thus both rewarded and frustrated at the same time which as has been noted by other researchers (Kintsch, 1980;Moynihan & Mehrabian, 1981), is a condition favourable for maintaining a high level of interest.

Finally, we conclude that the cognitive representation of background knowledge in story processing is not to be neglected if understanding in some depth of affect in the processing subject is sought for. More specifically, an account of such knowledge is needed in order to explain why narratives can be arresting when nothing of importance to goal directed action happens. Representation systems based on bottom-up processing, including Lehnert's proposal[5] seem to be built on primitives involving an implicit notion of goal-directedness and therefore, in need of extension with units representing the static 'given' (exposition ,characterization, atmosphere) in relation to the dynamic 'new' (goal-directed action).

NOTES

1. Two men travel to Scotland on a train. In the luggage rack there is an odd looking parcel. 'What have you there?', asks one of the men. 'Oh, that's a Macguffin.''What's a MacGuffin?' 'It's a device for trapping lions in the Scottish Highlands.' 'But there are no lions in the Scottish Highlands'. 'Well, then, I guess that's no MacGuffin.' (From Spoto, 1984).
2. Moynihan & Mehrabian present evidence suggesting that postponement of knowledge, provided that it is counterbalanced by a resolution, makes a story by far the most preferred among eight alternative types. In addition they quote sources from the theory of literature to the same effect.
3. We have, for convenience, dropped the distinction made by Lehnert between a 'family' and a 'cluster'. We only speak of 'clusters', employing the term in a rather loose sense synonymous with 'groups'.
4. Frijda (in press) coins the term 'expectation strength' which refers to one of the major determinants of the intensity of emotion in general. It is influenced by factors like temporal and spatial proximity.
5. In research not reported here, an empirically based representation of the same film's narrative was sought for by following the bottom-up procedure outlined by Lichtenstein & Brewer (1980). Subjects were asked to indicate in-order-to relations they felt to exist between events obtained before. Subjects reported that this seemed impossible and to make no sense, since many events did not serve any purpose in terms of other events.

REFERENCES

Bordwell, D. (1985). Narration in the fiction film. Madison (WI), University of Wisconsin Press.

Brewer, W.F. & Lichtenstein, E.H. (1982). Stories are to entertain: a structural affect theory of stories. Journal of Pragmatics, 6, 473-486.

Chatman, S. (1978). Narrative structure in fiction and film. Ithaca & London: Cornell University Press.

Comisky, P. & Bryant, J. (1982). Factors involved in generating suspense. Human Communication Research, 9, 1, 49-85.

Frijda, N.H. (in press). The Emotions. New York: Cambridge University Press.

Gee, J.P. & Grosjean, F. (1984). Empirical evidence for narra-

tive structure. Cognitive Science, 8, 1, 59-85.
Kintsch, W. (1980). Learning from text, levels of comprehension, or why anyone would read a story anyway. Poetics, 9, 87-98.
Lehnert, W. (1981). Plot units and narrative summarization. Cognitive Science, 4, 293-331.
Lichtenstein, H. & Brewer, W.F. (1980). Memory for goal-directed events. Cognitive Psychology,12, 412-445.
Moynihan, C. & Mehrabian, A. (1981). The psychological aesthetics of nararative forms. In: H.I. Day (ed.), Advances in intrinsic motivation and aesthetics. New York: Plenum.
Spoto, D. (1984). The dark side of genius: The life of Alfred Hitchcock. New York: Ballantine.
Sternberg, M. (1978). Expositional modes and temporal ordering in fiction. Baltimore & London: Johns Hopkins.
Tan, E. & de Wied, M.A. (in preparation). De normering van elementaire gebeurtenissen in de film 'Straf'. Receptie van Theater en Media, 14. Utrecht & Amsterdam: Instituten voor Theaterwetenschap. With a summary in English.

A Model for Parsing, Learning and Recognizing Objects in a Complex Environment

Arnold Trehub
Department of Psychology
University of Massachusetts, Amherst

ABSTRACT

A neuronal model is described that can parse, learn, and recognize objects in a complex visual environment. A computer simulation of the model network was tested with a variety of scenes and exhibits competent performance.

INTRODUCTION

The problem of cognitive adaptation without a prior knowledge base constitutes a ubiquitous and vexing issue in cognitive science. Imagine a person in an absolutely unfamiliar environment, one in which all visual patterns are completely novel. Where would the person look? Since any point of gaze would presumably be no more meaningful than another, how could one parse the scene into objects? How could the objects be learned and committed to memory? This paper presents a computer simulation of a detailed neuronal system that is plausible within biological constraints and can accomplish these fundamental visual-cognitive tasks. The model is composed of several putative neuronal mechanisms proposed in earlier papers (Trehub, 1975, 1977, in press) which have been organized in an integrated system that can deal competently with novel and complex visual environments. The neuronal model will be briefly described and then a computer simulation of the model's behaviour will be presented.

NEURONAL MODEL

Following is an outline of the principal processing elements in the model.

1. Center-surround mechanisms in the retina and lower-level visual nuclei extract simple contours from the light-intensity array.
2. There are cells which integrate contour excitation over small, discrete regions of the entire visual field. These are called flux detectors and serve to drive visual saccades to regions of maximum contour flux.
3. There is a visual field constriction mechanism that can limit the effective stimulus input to an area of variable retinal diameter centered on the foveal axis
4. There is a post-retinal dynamic visual buffer called a retinoid which can translate patterns of retinal stimulation over an egocentric coordinate space. This module locates and positions pattern centroids on a standard reference axis within the visual system.
5. There is an adaptive network called a synaptic matrix which can learn, recognize and image visual patterns.

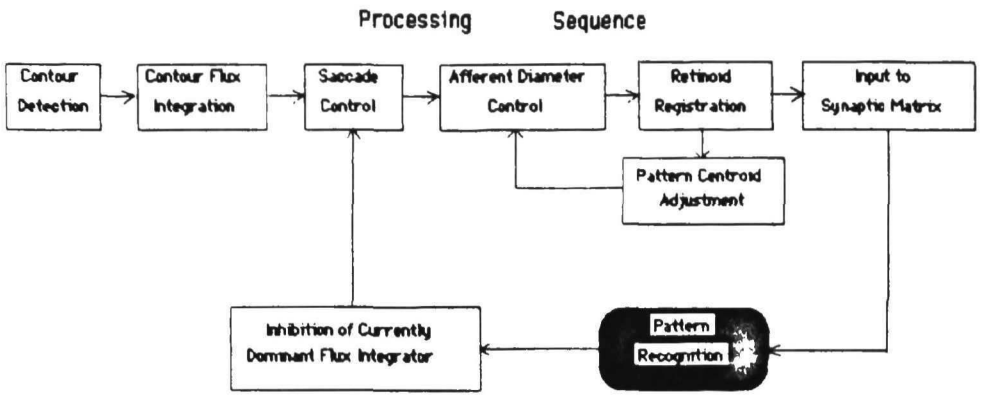


FIGURE 1. Block-flow diagram of processing sequence.

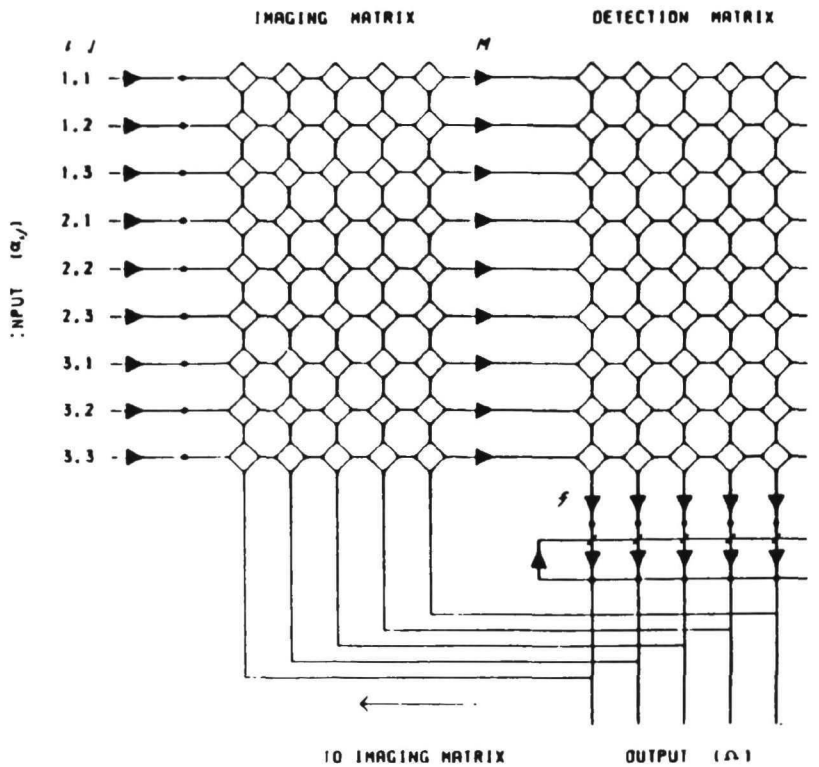


FIGURE 2. Schematic of a synaptic matrix. Afferent inputs from optic tract designated α_{ij}. Mosaic cells designated M. Dots represent fixed excitatory synapses. Short oblique slashes represent fixed inhibitory synapses. Lozenges represent adaptive excitatory synapses. Reset neurons marked (-) generates an inhibitory postsynaptic potential to reset all class cells (Ω) when discharged. Given an arbitrary pattern input, that class cell coupled with the filter cell (f) having the highest product-sum of afferent axon activity ($M_{\alpha ij}$) and corresponding synaptic transfer weights (ϕ_{ij}), will fire first and inhibit the output of all competing class cells.

The diagram shown in Fig. 1 gives a rough representation of the processing sequence. The major modules are outlined below. Space limitations preclude a more detailed presentation of their operating principles which can be found in other publications (Trehub, 1975, 1977, 1985, in press).

Synaptic Matrix. Figure 2 shows a basic version of the neuronal network that has the capability of learning complex retinal input patterns. If a pattern exemplar has been learned, subsequent stimulation by a similar pattern results in the discharge of a particular output cell (class cell) that has been associated with the original exemplar during the learning process. In effect, this cell represent the biological name of its associated pattern. Conversely, the discharge of a class cell alone can generate in an array of mosaic cells the afferent firing pattern (image) initially evoked only by the learned retinal stimulus. Learning occurs in the detection-matrix field when mosaic cells carrying an input pattern fire in virtual coincidence with the discharge of a previously unmodified filter cell, and in the imaging-matrix field when a class cell is fired in coincidence with discharge in the mosaic-cell array. The physical substrate of learning is an adaptive long-term change in the distribution of synaptic transfer weights (ϕ) on the dendrites of filter cells and mosaic cells.

Retinoids. The neuronal structure shown in Fig. 3 is a post-retinal mechanism called a retinoid because it represents visual space and projects afferents to the mosaic-cell array. This module may be thought of as a visual scratch-pad with phasic and dynamic content. The medium of storage is assumed to be a retinotopically organized sheet of excitatory autaptic neurons. Cells of this type have at least one of their axon collaterals in recurrent excitatory synapse with their own cell body or dendrite (Shepherd, 1974).

If there is a pattern of excitation evoked on a retinoid, this captured pattern can be spatially translated in any direction by appropriate pulses from the shift command cells. For example, each pulse from the shift-right line will transfer standing activity from each active autaptic cell to the adjacent autaptic cell on its right and, at the same time, erase activity in the previously active cell (the donor cell) unless that cell is also receiving transfered excitation from an autaptic cell to its immediate left. The more rapid the pulses, the more rapid will the pattern move; the longer the pulse train is sustained, the greater will be the distance over which the pattern is moved. Appropriate sequences of shift right/left, shift up/down, can move the pattern of cell activity to any position on the retinoid surface.

Imagine the retinoid as a quadrantally organized surface, with each quadrant receiving retinotopic afferents from its respective retinal quadrant. If the excitation of a standing pattern is summed independently over each quadrant, and if the relative magnitudes of the summed discharges are used to drive either the position of the eye or the shift control cells in the retinoid, then we have a neuronal mechanism which can align the centroid of any retinal stimulus with the central axis of retinoid space (Trehub, 1985). We define the normal foveal axis as that axis corresponding to the line of sight of the fovea when the eyes are straight ahead, the head unturned, and the shoulders square with the body. It is assumed that the central axis in retinoid space corresponds with the normal foveal axis.

The quandrantal summation fields for retinoid output are abbreviated as follows:

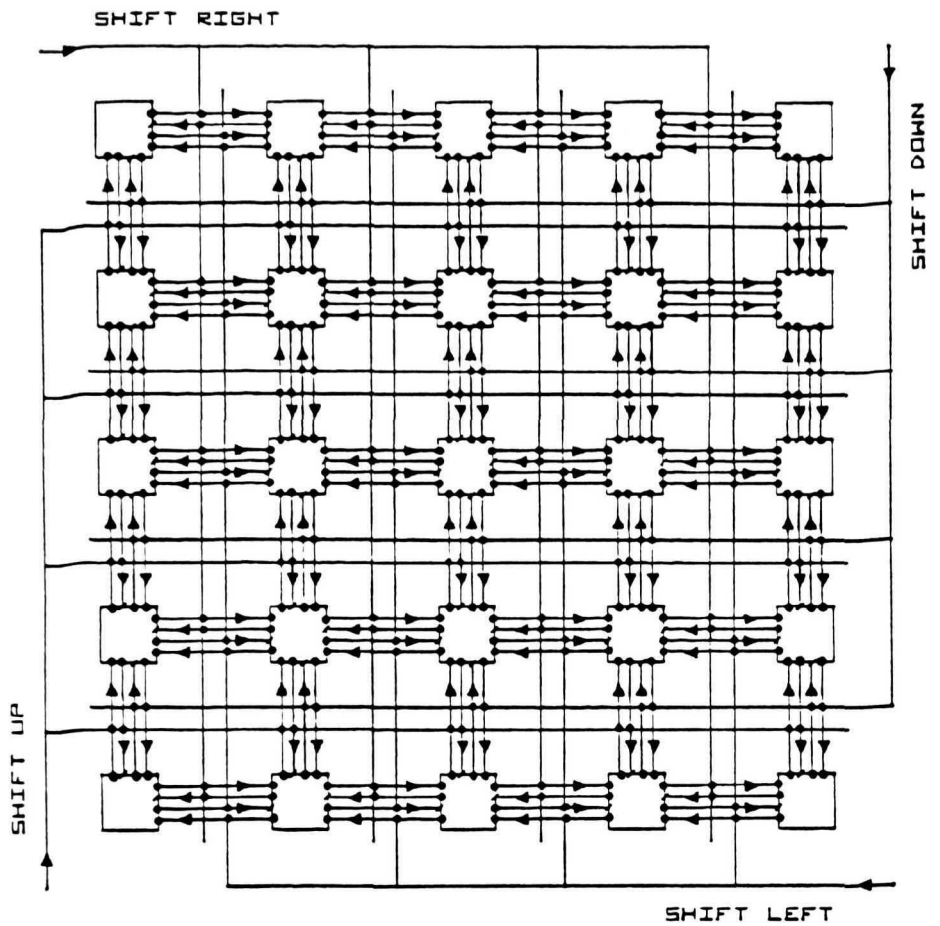


FIGURE 3. Translation retinoid. Large squares represent autaptic cells serving short-term memory. Small filled triangles represent interneurons. Shift-control cells designated by direction of effect.

LF = output from left retinoid field.
RF = output from right retinoid field.
TF = output from top retinoid field.
BF = output from bottom retinoid field.

If the difference between total output in RF-LF and TF-BF respectively were to drive an eyeball in the direction of the greater excitation in the hemi fields defined by the two orthogonal axes, the fovea would hunt until it targeted the contour centroid of any stimulus pattern presented to the retina. Alternatively, if the point of eye fixation does not change, then a pattern with a parafoveal centroid can be translated over the retinoid surface so that its centroid falls on the normal foveal axis of the retinoid. This is done by using the hemifield mismatches to drive the

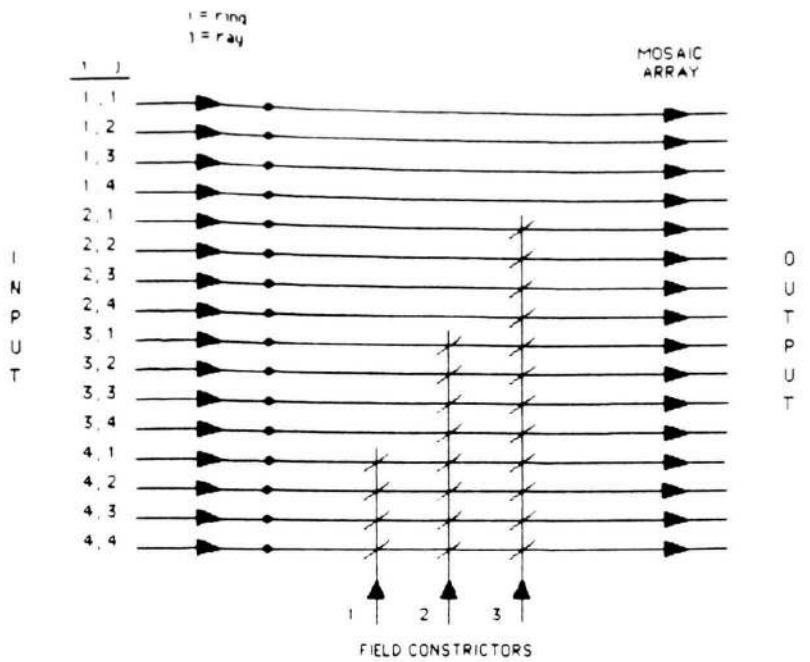


FIGURE 4. Controls for constricting effective visual field. Discharge of constrictor neuron 1 blocks input from ring 4 (outer ring); discharge of constrictor 3 blocks input from rings 4, 3, 2, restricting input to ring 1, the innermost ring of afferents.

shift-control cells so that excitation is balanced over retinoid quadrants.

Field Constrictor. It is possible to devise a number of different coordinate representations for retinotopic indexing, but I have found a ring-ray representation to be particularly useful and efficient. In this scheme, receptor cells in the retina and their associated afferent projections are indexed with respect to the central foveal axis in terms of their locations on imaginary concentric rings (i) centering on the axis, and imaginary rays (j) projecting from the axis and intersecting all rings. This retinal organization easily lends itself to CNS control of the afferent field aperture. Figure 4 shows how inhibitory neurons can impinge successively on entire rings of mosaic cells to constrict the diameter of the effective visual field.

COMPUTER SIMULATION

A 22x22 cell retina and the neuronal mechanisms outlined above were simulated in a digital computer. Indoor (near) and outdoor (far) environments were created in sketch-to-pixel conversions, and these environments were presented to the simulated visual system for parsing, learning, and object recognition.

At the start of each scene-parsing operation, the model first fixated on the retinotopic locus of the flux detector with maximum output, then

the afferent field aperture closed to the fully constricted state which was arbitrarily set at six retinal units in width and height. The fully expanded afferent aperture was limited to 22x22 retinal units. Whenever the visual aperture reached the state of full expansion, the excitation pattern on the retinoid was gated to the synaptic matrix for recognition (and learning if the pattern was incorrectly identified). Starting error tolerance was set at three units for quadrantal disparity over either the horizontal or vertical axes. At any fixed aperture, if error tolerance was exceeded on a given axis, the retinoid pattern was shifted in the appropriate direction to reduce hemifield disparity on that axis. When pattern position satisfied error tolerance for one axis, the pattern was shifted on the other axis, unless it was already within tolerance. If, now, shifting the image on the second axis resulted in an unacceptable error on the first, error tolerance was relaxed one unit. Whenever the pattern was brought within axial tolerance for both horizontal and vertical disparities, the afferent aperture expanded one unit and the process was repeated until full aperture was achieved. This operation was assumed to involve an expenditure of processing effort, and if a retinoid shift of nine units on any axis did not bring its disparity within tolerance limits, the system stopped trying at its current fixation and initiated a saccade to the next highest flux region.

In its initial state, the neuronal system is presented with a random visual pattern and taught to call this pattern "RANDOM". This simply means that one filter cell in the detection matrix and a spatially correlated array of mosaic cells in the imaging matrix have been synaptically tuned to the random exemplar.

The first "natural" environment learned was an outdoor scene consisting of trees, a house, several animals, a building, a car, and the outline of distant hills. Since the simulation does not incorporate mechanisms of visual accommodation or stereopsis (see Trehub, 1978), the operator is asked by the model to provide a rough estimate (in feet) of its viewing distance from the major elements of the scene. The operator estimates the distance as 200 feet and provides this information to the network. Parsing then proceeds according to the principles discussed above. After a pattern has been fixated and registered on the retinoid, it is passed to the synaptic matrix where it is identified and named as "RANDOM" because, in its utterly naive condition, this is its only available response. The model then asks the operator to inform it if the response is right or wrong. Let us say that the object it has happened to parse is a house or part of a house; then it is told that the response is wrong. At this point, the model changes the synaptic weights on a previously unmodified filter cell in accordance with the excitation pattern on its mosaic-cell array and the learning equation. It should be noted here that if there were no operator to inform the system about the correctness of its response, low frequency discharge of its filter cells can provide a signal that the current stimulus is novel, triggering the automatic learning of the novel object (Trehub, 1977). After the filter cell has been tuned to the stimulus, the model asks for a name to be associated with the class cell which is coupled to the just-modified filter cell. The operator then provides the appropriate name "HOUSE". This name then becomes part of a neuronal lexicon in which it is connected with the filter-cell-class-cell couplet which has just learned the exemplar of a house. The model then parses another object and if its recognition response is correct, parsing continues; if incorrect,

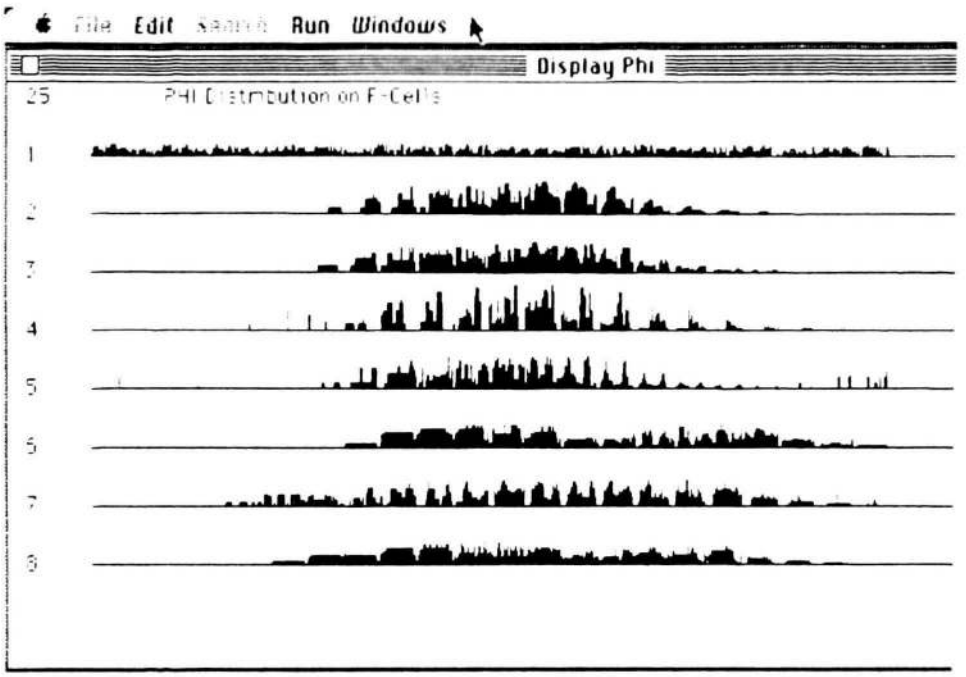


FIGURE 5. Examples of the distribution of synaptic transfer weights on dendrites of filter cells which have learned visual patterns. Each point on the dendritic line represents a particular synaptic location. Amplitude of each vertical line represents relative magnitude of transfer weight for that synapse. The objects learned by the cells shown are as follows: (1) a random visual pattern; (2) a car; (3) a different car; (4) an animal; (5) a different animal; (6) a building; (7) a house; (8) a different building.

the new object is learned (synaptic modification of another available filter cell, etc) and scene processing continues until a preset number of saccades are made, during which objects are fixated, translated to the normal foveal axis, recognized, and learned if necessary.

The second environment learned was a desktop with a book, telephone, ashtray, pencil, and bookmark. The viewing distance, in this case, was estimated to be five feet. Parsing and learning the objects in this scene then proceeded as in the outdoor environment. Variations of both kinds of environments were created and exposed to the model until a total of 25 exemplars of objects in these scenes were learned together with their appropriate names. Examples of synaptic transfer-weight (ϕ) distributions on filter-cell dendrites for the first eight patterns learned are shown in Fig. 5. The selectivity of recognition response is determined by the differences among such ϕ-distributions over the population of filter cells in the detection matrix. As the repertoire of exemplar-tuned filter cells increased, the frequency of recognition errors decreased.

Shown in Fig. 6 is a run of the simulation printed directly from the computer's CRT. In this case, the model was "looking at" an unfamiliar outdoor scene, in that all the patterns in the environment were new exemp

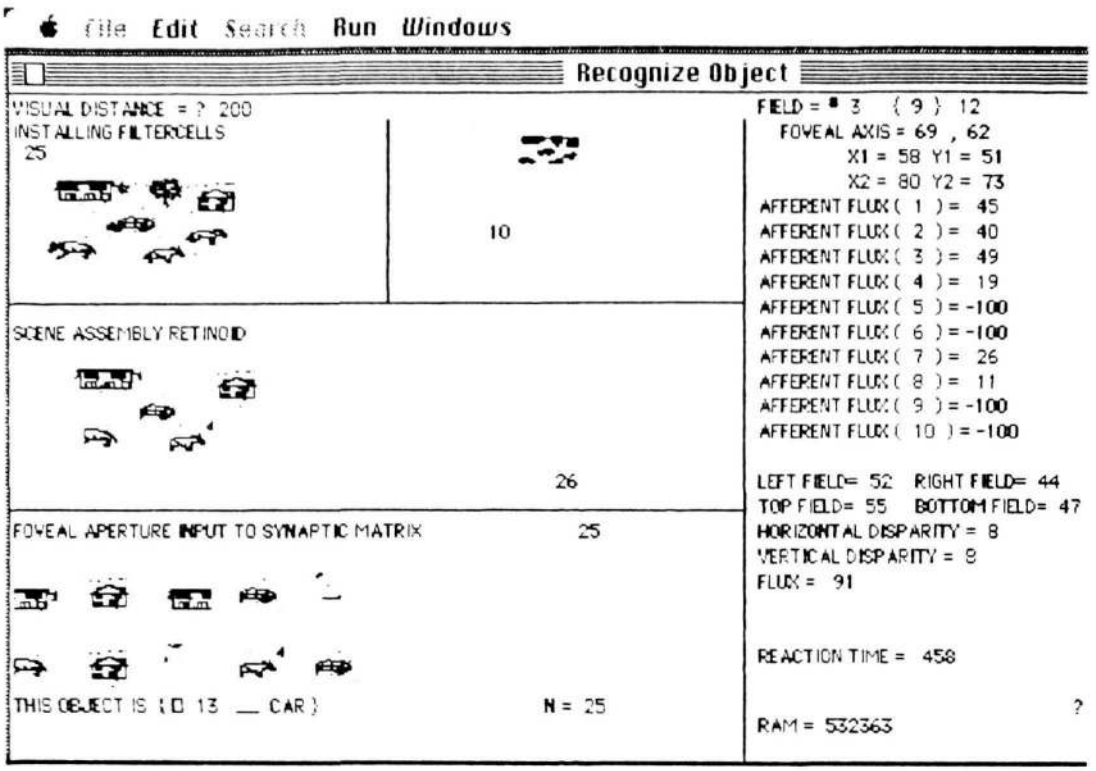


FIGURE 6. Passive recognition. Model's responses to outdoor scene. Top left frame is the scene presented. Bottom left shows objects parsed and recognized. All objects were correctly identified. Middle left is the visual reconstruction of the scene on a retinoid surface on the basis of the disparate fixations and parsings.

lars of previously learned objects and their locations were different. Figure 7 is a similar printout of a situation in which the model is "asked" to find named objects on a cluttered desktop. Here parsing and recognition is made even more difficult by the fact that a bookmark has been placed on the book and a substantial part of the book is covered with a sheet of paper. It has been conjectured that occlusions of this kind as well as the conjunction of nearby objects would make it impossible for template/filter models to operate properly (Pinker, 1984). The successful performance of the model described here suggests that the conjecture is incorrect.

In summary, computer simulation of an explicit and biologically plausible neuronal model demonstrates that a visual system that integrates (a) contour flux detection, (b) flux-driven saccades, (c) control of afferent-field aperture, (d) a retinoid for pattern centroid alignment, and (e) a synaptic matrix for pattern learning can start without an initial store of world knowledge, be exposed to novel and complex scenes, and build an appropriate knowledge base. Confronted with a rich, new visual environment, it isolates objects, learns them, and recognizes similar objects in other environments.

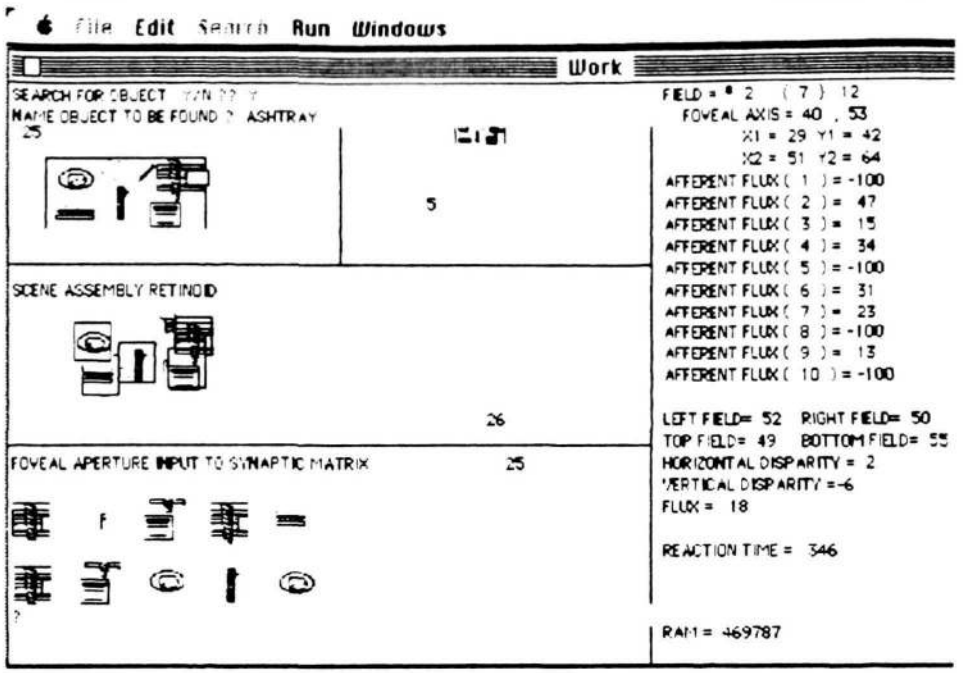


FIGURE 7. Active search and recognition. Model's responses to desktop scene. Small rectangular frame around parsed objects on the scene assembly retinoid indicates that a searched-for object has been found.

REFERENCES

Pinker, S. (1984). Visual cognition: An introduction. Cognition, 18, 1-63.

Shepherd, G. M. (1974). The Synaptic Organization of the Brain. New York: Oxford University Press.

Trehub, A. (1975). Adaptive pattern processing in the visual system. International Journal of Man-Machine Studies, 7, 439-446.

Trehub, A. (1977). Neuronal models for cognitive processes: Networks for learning, perception and imagination. Journal of Theoretical Biology, 65, 141-169.

Trehub, A. (1978). Neuronal model for stereoscopic vision. Journal of Theoretical Biology, 71, 479-486.

Trehub, A. (1985). A confusion matrix for hand-printed alphabetic characters: Testing a neuronal model. Eighth Symposium on Quantitative Analysis of Behavior, at Harvard. (Cambridge, Massachusetts).

Trehub, A. (In press). Visual-cognitive neuronal networks. In M. A. Arbib and A. R. Hanson (Eds.), Vision, Brain, and Cooperative Computation. Cambridge: MIT Press.

MENTAL REPRESENTATION OF SPATIAL INFORMATION: A PRODUCTION SYSTEM MODEL FOR PRIMING AND VERIFICATION

Karl F. Wender, Monika Wagener, and Bernd Wittmann
Institut für Psychologie
Technische Universität Braunschweig

ABSTRACT

In this contribution we investigate the mental representation of spatial information. In a previous paper (Wender & Wagener, 1985) we reported results that supported the idea of spatial information being mentally represented in an analogue format. This paper extends the previous results. Furthermore a simulation model is developed that describes the data from two different experimental tasks, a priming technique and a sentence verification test. These experimental tasks gave results that in part look contradictory. The model discussed is an attempt to account for this contradiction.

INTRODUCTION

The format of mental representations has been under discussion for several years now. We don't want to review this discussion. Our basic opinion is that propositional as well as analogue representations are possible. To some extent the format may be under the control of the person. Furthermore, the form of the representation and the processes working upon it interact. This interaction may also contain a tradeoff between building efforts and processing times. If a person expects a spatial task involving spatial judgments probably a spatial representation is formed which may cost more effort. If a person expects a more abstract task then a propositional representation may be preferred. Also, a person may be able to change the representation if the type of the task requires this.

The mental representation of spatial information is investigated in this paper by a priming technique and a sentence verification test. Both methods consider the relationship between decision times and spatial distances. If decision times depend on spatial distances but not on distances in a propositional network then this is taken as evidence for the analogue nature of the mental representation.

In the priming technique on each trial two words are presented to the subject who has to decide whether these two words belong to the learned material or not (Ratcliff & McKoon, 1978; Wagener & Wender, 1985). The idea behind this technique is that activation of the first concept in memory starts a spread of activation. It then depends on the distance between the first and second concept how much the decision is facilitated.

This research was supported by the German Science Foundation (DFG) under grant We 498/9-2.

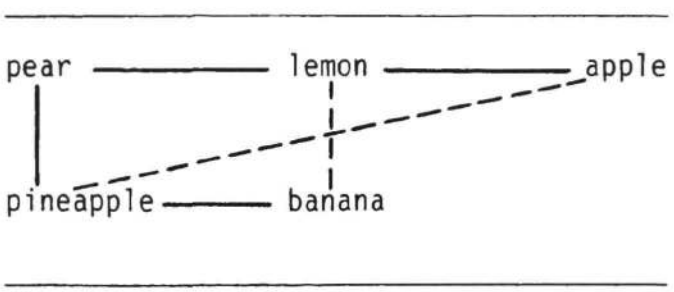


Figure 1: An example for the configurations. Solid lines correspond to sentences given to the subjects dashed lines mark relations not mentioned in the descriptions.

In the verification test a short sentence containing a spatial relation between two concepts is shown to the subject. The subject has to decide whether this sentence is true with respect to the to be remembered material or not. In linear orderings, e.g. the natural number system, a very reliable effect has been found: Decision times decrease with growing distances between concepts. This has been called the symbolic distance effect (Banks, 1977; Moyer & Dumais, 1978). Curiously, this effect seems to contradict the priming effect. Combined with the distance effect is the socalled end-anchor-effect. End anchors are the first and the last item of a finite list. If these items are combined with other items in a comparative sentence this leads to shorter decision times compared to other items with the same distance on the underlying scale.

It is the goal of this contribution to construct a model for the mental representation and processes that can explain the results from both experimental techniques. This model processes spatial information and simulates propositional as well as analogue aspects of the mental representation.

EXPERIMENT 1

Method

Subjects memorized desriptions of spatial arrangements of five common objects. Each description consisted of four sentences and was printed on a card given to the subjects. An example of one configuration is as follows.

"Fruits

The pineapple is to the left of the banana.
The pear is behind the banana.
The lemon is to the right of the pear.
The apple is to the right of the lemon."

Figure 1 shows the configuration described by these four sentences. In the experiment subjects learned eight descriptions plus two additional ones for warming up purposes. As a criterion for learning subjects were asked four questions about each configuration such as:

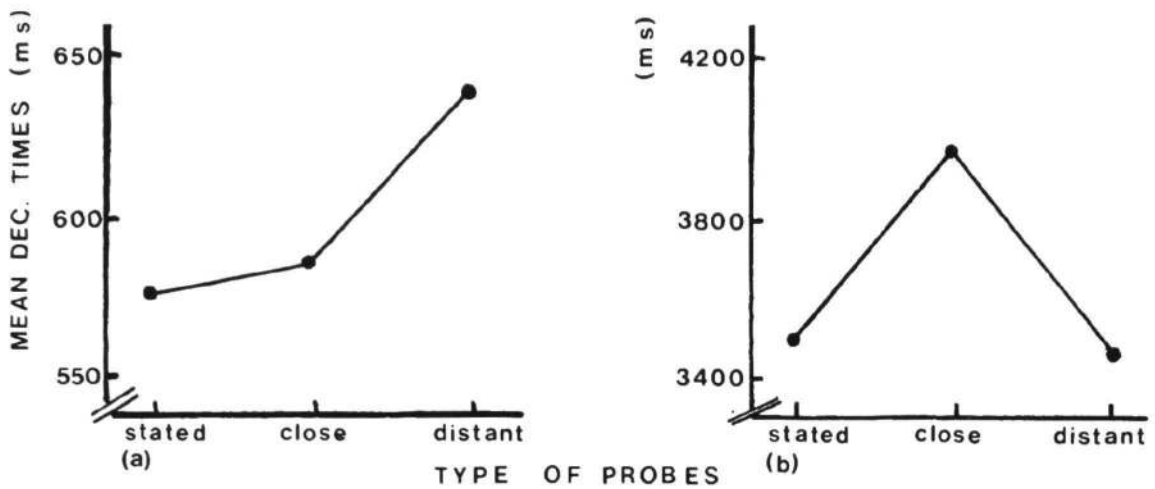


Figure 2: Mean decision times for Experiment 1 in (a) the priming task and (b) the verification task for stated, close, and distant probes.

"What is to the left of the banana?"

Learning continued until subjects could correctly answer all questions by heart.

After completing the learning task subjects participated in a priming phase. On each trial subjects were presented with the words for two objects from one configuration. The two words were shown successively on a computer screen with a 500 ms SOA. Subjects had to decide whether both objects belonged to the learned configuration or not. Decision time was measured starting with the presentation of the second word.

For each configuration two pairs of concepts were of special interest. These are called the close pair and the distant pair. In Figure 1 these pairs are marked by dashed lines. To balance possible semantic effects there were two versions for each configuration in which the objects for the close and distant pairs were interchanged.

In the verification test subjects were shown sentences of the following form:

"The pineapple is to the left of the banana."

Subjects had to decide whether this sentence was correct with respect to the given configuration or not. Again decision time was measured.

RESULTS

Figure 2 gives the results for different probes. Stated pairs correspond to sentences that were given in the description. An example from Figure 1 would be "lemon - pear". Close pairs contain words that originally did not appear in one sentence but are close together in the spatial configuration like "lemon - banana" in Figure 1. Note that distant pairs are differently constructed in

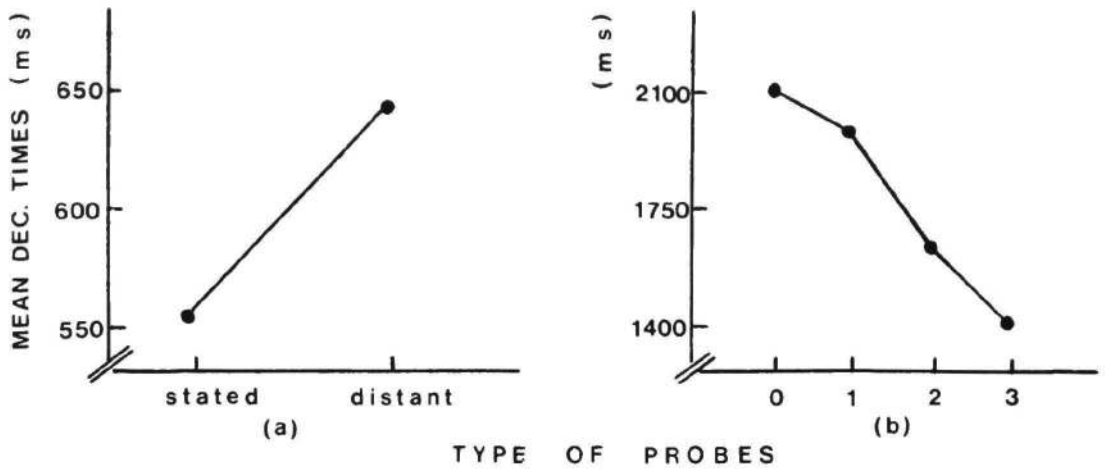


Figure 3: Mean decision times for Experiment 2 in (a) the priming task for close and distant probes and (b) the verification task for the four different distances determined by the number of objects between two test items.

the priming and verification task. In the priming task the distant pairs contain words that have the same propositional distance as the close pairs but a larger spatial distance. In Figure 1 this would be "pineapple - apple". For the verification task a distant probe would be "pear - apple". Again, these words had not appeared in one sentence. Hence, a proposition about their relationship may be called an inference. For the priming data there is an increase in reaction time from close pairs to distant pairs ($F1(1,22)=7.2$, $p < .02$; $F2(1,14)=5.51$, $p < .05$), whereas stated and close pairs are almost equal. We observed an opposite effect for the verification data: a clear decrease in reaction times from close probes to distant and stated probes ($F1(2,89)=4,53$, $p < .02$; $F2(2,21)=5.02$, $p < .05$).

EXPERIMENT 2

To investigate the processes involved in the priming and verification task in more detail we conducted a second experiment. The two dimensional configurations were reduced to one dimension.

Method

Subjects learned lists of five words each that were given to them printed on cards such as

"Fruits
pear pineapple lemon banana apple."

Each list was learned until subjects, after a short distractor task, could reproduce the complete list. The following priming task was identical to that in Experiment 1. The verification test was somewhat different from Experiment 1. Instead of a complete sentence subjects were shown two words on a screen printed on the same line. Subjects had to judge whether the words had appeared in the list in the same order. Different probes were constructed in such a way that the words had been 0, 1, 2 or 3 items apart in the original list.

```
(MEMBER-OF-LIST   A)          (NEIGHBOR A B)
 POSITION: 1.1                 ACTIVATION: 0

(MEMBER-OF-LIST   B)          (NEIGHBOR B C)
 POSITION: 0.6                 ACTIVATION: 0

(MEMBER-OF-LIST   C)          (NEIGHBOR C D)
 POSITION: 0.1                 ACTIVATION: 0

(MEMBER-OF-LIST   D)          (NEIGHBOR D E)
 POSITION: -0.4                ACTIVATION: 0

(MEMBER-OF-LIST   E)
 POSITION: -0.9
```

Figure 4: Contents of long-term memory after the encoding step

RESULTS

Mean decision times are shown in Figure 3. For the priming data times are longer for distant pairs (min $F'(1,30)=4,18$, $p < .05$) pairs whereas there is an inverse relationship for the verification data: Decision times are shorter for distant compared to adjacent probes (min $F'(3,94)=10,81$, $p < .001$).

DISCUSSION

The results from Experiment 2 suggest a rather straightforward interpretation. The priming data may be explained by a spread of activation starting from the first concept and reaching the second concept earlier in an adjacent pair than in a distant pair. The results of the verification test show a clear symbolic distance effect possibly combined with an end-anchor effect. Several models have been proposed for its explanation; but none of them combines priming and verification data. The model we propose makes use of the PRISM architecture by Langley (Langley, 1983). The model consists of one working memory, one long-term memory and three production memories.

The long-term memory stores the simulation of the mental representation. The working memory contains goals, test items, and items activated from long-term memory. The first production memory includes productions that input information to long-term memory and productions that control the work of the other two production memories. The second production memory holds the productions for the priming task. The productions used in the verification task are included in the third production memory.

Figure 4 shows the content of the long-term memory after a list of five items, A B C D E, has been learned. These elements are linked via common symbols. Our

model simulates the priming task by using the propagation process furnished by PRISM. The activation starts from the first element in the probe and successively activates neighboring elements. If the activation of an element reaches a threshold this element is transferred to the working memory and matched against the probe item.

The verification task has a different goal structure and needs a more complex and strategic process. An effective and short representation for this task is a list of elements where the ordering is represented by an attribute, as shown in Figure 4. We assume that the spreading activation is not relevant for this strategic process. Hence, the verification process needs only the "(MEMBER-OF-LIST Y)" elements. This representation corresponds to the model of semantic coding by Banks (1977). Because we assume an asymetrical end-anchor effect the first anchor has a higher priority than the last one. This has to be confirmed by further experiments. Furthermore, this simple model has to be modified and tested for data from the two dimensional configurations in Experiment 1.

Following the semantic code model we explain the symbolic distance effect by four steps of processing: first the encoding step, second the discrimination process, third the matching process and finally the decision by the subject. We assume that subjects have finished the encoding step before they start the task, since they have had time enough to remember the list and they were trained with two probe lists.

The crucial step is the discrimination process which utilizes the difference between the attribute values to discriminate between objects. If this difference is high, as for items A and E, this pair reaches a given threshold in fewer cycles than, for instance, items C and D.

The results from Experiment 1 pose an additional problem for the model. The not-stated close pairs lead to a fast decision in the priming but, compared to the stated pairs, to a slow decision in the verification task.

The priming data are simulated by including an additional element into long-term memory. For the configuration in Figure 1 this would be the element "(NEIGHBOR lemon banana)". For the verification data we assume that the attributes representing the spatial position have different values for the two dimensions of the configuration. Our model is not yet complete. It is conceivable that we will have to add elements to long-term memory representing surface information from the original sentences.

REFERENCES

Banks, W.P. (1977). Encoding and processing of symbolic information in comparative judgements. In G.H. Bower (Ed.), The Psychology of Learning and Motivation (Vol. 11, pp. 101-159). New York: Academic Press.

Langley, P. (1983). Exploring the space of cognitive architectures. Behavior Research Methods & Instrumentation, 15, 289-299.

Moyer, R.S., & Dumais, S.T. (1978). Mental comparison. In G.H. Bower (Ed.), The Psychology of Learning and Motivation (Vol. 12, pp. 117-155). New York: Academic Press.

Ratcliff, R., & McKoon, G. (1978). Priming in item recognition: Evidence for the propositional structure of sentences. Journal of Verbal Learning and Verbal Behavior, 17, 403-417.

Wagener, M., & Wender, K.F. (1985). Spatial representations and inference processes in memory for text. In G. Rickheit & H. Strohner (Eds.), Inferences in Text Processing (pp. 115-136). Amsterdam: North-Holland.

Wender, K.F., & Wagener, M. (1985). Spatial inferences and discourse comprehension. Proceedings of the Seventh Annual Conference of the Cognitive Science Society, 376-384.

INVERTING A CONNECTIONIST NETWORK MAPPING BY BACK-PROPAGATION OF ERROR

Ronald J. Williams
Institute for Cognitive Science
University of California, San Diego

ABSTRACT

The *back-propagation* learning algorithm (Rumelhart, Hinton, & Williams, 1986) for connectionist networks works by adjusting the weights along the negative of the gradient in weight space of a standard error measure. The back-propagation technique is simply an efficient and entirely local means of computing this gradient. Using what is essentially the same back-propagation scheme, one may instead compute the gradient of this error measure in the space of input activation vectors; this gives rise to an algorithm for inverting the mapping performed by a network with specified weights. In this case the error is propagated back to the input units and it is the activations of these units — rather than the values of the weights in the network — that are adjusted so that a specified output pattern is evoked. This technique is illustrated here with a small network which is a much simplified version of the NETtalk text-to-speech network studied by Sejnowski and Rosenburg (1986). The idea is to run this network backward so that it attempts to spell words based on their phonetic representations. This example further illustrates the use of this technique in a sequential interpretation setting in which phonemes are presented to the system one at a time and the system must refine its previous guess at the correct spelling as each new phoneme is presented.

INTRODUCTION

This paper explores the use of the technique of back-propagation of error (Rumelhart, Hinton, & Williams, 1986) to invert the mapping performed by a connectionist network. While the technique was introduced in that paper as a means of finding a set of weights which would achieve a certain mapping in a network of given topology, it is equally applicable to the problem of finding what input pattern would give rise to a specified output pattern in a network with given weights. Table 1 summarizes how it is possible to solve for any one of the three items *input pattern*, *weight matrix*, and *output pattern* given the other two, such that a network with that weight matrix maps that input pattern to that output pattern. The back-propagation input adjustment algorithm will be described further in

Table 1

Given	Solve For	Using
input pattern, weights	output pattern	forward propagation
input pattern, output pattern	weights	back-propagation learning
output pattern, weights	input pattern	back-propagation input adjustment

the next section, following which its use in inverting a particular network mapping will be explored.

Although no specific network implementation of the back-propagation mechanism is proposed here, these ideas hint at the intriguing possibility that there might exist a design for a network which contains within it the means for production, for comprehension, and for learning, all integrated together.

THE BACK-PROPAGATION INPUT ADJUSTMENT ALGORITHM

Just as the back-propagation learning algorithm (Rumelhart, Hinton, & Williams, 1986) is an incremental procedure for adjusting the weights in a network, the back-propagation input adjustment algorithm to be explored here is an incremental procedure. This means that it requires an initial "guess" at the input vector, which it successively modifies until the resulting input vector gives the desired output in the given network. The mathematical details of this algorithm, although straightforward, will not be given here, in the interest of brevity; suffice it to say that the algorithm moves down the negative of the gradient of the same squared-error performance measure used for the back-propagation learning algorithm. The difference is that this gradient is computed in the space of input vectors rather than in weight space. The derivation of this algorithm proceeds almost identically to the derivation of the learning algorithm, with the chain rule for partial derivatives giving rise to its back-propagation flavor, in which error-correction information is required to flow backwards along the connections in the network.

Just as in the learning case, back-propagation of error-correction information must be interspersed with forward propagation in the net to determine what additional adjustments are necessary. Thus the entire algorithm for inverting the network mapping for a particular specified output pattern consists of starting with an initial input pattern and modifying this pattern by repeated application of what will be called a *basic adjustment cycle*. Such a basic adjustment cycle consists of propagating activity forward in the network, back-propagating the error-correction information, and incrementing the input vector accordingly.

SOME SIMULATION RESULTS

The NETtalk text-to-speech network of Sejnowski and Rosenburg (1986) is a network whose input represents a seven-character window on a potentially much longer string of text and whose output represents the single phoneme which is appropriate for the character at the center of the window in the context of the remaining six characters. The network operates on an arbitrary-length text string by successively sliding its seven-character window along this string by one character at a time. Here we consider a drastically simplified version of such a net and study its ability to run "backwards" — i.e., its ability to spell a word given its phonetic representation. Furthermore, since the network represents only a single phoneme at a time, this problem will take on a sequential interpretation flavor: as successive phonemes are presented, the system will be forced to update its "preferred spelling" as these phonemes come along, rather than in parallel. It thus becomes interesting to examine the sequence of results obtained by the system.

The network used in these experiments is depicted in Figure 1. The input units encoded eight characters in each of three positions, with unused character/position combinations eliminated. The output units encoded nine phonemes. Both the input and the output representations were local rather than distributed. This was done strictly as a matter of convenience in setting up the network and also to make it easier to interpret arbitrary pattern vectors in a reasonable way, as will be discussed below.

The network was first trained (using back-propagation learning) to produce the appropriate phonemes for seven words. In the manner of NETtalk, each word was trained in each relevant position in the 3-character window. The training data is listed in Table 2. These words were chosen

Table 2

Input Word (3 positions)	Output Phonemes
can	/k/, /@/, /n/
car	/k/, /a/, /r/
con	/k/, /a/, /n/
wan	/w/, /a/, /n/
war	/w/, /c/, /r/
was	/w/, /^/, /z/
won	/w/, /^/, /n/

because of the interesting problems they present for a system trying to determine the spelling as the phonemes appear sequentially. Note that determination of the correct vowel cannot be made for some of these before the final consonant phoneme has been presented. Thus particular interest in these experiments was centered on the ability of the network to infer the vowel as each phoneme was presented.

Once the network had been trained to achieve essentially perfect performance, the weights were fixed; this network then formed the basis of the system on which all the experiments reported here were performed. In all these experiments, character positions to be determined had all their character units' outputs initialized to the same nominal value (typically .1). Also, for all the experiments the "solution" obtained by the system for any particular character position was interpreted to be that character having the largest output. This is consistent with the Sejnowski & Rosenburg interpretation of the "best guess" output of their system as the vector making the smallest angle with the output vector.

The results of the experiments are summarized in Table 3, where an asterisk is used to denote a character position which is to be determined. Experiments 4-7 involved presenting sequences of phonemes to the system, and these experiments were run as follows:

(1) The letter units were initialized to nominal values.
(2) The given phoneme was selected as the target output pattern to be achieved.
(3) The character units had their values adjusted via several iterations (typically 50, using a rate parameter of .1) of the basic adjustment cycle. At this point the target output was well matched by the output actually achieved using the adjusted input pattern.

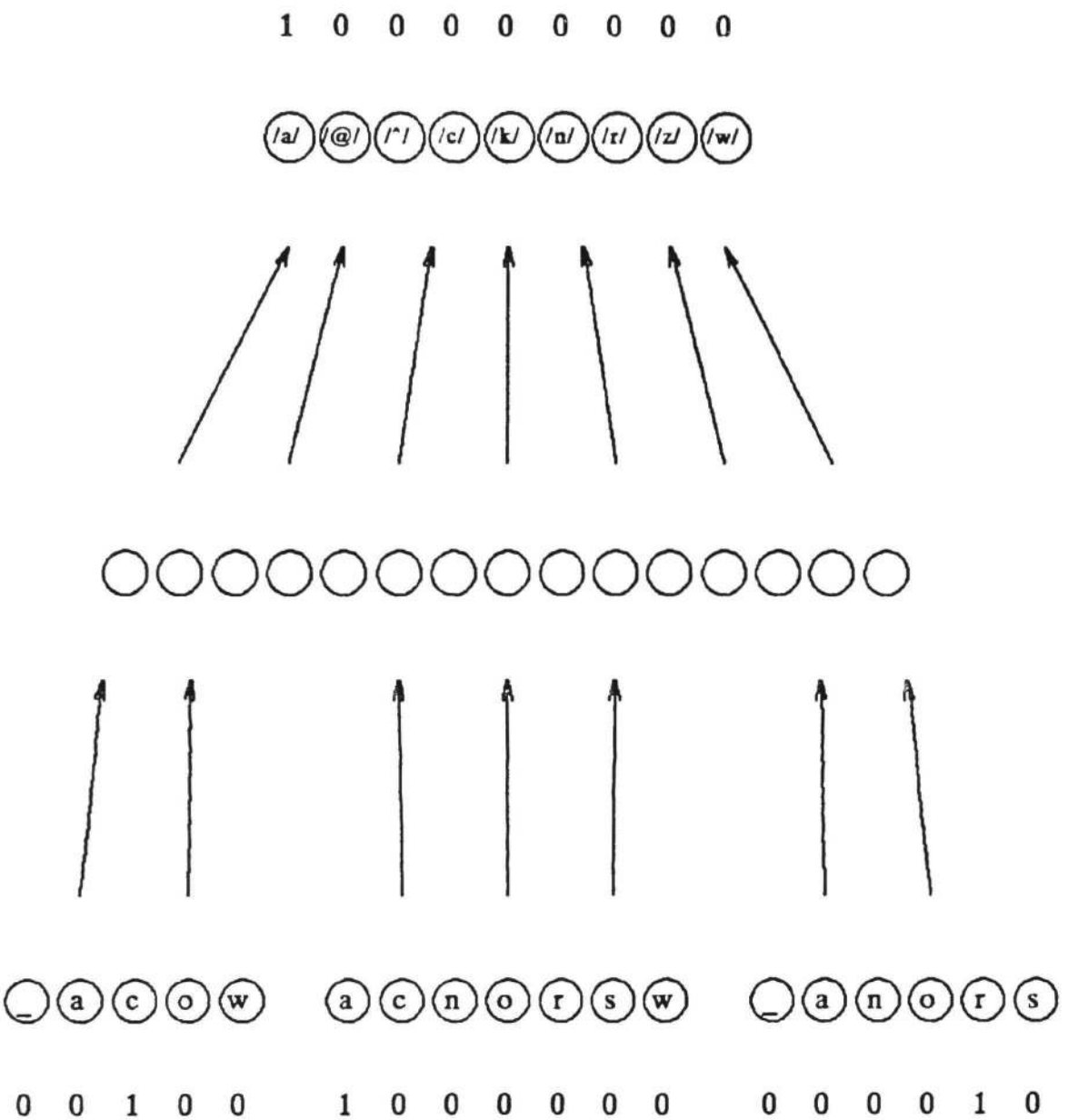


Figure 1. The network used in the experiments. The input layer is divided into three parts, one for each of three character positions. There is complete connectivity between layers. The input vector shown represents the word *car*. The corresponding output vector shown represents the appropriate phoneme /a/. The printed representation used here for these phonemes is taken directly from Sejnowski and Rosenburg (1986). The sounds they denote can be inferred from the description of the training data for the network given in Table 2. The underscore represents the <*space*> character.

(4) The output values of the character units were shifted one slot to the left in preparation for receipt of the next phoneme. This set of output values represented the starting point for the attempt to match the next phoneme. (The rightmost character slot had all its units' outputs set to the default nominal value.)

Steps (2)-(4) were repeated for each phoneme in the string of phonemes which the network was to spell. At each iteration of this cycle of steps the pattern of activity in the character units was examined at the end of step (3).

Table 3

Experiment	Target Phoneme	Character Units	
		Initial State	Winners After Convergence
1	/@/	•••	can
2	/a/	c•r	car
3	/a/	c•n	con
4	/w/ /^/ /n/	•••	_wa wos on_
5	/w/ /^/ /s/	•••	_wa wos as_
6	/k/ /a/ /r/	•••	_cr car ar_
7	/k/ /a/ /n/	•••	_cr cor on_

In every one of these sequential experiments there was ambiguity concerning the correct choice of vowel at the time that phoneme was presented; subsequent presentation of the next phoneme provided disambiguating information which enabled the system to choose the correct vowel even after the corresponding phoneme was no longer available to the system. In every case the system made the correct vowel its clear favorite once this disambiguating information was made available.

DISCUSSION

There are several remarks to be made here. First, the particular network used for these experiments was quite small and the inversion problems posed were quite simple. It will be interesting to see whether similar results are obtained if corresponding experiments are performed in a much larger network having many more input/output pairs "stored" in its weights.

Second, the problem of inverting a NETtalk-like grapheme-string-to-single-phoneme mapping was chosen because it illustrated not only the notion of inverting a mapping but also some other issues which are based on the observation that the network inversion problem and the network learning

problem are dual instances of the same general mathematical problem. As such, they are both underdetermined, in general, so that what must be sought are simultaneous solutions to multiple instances of such problems, or else solutions which are nearest, in some sense, to a given starting point.

Furthermore, the sequential nature of learning problems — in which the items to be learned are assumed to be experienced sequentially — is generally taken for granted, while the connectionist approach often suggests solutions to comprehension problems in which a great deal more parallelism is assumed (often by simply buffering a temporally extended input stream in order to make all components of it available for processing simultaneously). In the example considered here, a strictly sequential process was invoked in which the system was only allowed to examine one phoneme at a time. An interesting question is what to do at each step in order to get optimal convergence to the "correct" answer. The engineering topic of *recursive identification* (Ljung & Söderström, 1982) addresses such questions, although most such algorithms may be of limited applicability to these network problems since they are based on linear approximations. The importance of these sequential issues can be seen by noting that such algorithms applied to the learning problem would provide one-trial learning. While a parallel approach is possible even in the learning case, it is clearly inappropriate.[1]

Another remark concerns the implication of the duality between input vector and weight matrix for the connectionist approach. It is intriguing to speculate on how it might be possible to create models in which activations and weights play a more symmetric role. There are certainly precedents for such an enterprise. For example, models have been proposed which essentially replace weights by activation of units via *gating* connections on second-order *sigma-pi* units (Hinton, 1981; Rumelhart, Hinton, & McClelland, 1986; Williams, 1986). One such model is that of McClelland (1986). Also, some models have been studied which call for fast short-term changes in weights, which might be considered a means by which network weights are made to take on a role more like that of activation of units.

Finally, note that unlike most reports on connectionist-style research, this paper does not actually propose a network implementation of the computational technique suggested here. Rather, it suggests the utility of the back-propagation formalism in another setting besides that for which it was originally proposed. Results such as these suggest that treating back-propagation as a functional primitive in networks may lead to a number of elegant solutions to connectionist-style problems. It remains an open question just how such functionality may be implemented in more conventional forward-propagating network fashion. It is clear that the computation explored here has the flavor of a sequence of settlings, suggesting that a possible implementation might consist of a network designed to carry out this settling behavior.

---

1. In fact, it is not hard to devise a general procedure for constructing a larger network, with certain weights constrained to be equal, such that the sequential problem of determining the weights for a given collection of input/output pairs in a given network amounts to a single training instance in this new (possibly gigantic) network. Furthermore, the usual technique of interspersing the pattern presentations throughout training can be viewed as a time-shared, serial implementation of this parallel process.

REFERENCES

Hinton, G. E. (1981). A parallel computation that assigns canonical object-based frames of reference. *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, Vancouver, B.C., Canada, 683-685.

Ljung, L., & Söderström, T. (1982). *Theory and Practice of Recursive Identification*. Cambridge: MIT Press.

McClelland, J. L. (1986). The programmable blackboard model of reading. In: Rumelhart, D. E. & McClelland, J. L. (Eds.) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 2: Psychological and Biological Models*. Cambridge: MIT Press.

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). A general framework for parallel distributed processing. In: Rumelhart, D. E. & McClelland, J. L. (Eds.) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations*. Cambridge: MIT Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In: Rumelhart, D. E. & McClelland, J. L. (Eds.) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations*. Cambridge: MIT Press.

Sejnowski, T. J. & Rosenburg, C. R. (1986). *NETtalk: a parallel network that learns to read aloud*. Technical Report 86/01, Department of Electrical Engineering and Computer Science, Johns Hopkins University.

Williams, R. J. (1986). The logic of activation rules. In: Rumelhart, D. E. & McClelland, J. L. (Eds.) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations*. Cambridge: MIT Press.

COMPLEMENT SELECTION AND THE LEXICON IN JAPANESE

Toshiyuki Yamamoto
Department of Foreign Languages
West Virginia University

ABSTRACT

This study is on the extended line of Grimshaw 1979, which explains the complement selection in Japanese. By extending Grimshaw's analysis that the combination of predicates and their complements are explicable by imposing well-formedness conditions on two different levels of representation: one at the syntactic level; the other at the semantic level, the analysis given here which utilizes two semantic restrictive features under the semantic feature: [+presupposition] and [±factive], is able to explain the anomalies concerning the complementizer selection in Kuno 1973.

INTRODUCTION

As far as complement selection is concerned, Bresnan 1972 and Chomsky 1973 assumed that the selectional restrictions between verbs and types of complements are made solely on syntactic level. However, Grimshaw 1979 claims that the combination of predicates and their complements are explicable by imposing well-formedness condition on two different levels of representation, i.e., one at the syntactic level, and the other at the semantic level. That is, subcategorization gives restrictions between verbs and the type of complements, and the semantic selection restricts the combination between verbs and the semantic type of their complements. Grimshaw 1979 argues that it is not possible to reduce the two restrictions to either one of levels.

For ease of exposition, observe examples in (1).

(1) a. John wondered [who Bill saw].
    b. John wondered *[that Bill saw someone].
(2) a. John thought [that Bill saw someone].
    b. John thought *[who Bill saw].

The verb wonder takes only questions, but does not take that-complements. However, the verb think takes that-complements and not questions. At the syntactic level, what we need is for the verb wonder and the verb think to have the syntactic feature +[___$\overline{S}$]. At the semantic level, the verb wonder requires questions, that is, +[___Q], while the verb think requires propositions, that is, +[___P]. Figure 1 summarizes what has been discussed so far.

The plausibility of this analysis is attested by null complement anaphora such as the following:

(3) A: Did John leave?
    B: *I agree.
       I don't know.
(4) A: John is telling lies again.
    B: I agree.
       It's too bad.
       *I inquired.

In (3), B's answer is used in response to A's question. Therefore, in order for this discourse to be complete, B's answer

	syntactic level	semantic level
wonder	+[___$\overline{S}$]	+[___Q]
think	+[___$\overline{S}$]	+[___P]

Figure 1: **Well-formedness conditions of wonder and think**

requires a question as its interpretation. The ill-formed discourse will be ascribed to the fact that the verb agree does not take +[___Q] as its semantic feature. On the other hand, in (4), the proposition John is telling lies again occurs after I agree and It's too bad because they take the feature +[___P]. However, the ill-formedness of *I inquired is due to the fact that inquire only takes +[___Q]. In other words, using two levels of restriction makes it possible to conclude that a null complement anaphora has only to be controlled pragmatically and that there is no need for its antecedent. This is a much more elegant analysis compared to that of Hankamer and Sag 1976 in which they tried to explain the null complement anaphora as a kind of ellipsis. It will be enough to point out one example of Grimshaw 1979 to refute Hankamer and Sag 1976.

(5) John asked me the time.
=what time it was

(6) Bill asked me the time {(but) I didn't know. / so I inquired.}

(7) *I inquired the time.

The sentence (5) includes a concealed question. That is, the NP, the time is interpreted as what time it was. Therefore, the verb ask requires +[___Q]. In (6), the verb inquire requires only +[___$\bar{S}$] syntactically, as the ungrammaticality of (7) shows. However, sentence (6) is totally grammatical, even though a concealed question is not possible in the case of the verb inquire, as in (7). The grammaticality of (6) is attributed to the fact that both verbs ask and inquire require +[___Q] semantically.

COMPLEMENTIZERS IN JAPANESE

In the Japanese language, there are at least four complementizers (koto, no, to, ka). Unlike English complementizers, Japanese complementizers occur at the rightmost end of $\bar{S}$ as the following sentence shows:

(8) [[Mary-ga Jack-o aishite-iru $_S$][koto] $_{\overline{S}}$]-wa
SUB OBJ loves COMP SUB

shuuchi-no jijitsu da.
well-known fact is

'It is a well-known fact that Mary loves Jack.'

Following the same line of argument presented in Grimshaw 1979, let us assume that at the syntactic level, verbs or adjectives only have such subcategorization features as +[$\overline{S}$___] or +[$\overline{N}$___], and at the semantic level, verbs or adjectives specify semantic features like P or Q. However, we still have such ungrammatical sentences as follows:[1]

(9) Watashi-wa [[John-ga Mary-o butsu][{*koto / no / *to}]]-o
I SUB SUB OBJ hit COMP OBJ

mita.
saw

'I saw John hit Mary.'

(10) Watashi-wa [[nihongo-ga muzukashii][{koto / *no / *to}]]-o
I SUB Japanese SUB difficult COMP OBJ

mananda.
learned

'I learned that Japanese is difficult.'

(11) John-wa [[nihongo-ga muzukashii][{*koto / *no / to}]] itta.[2]
SUB Japanese SUB difficult COMP said

'John said that Japanese is difficult.'

The ungrammatical sentences suggest that we need further specifications along the line of analysis given by Grimshaw 1979. That is, the semantic feature +[P___] has to be more specific to restrict complementizer selections. Kuno 1973 gives partial analy-

[1]Data from (8) through (26) except (20) are from Kuno 1973.

[2]The particle o is deleted obligatory when preceded by the complementizer to.

sis concerning this topic. However, because he differenciates verbs according to whether verbs include presupposition or not, some listings of verbs are left unexplained as to why they take only one complement rather than the others.

The claim to be made under the analysis given here is that complementizer selections of verbs can be explicable by using two semantic restrictive features under the semantic feature +[P___], namely [±presupposition] and [±factive]. The feature [+presupposition] means that the verb requires P with a presupposition. The feature [+factive] means that P is based on fact. The whole combination is summarized in Figure 2.

In Kuno 1973, the verb omow-u 'think' has the semantic feature [-presupposition] and the verb wasure-ru 'forget' has the feature [+presupposition]. These carry the grammaticality and the ungrammaticality of (12) and (13).

(12) John-wa [Mary-ga baka da {*koto / to (COMP)}]-o omotta.
SUB SUB stupid is OBJ thought

'John thought that Mary was stupid.'

(13) John-wa [Mary-ga tunbo de aru {koto / no / *to (COMP)}]-o
SUB SUB deaf is OBJ

wasurete-ita.
forgot

'John forgot that Mary was deaf.'

Complementizer	+[P___]	
	Presupposition	Factive
no	+	+
koto	+	-
to	-	+
	-	-

Figure 2: Semantic restrictive features under +[P___]

Verbs like haya-gaten-s-uru 'form a hasty conclusion', iw-u 'say', kanchigai-s-uru 'make a wrong guess', and gokai-s-uru 'form the wrong notion' have the semantic feature [-presupposition]. These verbs take only to clause. Observe the following examples.

(14) John-wa [Mary-ga shinda {to / *koto-o / *no -o}]
SUB SUB died COMP OBJ

{haya-gaten-shita. / formed-a-hasty-conclusion
itta. / said
kanchigai-shita. / made-a-wrong-guess
gokai-shita. / formed-the-wrong-notion}

'John {formed the hasty conclusion / said / made the wrong guess / formed the wrong notion} that Mary had died.'

The following verbs take only the [+presupposition] feature, which tells that the complementizer to is not used.

(15) John-wa [Mary-ga tsunbo de aru {koto / no / *to}]-o
SUB SUB deaf is COMP OBJ

{omoidashita. / recalled
wasurete-ita. / had-forgotten}

'John had forgotten that Mary was deaf.'

So far, Kuno 1973 and the analysis given here predict the same data. Kuno 1973 sets the verb shiru 'know' as an exception and lists the following examples.

(16) *Anata-wa [Mary-ga tunbo da to] shitte-imasuka?
you SUB SUB deaf is COMP know

'Do you know that Mary is deaf?'

(17) Watashi-wa [Mary-ga tsunbo da to] sonotoki
I SUB SUB deaf is COMP then

shitta.
got-to-know

'I got to know then that Mary was deaf.'

(18) *[Mary-ga konna baka da to] shitte-imashitaka?
SUB such-a fool is COMP know

'Do you know that Mary was such a fool?'

(19) [Mary-ga konna baka da to] shirimasen deshita.
SUB such-a fool is COMP know not past

'I did not know that Mary was such a fool.'

However, a close examination shows that shir-u, which is the present form, means 'come to the state of knowing' or 'get to know', and the past tense shitta means 'came to the state of knowing' or 'got to know'. Shirimasen, which is the negative counterpart of shir-u, means 'not come to the state of knowing' or 'not get to know'. On the other hand, shitte-iru means 'be in the state of knowing'. That is, Kuno 1973 fails to assume that the verb shitte-iru is a stative verb unlike the verb shir-u, which is an action verb. The semantic restrictions of the verbs shir-u and shitte-iru will be as follows:

(20) shir-u: [±presupposition, ±factive]
shitte-iru: [+presupposition, ±factive]

With (20), we can now clearly explain the grammaticality from (16) through (19).

By extending the semantic restrictions, the analysis given here can now explain anomalies in Kuno's paradigm. First, verbs of perception can take only a no clause.

(21) Watashi-wa [John-ga Mary-o butsu {no / *koto}]-o mita.
I SUB SUB OBJ hit COMP OBJ saw
'I saw John hitting Mary.'

(22) Watashi-wa [John-ga piano-o hiku {no / *koto}]-o kiita.
I SUB SUB OBJ play COMP OBJ heard
'I heard John playing the piano.'

(23) Watashi-wa [sesuji-ga samuku naru {no / *koto}]-o
I SUB spine SUB cold become COMP OBJ
kanjita.
felt
'I felt a cold shiver running down my spine.'

The analysis given here predicts that the verbs in (21)-(23) have the semantic feature, +[P___], [+presupposition, +factive]. This feature specification clearly distinguish no and koto.

Second, verbs of ordering are specified by the semantic feature, [+presupposition, -factive], which restricts the complementizer to only koto.

(24) Watashi-wa John-ni [hataraku {koto / *no / *to}]-o
I SUB to work COMP OBJ
{yookyuu-shita. (demanded) / tanonda. (asked) / kyoosei-shita. (forced)}
'I {demanded / asked / forced} John to work.'

Third, verbs of expecting have implication that things to expect have not yet come true. Therefore, the feature [-factive] is assigned to these types of verbs. To put it more precisely, the semantic restriction is +[P___], [±presupposition, -factive], which allow complementizers koto and to.

(25) Mary-wa [John-ga kuru {koto-o / to}] kitai-shite-ita.
SUB SUB come COMP OBJ was-expecting
'Mary was expecting that John would come.'

Finally, the verbs of waiting take no and koto. This seems to be a natural consequence because there have to be things to wait for when one waits for something. Therefore, the feature [+presupposition] is marked at the semantic level. Whether the complementizer no or koto is specified is based on the things to wait for. The following examples show that the highly abstract concept is marked [-factive], while waiting for John to come is less abstract, and rather more factive because the reason that one can wait for John to come is due to the fact that John is coming.

(26) a. Watashi-wa [John-ga kuru {no / *koto}]-o matta.
I SUB SUB come COMP OBJ waited
'I waited for John to come.'

b. Watashi-wa [sekai-ni heiwa-ga otozureru {no / koto}]
I SUB world to peace SUB visits COMP
-o matte-imasu.
OBJ am-waiting
'I am waiting for peace to descend on the world.'

CONCLUSION

By subcategorizing the semantic feature +[P___] into the combinations of [±presupposition] and [±factive], Kuno's anomalies come under the regular pattern. It seems that the solution we have made argues for non-autonomous syntactic hypothesis. However, the point to be made is that assuming syntax and semantics are two different autonomous systems and that each module serves to generate grammatical sentences, we will get higher generalizations in rules of grammar.

REFERENCES

Akmajian, A. (1977), "The Complement Structure of Perception Verbs in an Autonomous Syntax Framework," in Culicover, et. al., Formal Syntax., Academic Press, New York, 427-460.

Bresnan, J. (1972), Theory of Complementation in English Syntax, Doctoral Dissertation, MIT, Cambridge, Massachusetts.

Chomsky, N. (1965), Aspects of the Theory of Syntax, MIT Press, Cambridge, Massachusetts.

----------. (1973), "Conditions on Transformations," in S. R. Anderson and P. Kiparsky, eds., A Festschrift for Morris Halle, Holt, Rinehart and Winston, New York, 232-286.

Grimshaw, J. (1979), "Complement Selection and the Lexicon," Linguistic Inquiry, vol. 10: 2. 279-326.

Grosu, A. & A. Thompson. (1977), "Constraints on the Distribution of NP Clauses," Language 53: 1. 104-51.

Hankamer, J. & I. Sag. (1976), "Deep and Surface Anaphora," Linguistic Inquiry 7, 391-428.

Katz, J. & P. Postal. (1964), An Integrated Theory of Linguistic Descriptions, MIT Press, Cambridge, Massachusetts.

Kiparsky, P. & C. Kiparsky. (1970), "Fact," in M. Bierwisch & K. E. Heidolph, eds., Progress in Linguistics, Mouton, The Hague, 143-173.

Kuno, S. (1973), The Structure of the Japanese Language, MIT Press, Cambridge, Massachusetts.

Lightfoot, D. (1977), "Syntactic Change and the Autonomy Thesis," Journal of Linguistics 13: 191-215.

Newmeyer, F. J. (1983), Grammatical Theory, Its Limits and Its Possibilities, The University of Chicago Press, Chicago.

Williams, E. (1976), "Abstract Triggers," ms., University of Massachusetts at Amherst, Amherst, Massachusetts, to appear in the Journal of Linguistic Research.

I, me, mine (1)
Psycholinguistic Constraints of French Clitics in Sentence Generation

Michael ZOCK
Gérard SABAH

LIMSI, Langues Naturelles
B.P. 30 - 91406 ORSAY Cédex/France

Abstract :

This paper describes an implemented tutoring system, designed to show various ways of converting a given meaning structure into its corresponding surface expression. The system is meant to be a teaching tool for students who learn French as a foreign language.

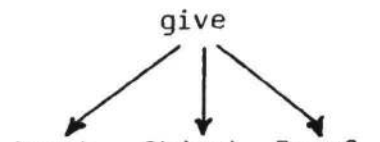
While showing various ways of converting a given meaning structure into its corresponding surface expression, the system helps not only to discover WHAT data to process but also HOW this information processing should take place. In other words, we are concerned with efficiency in verbal planning (flexibility and economy of performance).

Recognizing that the same result can be obtained by various methods, the student should find out which one is best suited to the circumstances (what is known, task demands etc). Informational states, hence the processor's needs, may vary to a great extent, as may his STRATEGIES or cognitive styles. In consequence, in order to become an efficient processor, the student has to acquire not only STRUCTURAL or RULE-KNOWLEDGE but also PROCEDURAL-KNOWLEDGE (skill).

With this in mind we have designed three modules in order to foster a reflective, experimental attitude in the learner, helping him to discover insightfully the most efficient strategy.

1 THE PROBLEM :

It is well known that children or students learning French as a foreign language need a great deal of practice before they master the French pronoun system well enough to fluently produce correct sentences with 1 or 2 pronoun complements, such as:



2 3 1

a) Tu me le donnes? S-IO-DO-V
Do you give it to me ?

b)Donne-le moi ! V-DO-IO
Give it to me !

c) Ne me le donnes pas! neg-IO-DO-V-neg
Dont' give it to me

The student's problem can be stated in the following terms: he has to learn how to determine both form and position for the French pronouns. Basically he is faced with the following problem:

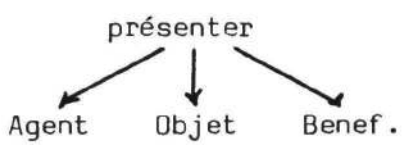
- he has to LEARN two lists, one for morphology the other for syntax (frames)
- he has to DISCOVER under what conditions each of these elements applies.

List of MORPHOLOGY		list of SYNT. FRAMES
SPEAKER:	je, me, moi, nous	S-DO-IO-V
LISTENER:	tu, te, toi, vous	S-IO-DO-V
3d PERSON:	le, la, les, lui, leur, eux	S-DO-V-prep-IO
	il, elle, ils, elles, on, se soi	etc.

As we will see, the student's task is not all that easy. These are some of the reasons why:

A) SYNTAX depends upon MORPHOLOGY:

By modifying the form of a given element one may also alter its position. In other words, certain features influence simultaneously morphology and syntax (see d-e).



3 1 3 d) Il me présente à elle S-DO-V-prep-IO
He presents me to her

3 3 3 e) Il le lui présente S-DO-IO-V
He presents him to her

B) The determination of FORM and POSITION requires complex FEATURE-CLUSTERS:

Any form or sentence-frame depends upon a number of conditions or features (2). Even a simple concept such as SPEAKER requires quite a lot of proces-

sing. It may be expressed by a variety of forms (je,me,moi); each may affect the choice of a sentence frame.

Furthermore, features such as PERSON, NUMBER, CASE and GENDER (3) which come readily into mind, are by no means sufficient. As the examples "a-c" clearly show, many other variables like SENTENCE-MODE (a,b), NEGATION (b,c), etc. come into play.

C) Some MORPHEMES are MULTIPLE DEPENDANTS (4):

The form of some morphemes depends not only upon features inherent in the coreferent (vertical dependancy), but also upon features coming from another referent (horizontal dependancy). This is the case of the indirect object, whose form depends upon the value of the direct object (see d,e). This fact is interesting for its procedural implications, namely it excludes any word-to-word processing.

In the light of these facts, one must admit, that what looked easy at first sight turned out to be a complex enterprise.

2 OBJECTIVE :

The system described here (5) is an attempt to help the student to acquire the necessary structural and procedural knowledge in order to economically generate pronoun-constructions in French (6).

While converting a given meaning structure into its corresponding surface expression, the student should not only learn WHAT data to process, but also HOW this information processing should take place. Recognizing that the same result can be obtained by various methods (strategies), the student should find out which one is best suited to the circumstances.

Particular emphasis is placed upon the discovery of operating principles (7) and the building of larger blocks (schematas). This chunking method should not only help to avoid unnecessary disruptions and memory load but also allow evolution from serial to simultaneous processing.

3 DESCRIPTION OF THE SYSTEM :

The heart of the system is a knowledge base which contains, in the form of production rules, the structural information necessary to incrementally determine form as well as position. Furthermore the system contains an inference-mechanism, i.e. a set of rules, whose function is to deduce new facts from information given to the system.

The base can be accessed in various ways, thus allowing for varying usage of the knowledge according to the objectives.

3.1 THE SOCRATIC METHOD :

The system guides the student in the form of a dialogue, by showing him what and how to process in order to get from an input to the output. The user starts by providing the input (verb-pattern composed of a verb, its complements and eventual prepositions):

donner (qn,qc,à qn) give (so, sth, to so)

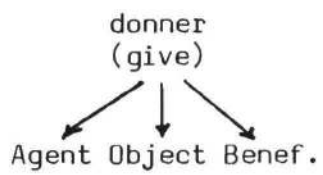
Then the system takes over, asking for more information about these basic elements. By asking specific questions (person, gender, number), the system shows which information is relevant when determining form as well as position. While answering these questions the student incrementally determines the final form of the sentence.

3.2 GUIDED DISCOVERY :

The system still controls the nature of the operations but no longer controls their order. The latter is controlled, via strategies, by the user. He decides in what order to process the data. Having determined the subject, whose position is invariable, one can choose from three strategies:

- a syntactical one (syntactic-driven processing)
- and two morphological ones (lexical-driven processing).

If priority is given to syntax (strategy 1), no further reordering of syntactic constituents is meant to take place, i.e. all information necessary to determine word order will have been processed. The result is an ordered categorial structure or syntactical frame (T1) which will be filled in by the morphological values determined later (T2), for example:



(T1) sentence frame :Subject - Dir.Obj.- Ind.Obj.- Verb
(T2) morphology : il - le - lui - donne

3 3 3 (he gives it to her)

If priority is given to morphology (lexically driven generation), the form is determined before the relative order of the constituent elements. In this case two strategies are possible: either one processes the direct (strategy 2), or the indirect object (strategy 3).

This experimental method should make the student aware of the fact that there are several ways of arriving at a particular solution (sentence).It is precisely his task to find out which strategy is the best suited. By applying performance criteria such as:

- number of steps necessary to generate the sentence

- what is known when? (form/position)
- congruence of input/output order (are permutations necessary? LIFO/FIFO)
- are there any conceptual disruptions?

the student hopefully discovers the procedural knowledge necessary for economic production of pronoun constructions.

3.3 USER DRIVEN EXPERIMENTATION :

This method, like the previous one, is empirical. By playing with the system the student may gain certain insights about processing order.

A matrix appears on the screen, whose blank spaces have to be filled in by the student. The horizontal line shows the syntactic information given with the input (verb, subject, object, preposition), -more information is needed about those elements- the vertical line shows the nature of the information necessary to arrive at the output.

Thus the processing once again consists of the specification of the values of a list of attributes. However there is a fundamental difference between this approach and the former, namely that the system has an inference mechanism. Each item of information given to the system is considered for its meaning potential, i.e. the system tries to find out whether some new facts can be deduced from the old fact.

It should be noted that the inference power varies with the nature of the data as well as with their order. There are cases where a single fact enables 3 other facts (reflexives) to be deduced. A given inference may allow further deductions (inference-chain/knowledge propagation). This has of course an effect on the process, namely the greater the inference power, the greater the economy of processing. This suggests the following operating principle:

> the greater the inference power of a given piece of information, the earlier it should be processed.

This method is interesting in that, by testing different items and different orders, it makes it possible to see on the screen which items allow what inferences. Since those inferences depend upon the nature of the input as well as on the moment at which that information is given, we believe that this system is particularly useful in helping to discover the optimal order of processing.

Furthermore we think that this method has another virtue, namely that it can simulate literally any knowledge state, thus making it possible, by experimental means to discover the shortest path between a given informational state (input) and the solution (output).

4 CONCLUSIONS :

We have stressed the need for teaching procedural knowledge (strategies) as well as structural knowledge (linguistic rules). In order to achieve this goal we have designed three modules, one demonstrative, the other two experimental.

Each module is intended for a different learning stage or learner type, as the cognitive styles may vary both among individuals and within the same individual. By progressively moving the control from outside (system) to inside (student), i.e. in integrating the student into the learning-process (8) we hope to make him:

- actively curious (testing of hypotheses - learning by discovery);
- conscious of the need for planning (how far should one plan ahead?),
- selective about the means he should use (Which strategy is best under what circumstances ?).

The whole idea of having different strategies compete has been largely ignored by current work on language generation. While this aspect may be only of secondary interest for automatic generation in general, it certainly is not an unimportant issue in cognitive modelling, whether it be language learning or usage.

NOTES :

1° In memory of one of the Beatle's songs.

2° The average number for a given pronoun is about three.

3° We call these inherent features, as opposed to features like sentence-mode, negation etc. which also determine form but whose information is not explicited in the morpheme.

Note that only PERSON and NUMBER are inherent and well marked in all pronouns. A feature like GENDER is only present in some forms (mostly 3d person), whereas the feature CASE is ambiguous for many direct and indirect object pronouns, in particular for 1st and 2nd person.

4° The way we use the term dependancy here is not to be confounded with Tesnière's dependancy theory (1959). Tesnière uses the word dependacy to signify "valency" whereas we use the word in its literal meaning.

5° The modules described are written in Simula and Prolog. They were implemented by G.Sabah and C.Alviset.

6° By economy we mean: number of operations, necessary permutations and the number of items to be stored.

7° Among those operating principles are the following:

- avoid disruptions by grouping together what belongs conceptually together;
- start with the most informative items
  (feature hierarchy: PERSON > CASE > NUMBER > GENDER);
- avoid unnecessary storage - start with the leftmost item.

8° The system could be greatly improved if it contained a module, capable to analyse the students performances (see Woolf & McDonald).

REFERENCES :

L.Tesnière
1959 Eléments de Syntaxe Structurale, Klinksieck, Paris

B.Woolf & D. McDonald
1984 Context-dependent Transitions in Tutoring Discourse
Proceedings of AAAI

In the best tradition of cognitive science inquiry – Academic Press.

Bringing you important studies that make a difference.

Cognitive Sciences: Basic Problems, New Perspectives, and Implications for Artificial Intelligence

MARIA NOWAKOWSKA

The topics covered in this highly original work range from such relatively unexplored ones as observability networks and their restrictions, random and semi-random fuzzy objects, and relations between subjective and objective time, through judgement formation, descriptions, and texts, to more traditional subjects studied by cognitive scientists, namely, visual perception, memory, and communication. In her treatment of each topic, the author develops new conceptual representations of the phenomenon analyzed, postulates appropriate mechanisms, and formulates explanatory hypotheses.

Dr. Nowakowska treats cognitive processes through a set of interrelated theories, from different perspectives and on various levels of generality. She departs from the information processing paradigm and proposes new concepts and solutions based on psychological intuitions and mechanisms. Her work stresses the problems of cognitive limitations and distortions as well as the generative, controlled, and dynamic character of cognition.

In Paperback
1986, 400 pp., $29.95/ISBN: 0-12-522621-7
$59.50/ISBN: 0-12-522620-9 — Casebound

Pattern Recognition by Humans and Machines

Edited by **EILEEN C. SCHWAB and HOWARD C. NUSBAUM**

Volume 1 delves into • basic theoretical issues in speech perception • perception of phenomes and words • the role of attention in speech processing • computation approaches to modeling speech perception • general computational approaches in language processing. **Volume 2** highlights • basic theoretical issues in form perception and feature integration • perception and psychophysics of form recognition • the role of attention of form perception • computational approaches to modeling recognition and parsing of visual form • general computational concerns in modeling perception. Both **Volumes 1 and 2** are in the *Academic Press Series in Cognition and Perception.*

Volume 1
Speech Perception
In Paperback
June 1986, 336 pp.
$29.95/ISBN: 0-12-631403-9
$59.50*/ISBN: 0-12-631401-2 — Casebound

Volume 2
Visual Perception
In Paperback
July 1986, 264 pp.
$23.95/ISBN: 0-12-631404-7
$46.50*/ISBN: 0-12-631402-0 — Casebound

SAVE over $16.00 by purchasing the entire casebound set! (Set price: $89.50)

Reasoning and Discourse Processes

Edited by **TERRY MYERS, KEITH BROWN, and BRENDAN MCGONIGLE**

This book reflects the contemporary effort to unite formal and psychological approaches to the understanding of the mind of the a language user engaged in discourse and reasoning.

In Paperback
May 1986, 312 pp.
$29.95/ISBN: 0-12-512321-3
$59.00/ISBN: 0-12-512320-5 — Casebound
A Volume in the COGNITIVE SCIENCE Series

Complete with a cassette tape of illustrative examples...

Music Cognition

W. JAY DOWLING and DANE L. HARWOOD
1986, 258 pp., $35.00/ISBN: 0-12-221430-7
A Volume in the ACADEMIC PRESS SERIES IN COGNITION AND PERCEPTION

Musical Structure and Cognition

Edited by **PETER HOWELL, IAN CROSS, and ROBERT WEST**
1985, 338 pp., $57.50/ISBN: 0-12-357170-7

Academic Press Journals...

Journal of Memory and Language

Formerly Journal of Verbal Learning and Verbal Behavior
Editors: **MARCEL ADAM JUST** and **PATRICIA A. CARPENTER**
Volume 25 (1986), 6 issues (including annual subject index)
ISSN: 0749-596X
***Annual subscription rates:**
In the U.S.A. and Canada: $113.00
All other countries: $130.00

International Journal of Man–Machine Studies

Executive Editors: **B. R. GAINES** and **D. R. HILL**
Volumes 24–25 (1986), 12 issues
ISSN: 0020-7373
***Annual subscription rate: $337.00**

Cognitive Psychology

Editor: **EARL HUNT**
Volume 18 (1986), 4 issues
ISSN: 0010-0285
***Annual subscription rates:**
In the U.S.A. and Canada: $99.00
All other countries: $116.50

Journal of Mathematical Psychology

Editor: **JAMES T. TOWNSEND**
Volume 30 (1986), 4 issues
ISSN: 0022-2496
***Annual subscription rates:**
In the U.S.A. and Canada: $127.00
All other countries: $149.00

Computer Speech and Language

Editors: **FRANK FALLSIDE** and **STEPHEN E. LEVINSON**
Volume 1 (1986), 4 issues
ISSN: 0885-2308
***Annual subscription rate: $97.50**

Journal subscriptions are for the calendar year and are payable in advance. Please send all orders and inquiries to: Academic Press, Journal Promotion Department, 1250 Sixth Avenue, San Diego, CA 92101.

For more information on these titles, contact Karen Brogan, Product Manager, Academic Press, Marketing Department, Orlando, FL 32887 *or* call (305) 345-2746.

ACADEMIC PRESS
Harcourt Brace Jovanovich, Publishers
Orlando, Florida 32887-0405

1-800-321-5068
To Place An Order From Florida, Hawaii, Or Alaska CALL 1-305-345-4100.

Orlando San Diego New York Austin Boston London Sydney Tokyo Toronto

Prices are subject to change without notice. 12086 HR/IY

Books that Point the Way

Advanced Computer Science and Artificial Intelligence

Morgan Kaufmann Publishers, Inc. specializes in upper division and graduate texts, research and reference works, and professional conference proceedings in computer science. With an emphasis on innovative books of the highest quality, we maintain rigorous standards of selection and review and utilize modern publication methods to bring you the finest in advanced computer science and artificial intelligence.

AAAI and IJCAI Conference Proceedings

new AAAI-86
Philadelphia, Pennsylvania
The proceedings of the Fifth National Conference on Artificial Intelligence presents important, current work in the field. For the first time, the proceedings have been expanded to include two major categories of papers. The science track presents fundamental research results; the engineering track includes the best work in applied AI. All 188 papers have been refereed, and many include illustrations, references, and other useful resources.
☐ Paper; 2 volumes. ISBN 0-934613-13-3; $55 ($40 AAAI members)

AAAI-84
Austin, Texas
☐ Paper; ISBN 0-86576-084-3; $45 ($30 AAAI members)

AAAI-83
Washington, D.C.
☐ Paper, ISBN 0-86576-083-3; $45 ($30 AAAI members)

AAAI-82
Pittsburgh, Pennsylvania
☐ Paper; ISBN 0-86576-043-8; $45 ($30 AAAI members)

AAAI-80
Stanford, California
☐ Paper; ISBN 0-86576-080-1; $40 ($25 AAAI members)

new AAAI Two-Year Set
1984 and 1986
☐ Paper, 2 volumes; ISBN 0-934613-25-7; $70 ($50 AAAI members)

IJCAI-85
Los Angeles, California
The proceedings of the Ninth International Joint Conference on Artificial Intelligence presents leading edge research in such areas as AI and education, automated reasoning, expert systems, knowledge representation, logic programming, natural language, and robotics.
☐ Paper, two volumes; ISBN 0-934613-02-8; $55 ($40 AAAI members)

IJCAI-83
Karlsruhe, West Germany
☐ Paper, two volumes; ISBN 0-86576-83-8; $50 ($35 AAAI members)

IJCAI-81
Vancouver, B.C.
☐ Paper, two volumes; ISBN 0-86576-81-7; $40 ($25 AAAI members)

IJCAI-79
Tokyo, Japan
☐ Paper, two volumes; ISBN 0-86576-79-6; $40 ($25 AAAI members)

IJCAI-77
Cambridge, Massachusetts
☐ Paper, two volumes; ISBN 0-86576-77-5; $40 ($25 AAAI members)

IJCAI-75
Tbilisi, Georgia USSR
☐ Paper; ISBN 0-86576-75-4; $65 ($40 AAAI members)

IJCAI-73
Stanford, California
☐ Paper, ISBN 0-86576-73-3; $65 ($40 AAAI members)

IJCAI-71
London, England
☐ Paper, ISBN 0-86576-71-2; $80 ($50 AAAI members)

IJCAI-69
Washington D.C.
☐ Paper; ISBN 0-86576-69-1; $65 ($50 AAAI members)

AAAI/IJCAI Eleven-Year Set 1969–84
☐ Paper, 16 volumes; ISBN 0-934613-23-0; $525 ($340 AAAI members)

ORDER FORM — *Morgan Kaufmann Publishers, Inc.*

Order Fulfillment Center, Dept. 25, P.O. Box 50490, Palo Alto, CA 94303

Please indicate your order below. Enclose full payment plus $1.50 for the first volume and $.50 for each additional volume for book rate U.S. shipment. (For surface shipment out of the country enclose full payment plus $2.00 for the first volume and $.75 per each additional.) California residents please add sales tax. If not satisfied, you may return books within 30 days for a complete refund. (Software is non-returnable.)

QTY TITLE PRICE

☐ I am a member of AAAI

SHIPPING _____

TAX _____

TOTAL ENCLOSED _____

Name
Address
City
State
Zip

Morgan Kaufmann Publishers, Inc.

Research Notes in Artificial Intelligence

Copublished with Pitman Publishing Ltd.

The Research Notes series consists of technical monographs in specialized areas of AI research for use in graduate courses or as professional references. Manuscripts submitted for the series are evaluated by an international editorial board with Derek Sleeman of Stanford and NS Sridharan of BBN as the main editors.

A Critiquing Approach to Expert Computer Advice: ATTENDING By Perry L. Miller
Paper, ISBN 0-273-08665-0; $19.95

Heuristic Reasoning About Uncertainty: An Artificial Intelligence Approach
By Paul R. Cohen
Paper, ISBN 0-273-08667-7, $22.95

Searching With Probabilities
By Andrew Palay
Paper, ISBN 0-273-08864-2; $22.95

Knowledge-Based Interpretation of Outdoor Natural Color Scenes
By Yuichi Ohta
Paper, ISBN 0-273-08673-1, $19.95

Learning to Solve Problems by Searching for Macro-Operators By Richard Korf
Paper, ISBN 0-273-08690-1, $19.95

Empirical Analysis for Expert Systems
By Peter Politakis
Paper, ISBN 0-273-08663-4, $22.95

The Mathematics of Inheritance Systems
By David Touretzky
Paper, ISBN 0-934613-06-0; $22.95

Shape From Texture By John Kender
Paper, ISBN 0-934613-05-2; $22.95 (Aug. '86)

A Deduction Model of Belief
By Kurt Konolige
Paper, ISBN 0-934613-08-7, $24.95 (Aug. '86)

Representing and Acquiring Geographic Knowledge By Ernest Davis
Paper; ISBN 0-934613-22-2; $22.95 (Aug. '86)

Design Problem Solving: Knowledge Structures and Control Strategies
By David Brown and B. Chandrasekaran
Paper, ISBN 0-943613-07-9, $22.95 (Nov. '86)

Intention-Based Diagnosis of Errors in Novice Programs By Lewis Johnson
Paper, ISBN 0-934613-19-2, $24.95 (Fall 86)

Essential Books

new **Readings in Software Engineering and Artificial Intelligence** Edited by Charles Rich and Richard C. Waters
A collection of important articles dealing with issues at the intersection of software engineering and artificial intelligence. The editors provide an overview, introductions to each section, and a complete bibliography. Topics include program verification, transformational approaches, very high level languages, intelligent assistants, programming knowledge, and artificial intelligence programming.
☐ Paper; ISBN 0-934613-12-5;
(August 1986, Prepublication Price $26.95)

new **Readings in Natural Language Processing** Edited by Barbara J. Grosz, Karen Sparck Jones, and Bonnie Lynn Webber
Focusing on the central issues of natural language processing (representation, reasoning, and recognition), these articles were selected for the importance of their problem analyses and proposed solutions. Six sections, introduced by the editors, cover parsing and grammars, semantic interpretation, discourse interpretation, language actions and intentions, language generation, and systems. A valuable collection for students and researchers in AI, NLP, and linguistics.
☐ Paper; ISBN 0-934613-11-7;
(August 1986, Prepublication Price $26.95)

Machine Learning: An Artificial Intelligence Approach, Vol. II Edited by Ryszard Michalski, Jaime Carbonell, and Tom Mitchell
Recent advances in machine learning since the publication of Volume I. Reflects the major expansion of the field through presentation of a wide range of results by leading authorities.
☐ Cloth; ISBN 0-934613-00-1; $39.95

Readings in Knowledge Representation
Edited by Ronald Brachman and Hector Levesque
A carefully selected collection of seminal and recent papers for text and reference use. Sections include associational, structured object, logic-based, procedural, and common sense knowledge representations. Introductions to the papers and an extensive bibliography are provided.
☐ Paper; ISBN 0-934613-01-X; $26.95

Machine Learning: An Artificial Intelligence Approach
Edited by Ryszard Michalski, Jaime Carbonell, and Tom Mitchell
A collection of original articles by leading scientists on contemporary research in machine learning.
☐ Cloth; ISBN 0-934613-09-5; $39.95

Principles of Artificial Intelligence
By Nils J. Nilsson
This classic text lucidly explains the core ideas and fundamental principles of artificial intelligence, filling the gap between theory and practice.
☐ Cloth; ISBN 0-934613-10-9; $32.95

Readings in Artificial Intelligence
Edited by Bonnie Webber and Nils J. Nilsson
Thirty-one important papers on search, planning, expert systems, and other topics are gathered in this convenient volume for class and reference use.
☐ Paper; ISBN 0-934613-03-6; $26.95

Software

PCTEX™ A full implementation of Donald Knuth's TEX™ for the IBM PC/XT or AT™ and compatibles, PCTEX allows typesetting of books, technical articles, documentation, and correspondence on the personal computer. The basic PCTEX package includes three macro packages: PLAIN for beginning and intermediate users, LATEX for document preparation, and AMSTEX for mathematical material. INITEX, a system enhancer, font metric files, 5 DS/DD diskettes, and the PCTEX manual complete the package.
Device drivers required to interpret PCTEX commands for finished output in any of over 200 fonts. *PCDOT* drives numerous dot-matrix printers by IBM, Epson, and Toshiba; *PCLaser* drives laser printers by Corona, Apple, HP, Imagen, and QMS. A *Preview* program (with graphics adapter) allows viewing the final output before printing, and *MF Medley* provides 44 additional fonts from the Metafoundry library. Write for hardware requirements and full price list.
☐ PCTEX $249.00; PCDOT $95.00; PCLaser $275.00

Conference Proceedings

Theoretical Aspects of Reasoning About Knowledge: Proceedings of the 1986 Conference
Edited by Joseph Y. Halpern
A chronicle of the first multidisciplinary conference in this area, this volume explores the theory and application of reasoning about knowledge in distributed system analysis, natural language processing, knowledge representation, and game theory. Original papers and overviews from leading researchers document the emergence of an important field.
☐ Paper; ISBN 0-934613-04-4; $18.95

VLDB-86 Kyoto
The written record of the Twelfth Conference on Very Large Data Bases. Rigorously refereed papers from an international community of researchers explore fundamental issues and current development trends in data base research.
☐ Paper; ISBN 0-934613-18-4; $35 ($55 Institutions)

VLDB-85 Stockholm
☐ Paper; ISBN 0-934613-17-6;
$35 ($55 Institutions)

VLDB-84 Singapore
☐ Paper; ISBN 0-934613-16-8;
$35 ($55 Institutions)

VLDB-83 Florence
☐ Paper; ISBN 0-934613-15-K;
$35 ($55 Institutions)

VLDB-82 Mexico City
☐ Paper; ISBN 0-934613-14-1;
$35 ($55 Institutions)

ABLEX PUBLISHING CORPORATION

Ablex will be displaying the following new titles at the 1986 Cognitive Science Society Conference:

ADVANCES IN HUMAN/COMPUTER INTERACTION VOLUME ONE
R. Hartson, ed.

EMPIRICAL STUDIES OF PROGRAMMERS
E. Soloway and S. Iyengar, eds.

FORMAL THEORIES OF THE COMMONSENSE WORLD
J. Hobbs and R. Moore, eds.

FROM MODELS TO MODULES
I. Gopnik and M. Gopnik, eds.

FROM PIXELS TO PREDICATES
A. Pentland, ed.

IMAGE UNDERSTANDING 1986
W. Richards and S. Ullman, eds.

INFINITE SYNTAX
John R. Ross

LANGUAGE LEARNING AND CONCEPT ACQUISITION
W. Demopoulas and A. Marras, eds.

LOGIC PROGRAMMING AND ITS APPLICATIONS
M. VanCanegham and D. Warren, eds.

MANAGING COMPUTER IMPACTS
N. Bjorn-Andersen et al, eds.

MEANING AND COGNITIVE STRUCTURE
Z. Pylyshyn and W. Demopoulas, eds.

NEW DIRECTIONS IN COGNITIVE SCIENCE
T. Shlechter and M. Toglia, eds.

NEW DIRECTIONS IN DATABASE SYSTEMS
G. Ariav and J. Clifford, eds.

PLANNING, GEOMETRY, AND COMPLEXITY OF ROBOT MOTION
J. Schwartz et al, eds.

STAGE AND STRUCTURE
I. Levin, ed.

STRUCTURES AND PROCEDURES OF IMPLICIT KNOWLEDGE
A. Graesser and L. Clark, eds.

UNDERSTANDING COMPUTERS AND COGNITION
T. Winograd and F. Flores

WHAT IS INTELLIGENCE?
R. Sternberg and D. Detterman, eds.

A wide range of journals, including COGNITIVE SCIENCE, will also be exhibited.

We look forward to assisting you at our book exhibit.



LAWRENCE ERLBAUM ASSOCIATES, INC.

Visit our booth and examine these new titles

BASIC PROCESSES OF LEARNING, COGNITION, AND MOTIVATION

Stephen M. Cormier

Research in learning, cognition, and motivation is analyzed at both biological and neurological levels in this sophisticated study. Cormier identifies the fundamental processes of information and motivation in humans and animals and shows how these basic constructs influence phenomena such as cognitive development, aptitudes, cue selection, habituation, and the information processing and behavioral functions of the limbic system.

0-89859-689-0 1986 / 472 pp. / $29.95

SEX DIFFERENCES IN COGNITIVE ABILITIES

Diane F. Halpern

A critical review of a highly controversial topic, Halpern's latest book synthesizes theories, research, and research methods that explore how and why selected intellectual abilities vary as a function of sex. The practical implications of these differences are examined, as are the validity and importance of biological, social, familial and educational factors toward the cognitive abilities of females and males.

0-89859-579-7 1986 / 208 pp. / $29.95 cl
0-89859-838-9 $16.50 ppr

EXPLANATION PATTERNS: Understanding Mechanically and Creatively

Roger Schank

An exploration of the human mind's basic mechanisms and the attempts of researchers to simulate the most human mechanism of all — creative thought. Cognitive science and AI perspectives are represented in this stimulating investigation of thought, creativity, memory, learning and AI capabilities.

0-89859-768-4 Fall 1986
256 pp. (tent.) / $24.50

APPLICATIONS OF COGNITIVE PSYCHOLOGY Problem Solving, Education, and Computing

Edited by Dale E. Berger, Kathy Pezdek, William P. Banks

Here is a detailed account of practical applications for cognitive research. Issues in education, problem solving, and computer science are addressed in light of current advances in cognitive psychology, touching on questions of classroom technology, thinking skills, and computer interface design. The book is a classic example of how cognitive theory can be used to solve applied problems, and how the study of applied problems can further cognitive research.

0-89859-710-2 Fall 1986
264 pp. (tent.) / $29.95 (tent.)

Forthcoming

MECHANISMS OF LANGUAGE ACQUISITION
The 20th Annual Carnegie Symposium on Cognition
Brian MacWhinney
596-7 cl *in press*
973-3 ppr *in press*

THE CEREBRAL COMPUTER: An Introduction to the Computational Structure of the Human Brain
Robert J. Baron
824-9 *in press*

Now in paperback

THE PSYCHOLOGY OF HUMAN-COMPUTER INTERACTION
Stuart K. Card, Thomas Moran, Allen Newell
859-1 488 pp. / $24.95

Best-Selling Text

THOUGHT AND KNOWLEDGE
An Introduction to Critical Thinking
Diane F. Halpern
381-6 1984 / 424 pp. / $39.95 cl
514-2 $19.95 ppr

Now Available

EXPERIENCE, MEMORY, AND REASONING
Janet L. Kolodner and Christopher K. Riesbeck
A volume in the Artificial Intelligence Series
664-0 1986 272 pp. / $29.95

Mention this ad with your order and save 15% when accompanied by payment. Write:

365 Broadway • Hillsdale, New Jersey 07642

Prices are subject to change without notice

686071

NOW AVAILABLE

Proceedings of the

Fourth (1982)
Fifth (1983)
Sixth (1984)
Seventh (1985)

Annual Conference of the Cognitive Science Society

$25.00 each, paperbound



LAWRENCE ERLBAUM ASSOCIATES, INC.
365 Broadway • Hillsdale, N.J. 07642
Prices are subject to change without notice.

0-89859-864-8

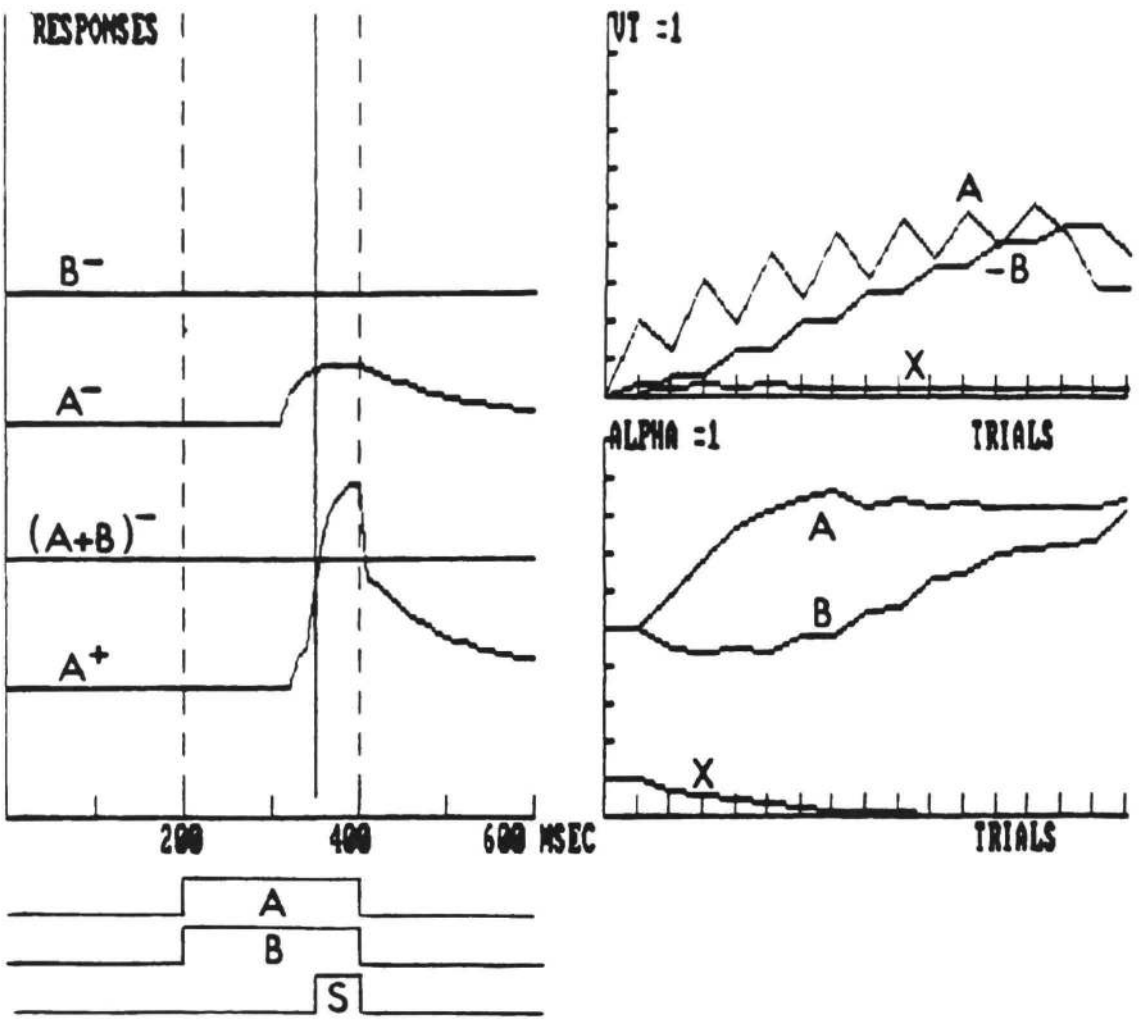


Figure 2. Conditioned Inhibition. A : CS(A). B : CS(B). X : Context. Left Panels: NM response topography in A+, (A+B)-, A-, and B-trials, after 10 alternated A+ and (A+B)- trials. Upper-Right Panels: Net associative values (VT) at the end of each trial, as a function of trials. Lower-Right Panels: Associability (ALPHA) at 350 msec, as a function of trials.

1983)

Conditioned Inhibition. Figure 2 shows simulations of a conditioned inhibition paradigm. During conditioned inhibition two types of trials were alternated: reinforced trials consisted of a single reinforced CS (A), and nonreinforced trials consisted of a compound CS (A and B). Stimulus B was the conditioned inhibitor. After 10 simulated trials, the CR elicited by A and B together was smaller than that elicited when A was presented alone because B has acquired inhibitory associative value. Associabilities of both CSs increased and context associability decreased over trials. Simulation results agree with data on conditioned inhibition reported by Marchant, Miss, and Moore (1972).

Blocking. Figure 3 shows simulations of a blocking paradigm. Experimentals received 5 trials with one CS (blocker) paired with the US followed by 5 trials with the same CS and a second (blocked CS) paired with the US. Controls received 5 two-CS trials in which both CSs were presented together and paired with the US. Controls were subject to mutual overshadowing between the two component CSs. The network showed simulated blocking because CR for the designated blocked CS was smaller than CR for the blocker CS, both after 5 training trials. The results agree with blocking data in the rabbit NM response preparation as reported by Marchant and Moore (1973).

Sensory preconditioning. Figure 4 shows simulations of a sensory preconditioning paradigm. In the first phase, 5 nonreinforced trials with a compound CS(A and B). During the second phase, one of the nonreinforced CSs (A) was reinforced for 5 trials. A test trial assessed the CR to CS(B) never paired with the US. Simulations showed that context associability decreases during preconditioning. In the