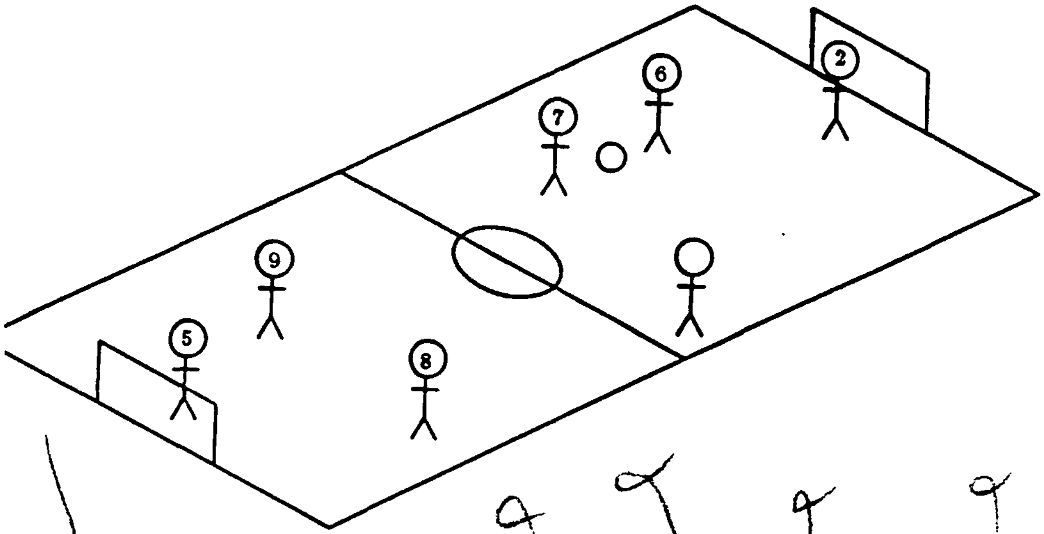
(2)

REPRESENTATIONS IN MENTAL MODELS (FINAL REPORT) 1990

Investigator:
(617) 253-5776
WHIT@EGO.MIT.EDU
Contract No. 90-0177

Whitman Richards E10-120
Massachusetts Institute of Technology
Cambridge, MA 02139



Abstact: On March 12-13 an interdisciplinary group of thirty-five, composed of computer scientists, experimental psychologists, linguists, philosophers and "connectionists" met to share views on representations and their role in mental models. Although at least two books and several papers directly address these issues, the nature of mental models is far from clear. The meeting shed some light on why "understanding" mental models is difficult. Simply put, the reason is that mental processes are described in many different ways and at quite different levels of abstraction, depending upon the researcher. For example, some emphasize the cognitive properties of mental models, whereas others are more concerned with the internal data structures. Still others may stress the logical form and content of the mental process, as contrasted with the actual computational machinery. The diversity of these viewpoints is clear upon reading the abstracts prepared by the participants. Further study is needed to examine how these diverse viewpoints fit together into a useful, integrated framework.

artificial intelligence
Key words: ~~AI~~, Cognition, Data Structures, Knowledge Representation, Language, Mental Models, Neural Nets, Perception, Reasoning , (KR)

September 1990

Contents

Cover figure: courtesy of Alan Mackworth

15 May 1990

1.0 Overview

On March 12-13, an interdisciplinary group of thirty-five, composed of computer scientists, experimental psychologists, linguists, philosophers and "connectionists" met to share views on representations and their role in mental models. The meeting was held at the American Academy of Arts and Sciences in Cambridge, Mass., under the sponsorship of the AFOSR. Although at least two books and several papers directly address these issues[1], the nature of mental models and their representations is far from clear. The meeting shed some light on why understanding "mental models" is difficult. At the same time some insight was obtained as to why the Cognitive Sciences have had difficulty in developing an integrated, focussed discipline typical of the older Natural Sciences or of the more recently formed Neurosciences.

The Problem for Consideration

"There is a strong intuition that there are at least two quite different types of representations in cognitive systems, one 'iconic' based upon pictures and the other 'symbolic' based upon links among symbols. (Distributed, connectionist representations might be still another form.) Pictures seem more suitable for geometric or spatial reasoning; whereas the symbolic, language-like form is common in language and reasoning where quantification is important. A 'Mental Model' may use one or both forms, or perhaps still another. What is the role of such representations in the functioning of mental models?"

The above statement was posed to all participants, who were required to submit an abstract of their view of the role of representations in mental models. Prior to the meeting, the abstracts were circulated to all to accelerate interchanges. (These abstracts are appended.) The meeting then began with three tutorial-like presentations of different types of representations, followed by case studies which were designed to contrast the use of quite different representations in supporting the same type of problem. For example in one reasoning session, Erik Sandewall's more sentential, logic approach was contrasted with Bernard Meltzer's analog or iconic reasoning with "pictures", and these in turn were to be contrasted with a more connectionist view as represented by Paul Rosenbloom (or in the language area by David Touretzky or Paul Smolensky). The program is also appended.

Unfortunately, illness prevented many of the connectionists from attending. This was a serious loss which crippled discussions of such issues as the role of working memory, for example. Consequently most of the discussion centered upon the spectrum of representations spanning pictures to sentences. After some delay in defusing the obvious fact that all such (feasible) computations are Turing reducible and hence "symbolic," more attention was drawn to the utility of the form of the representation in supporting particular tasks. For example, Levesque's definition of a vivid representation directly addresses this issue, with the definition applying to both the extreme iconic or sentential forms. (However, see Davis's concern where conclusions based upon partial information is needed.) In many reasoning tasks, logic is the preferred vehicle, but all logical representations are not vivid, nor will all logics support the human mental model process, as pointed out by Johnson-Laird. The utility of pictorial representations in recognition was stressed by Ullman. Mackworth stressed that both symbolic and pictorial forms commonly appear together, such as in maps, diagrams or plans, and showed how a logical model can help to understand their relationships. Because logic was the most universal language of the group, the logical form of the various representations tended to dominate the discussions at the detriment of understanding the utility of the representation per se with respect to various tasks.

As the meeting progressed, attention became more focussed upon what kind of computation a representation actually supported best, rather than with the logical content of the representation. For example, Krumhansl presented evidence for several representations other than iconic or symbolic in the perception and production of music; Norman stressed the role of artifacts in the choice of representation, thereby opening the spectrum of possibilities still further; Hayes raised the issue of external versus internal representations; and Smolensky attempted to distinguish between representations and to show the richness of possibilities, as did Philip Johnson-Laird in his overview of "What is a Mental Model?" By the end of the meeting, there seemed to be a consensus that the spectrum of representations is very rich and that they can be distinguished on many dimensions. As aptly summarized by McDermott, even if one wishes to consider only the simple iconic-symbolic dichotomy, as originally proposed, then possible distinctions might include "detailed vs. non-committal", "spatially vs. non-spatially indexed", "intrinsically vs. extrinsically" constrained. These distinctions provoke a reexamination of the utility of a representation independently of its reduction to a logical form. (See Johnson and Rosenschein, for example.)

In the same spirit, McDermott also attempted to distinguish between views of "models" in reasoning, recognition, or communication. (See Pinker also for distinctions between rule-based and associative models.) Here the group failed to reach any clear agreement. Some members emphasized the cognitive properties of mental models, whereas others were more concerned with the internal data structures. Similarly, some were mostly concerned with the logical form and content of the mental process and its definition (see Reiter), as contrasted with the actual computational machinery (see Hopfield, Jordan or Richards for example). Consensus was not reached here simply because the mental process was being described in quite different ways and at quite different levels of abstraction. Further study is needed to examine how these different descriptions may be combined into a useful, integrated framework. (See Peters and Kaplan.)

The Nature of Cognitive Science

The difficulty the group had in sharing a common framework for "What is a Mental Model?" provides insight into why some believe that disciplines within the Cognitive Sciences are tending to diverge, rather than to converge. These sciences are not like its sibling the Neurosciences, whose disciplines in the early days shared the common bond of wishing to understand neural machinery. Much like the connectionists today, neural systems (or "nets") at that time could be viewed as machinery similar in principle across a wide spectrum of biological systems. This is not true of mental operations, where quite different representations are needed to support a variety of cognitive tasks. Furthermore, in the Neurosciences the types of descriptions and levels of understanding desired were similar – in particular, emphasis was placed upon mechanisms and algorithms – whereas in the Cognitive Sciences the scope of desired descriptions and explanations include not only these, but also the more cognitive operations and conventions used to infer or reason about a variety of aspects of the world. This spectrum of interests and descriptions in the Cognitive Sciences was quite apparent in the dynamics of the meeting. Because logic was the language of many, and because the logical form and completeness of a representation cuts across many problems, the discussions tended to center upon logical content and favored those fluent in this language. But this focus excluded explanations of the mental process in terms of its computational or psychological properties. Hence seldom did the experimental psychologists or connectionists become involved, and even the linguists were largely silent. The reason in retrospect is clear: at this meeting we chose a diverse group, with each participant proficient in one of three types of representations (iconic, sentential, connectionist) for use in one of three problem areas (reasoning, recognition and language) with interests in one of three levels of explanation (logical form, algorithms, or psychological properties). We only need to scan the abstracts to see the wide diversity in viewpoints, problems, and the language or "tools " used to achieve an understanding of these problems. "Natural" grammars may be an appropriate vehicle for the linguist to study language, but of what use is a grammar to the individual studying object recognition or iconic reasoning when morphological operations or geometric transformations currently seem more suitable? Different cognitive tasks create different classes of problems which generally require quite different tools and expressions for solutions. Clearly each discipline within Cognitive Science has its own, quite different "Mental Model." The situation is not at all like the early days of the Neurosciences, where everyone's ideas were reducible to the behavior of neurons.

Cognitive Management

Should Cognitive Science thus be resigned to accept a "divide and conquer" strategy, with each subdiscipline studying its different problems independently, with each developing their own framework and "language"? Such a commitment might well penalize our understanding of the essence of our own mental processes. What seems remarkably clear is our ability to flow so easily from one type of representation to another, such as when we move from symbols to pictures as in navigation or spatial maps (Mackworth, DeJong), or when we solve problems by visualizing our actions in the world together with simple logics (Etchemendy, Johnson-Laird, Jepson, McDermott) or perhaps even when we learn a language (Johnson, Webber, Prince). In many of these cases the type of the representation (symbolic-iconic) is not the crucial issue; rather it is how constraints in the environment and agent-environment interactions can be exploited and embedded implicitly in the reasoning or perceptual process, (See Levelt, for example.) Perhaps at this early juncture in our understanding of representations in mental models, it will be prudent not to always focus too narrowly upon representational issues at the exclusion of the management of information flow across representations.

Whitman Richards[2]
May 1990

Acknowledgement: This meeting was supported by a grant from the Air Force Office of Sponsored Research (AFOSR-90-0177). Joshua Stern and Jacob Feldman were most helpful in criticizing earlier drafts of this report.

[1] Johnson-Laird, P.N. (1983) *Mental Models.* Cambridge, MA: Harvard University Press.

Gentner, D. & Stevens, A.L. (Eds.) (1983) *Mental Models.* Hillsdale, NJ: L. Erlbaum Assoc.

Rumelhart, D. & Norman, D.A. (1988) Representation in memory. In: R.C. Atkinson, R.J. Herrnstein, G. Lindzey, R.D. Luce (Eds.), *Stevens Handbook of Experimental Psychology* Vol. 2, New York: Wiley Press.

[2] Address: Dept. of Brain and Cognitive Sciences
Mass. Institute of Technology
79 Amherst St. E10-120
Cambridge, MA 02139

Advisory Committee:

Geoffrey Hinton, Univ. Toronto, Dept. of Computer Science

John Hopfield, Cal. Tech., Crellin Laboratory

Philip Johnson-Laird, Princeton University (formerly MRC, Cambridge, U.K.)

Drew McDermott, Yale, Dept. Comp. Science

Alan Mackworth, UBC, Dept. Comp. Science

Stanley Peters, Stanford, Dept. of Linguistics

Whitman Richards, MIT, Dept. Brain & Cognitive Sciences

Final Program*

Representations

Hector Levesque: Vivid Representations

Shimon Ullman: Pictorial and Symbolic Representations in Object Recognition

Alan Mackworth: Depiction Theory

Mental Models

Philip Johnson-Laird: What Is a Mental Model?

"Language" Understanding

Carol Krumhansl: Internal Representations for Music

Ronald Kaplan: Representational Transformations

Paul Smolensky: A Connectionist View

Reasoning

Eric Sandewall: Does Persistence Occur Outside the Mind of the Beholder?

Bernard Meltzer: Direct Representations of Naive Physics

Donald Norman: Cognitive Artifacts

John Etchemendy: Heterogenous Reasoning

Comment: Willem Levelt: Perspective is Free

Issues – Recapitulation

Drew McDermott

[*] Due to illness, three connectionists originally scheduled for presentations are omitted – an unfortunate bug in their network!

3.0 Problem Statement

There is a strong intuition that there are at least two types of representations in cognitive systems, one "Iconic" based upon pictures, and the other "Symbolic" based upon links among symbols. ("Distributed, connectionist" representations might be still another form.) Pictures seem more suitable for geometric or spatial reasoning; whereas the symbolic, language-like form is common in language and reasoning where quantification is important (for all ..., there exists ...). A "Mental Model" may use one or both of these forms, or perhaps still another. Although we propose to examine the definitions and essential features of these alternate representations, our primary aim will be to explore their roles in the functioning of mental models. The method we will use to address these issues will be selected case studies of mental models in humans and artifacts. (Abstracts of each participant's position with respect to the above follow.)

E. Charniak
E. Davis
G. DeJong
J. Etchemendy
P. Hayes
G. Hinton*
J.J. Hopfield
A. Jepson*
M. Johnson
M. Jordon
P. Johnson-Laird
R. Kaplan
S. Kosslyn
C. Krumhansl
H. Levesque
W. Levelt
A. Mackworth
D. McDermott
B. Meltzer
D. Norman
S. Peters
S. Pinker
A. Prince
Z. Pylyshyn
R. Reiter
W. Richards
P. Rosenbloom*
S. Rosenschein
D. Rumelhart*
E. Sandewall
G. Sperling
P. Smolensky
D. Touretzky*
A. Treisman
S. Ullman
B. Webber

*absent

Probabilities as Poor-Man's Connectionism

Eugene Charniak

Brown University
Department of Computer Science

I am, at heart, a connectionist. However, with a single exception, every paper I have published (and continue to publish) has assumed the "standard" AI approach with its close coupling between meanings and the data-structures which encode them. In this note I will try to explain the seeming contradictions therein.

I find connectionism appealing on several grounds. First, I am all too aware of the problems inherent in the traditional view, having butted my head against them for more than twenty years now. (How time flies.) Secondly the argument for parallel processing in brain computation seems very strong, and I see no other even semi-plausible model for such computation. Lastly, I find the abstraction from neurons to connectionist "units" semi-reasonable. It may be that none of the current abstractions is correct, but I believe that the correct abstraction will have the same "spirit." (In particular, I would be very surprised if the correct answer had anything resembling "cons.")

My problem is that while I believe all of this, I also believe the criticism of existing connectionist models of high-level cognition as put forward by Fodor, Pylyshyn, McDermott, etc. Where I differ, say, from Fodor and Pylyshyn, is that I view these inadequacies as problems for future research, not as reasons to reject the paradigm. Unfortunately I do not see any way to attack these problems directly, and thus I am reduced to my current research strategy, which is to keep the standard model, with its explicit encoding of meanings, and try to cater to my connectionist inclinations by bringing other aspects of my models closer to connectionism.

In particular, I have now become a probabilist in that I have adopted probability theory as a method of deciding on the best interpretation of a text. This is to say, questions of word-sense disambiguation, pronoun reference, etc. are recast as questions like, "what is the posterior probability that this token of the word 'bark' means DOG-NOISE (as aoppsed to TREE-COVER), given the evidence provided by the text?" Then the alternative with the highest posterior probability is taken to be the correct disambiguation for the word. Naturally how the relevant probability distributions are created, particularly when the various decisions are quite dependent on context, is an interesting question, but one beyond the bounds of this note. Rather it is important here to note the similarity with connectionist models in which several alternative interpretations can be represented simultaneously, but with different activity levels. I am hoping that these activity levels can be given something like a probabilistic semantics since it is hard to see what the alternative could be.

Thus I see my work on probabilistic models of language comprehension as as sort of compromise between my connectionist inclinations, and my complete inability to overcome the representational inadequacies of connectionism. This is not to say, however, that one must buy connectionism to see the desirability of a probabilistic model of abductive choices in language. Since there is no other formal model of such choices, probability theory is really the only game in town.

Partial Information and Vivid Representations

Ernest Davis

New York University
Courant Institute

My research is in commonsense reasoning and representation of commonsense knowledge. From this viewpoint, iconic representation, or more generally vivid representations (in the sense of Levesque), are attractive in that they support inference procedures that are efficient and involve little search. However, vivid representations do not easily support the inference of valid conclusions from partial information. This ability to use partial information is essential to commonsense reasoning.

Various techniques have been proposed to carry out reasoning from partial information using vivid representations; none of these are very satisfactory:

1. The vivid representation may be given labels indicating which aspects are known, and which have been added arbitrarily to make it vivid. However, if an inference procedure uses such labels respectfully, inferring a fact as known only if it is supported by information labelled known, then it could as well use just the labels; the vivid representation adds nothing.

2. Partial information may be recorded in terms of a collection of vivid representations; the certain information is that which is common to the collection. This significantly reduces both the efficiency and the intuitive appeal of vivid representations. Worse, it is difficult to combine two such collections which express overlapping information.

3. A third possibility is to use non-vivid constraints as permanent knowledge structures, and to perform inferences by using Monte Carlo techniques to generate a sampling of vivid representations satisfying the constraints. Again this sacrifices the intuitive appeal of vivid representations. The computational problems are even worse than in (2); it is often as difficult to find any vivid representation satisfying a system of non-vivid constraints as to perform any other basic inference on those constraints. Finally, this technique runs the risk of overlooking some case which occupies only a small area in a Monte Carlo search but in reality should be a salient possibility. For example, such a process might conclude that a key could not fit inside an unseen lock, because an extensive search among combinations of cylinder sizes had failed to reveal one fitting the key.

Heterogeneous Reasoning

John Etchemendy

Stanford University
Department of Philosophy

For the last year and a half, Jon Barwise and I have been studying (from the point of view of logicians, not psychologists) what we call *heterogeneous reasoning*. What we mean by this is reasoning in which information is provided or manipulated in multiple forms, for example, in the form of both sentences and diagrams. During the past 100 years, logicians have made great progress in the study of homogeneous, linguistic reasoning -- reasoning that begins, ends, and never strays from sentences in some language (generally sentences of a formal language like the first-order predicate calculus). But few, if any, of the mathematical tools developed to study purely linguistic inference can be applied to the heterogeneous case.

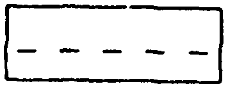
We have three goals in this research. First, we hope to convince logicians of the importance *and legitimacy* of reasoning that makes essential use of nonlinguistically represented information (diagrams, pictures, and so forth). Second, we are trying to *develop the requisite mathematical tools for assessing the validity or invalidity of such* reasoning. Finally, we are creating a computer program, called *Hyperproof*, for use in teaching a simple form of heterogeneous reasoning. In my talk, I will try to convince you of the first, briefly describe the second, and show a quick demo of the third.

The sort of reasoning we have examined most closely is typified by problems found in puzzle magazines or on the analytical reasoning section of the Graduate Record Examination A very simple example of this kind of problem is described in Example 1.

Example 1 You are to seat four people, *A*, *B*, *C*, and *D* in a row of five chairs. *A* and *C* are to flank the empty chair. *C* must be closer to the center than *D*, who is to sit next to *B*. From this information, show that the empty chair is not in the middle or on either end. Can you tell who must be seated in the center? Can you tell who is to be seated on the two ends?

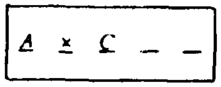
I urge you to solve this problem before reading on. As simple a it is, you will no doubt find that the reasoning has a large visual component. Probably you will find it useful to draw some diagrams. With more complex problems of this sort, diagrams become even more essential.

One line of reasoning that can be used in solving this problem runs as follows. Let us use the following diagram to represent the five chairs.

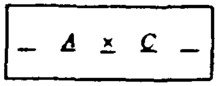


Our first piece of information tells us that *A* and *C* are to flank the empty chair. Let us use "x" to signify that a chair is empty. Then we can split into six cases. Or, since the problem does not distinguish left from right in any way, we can limit our attention to three cases, the other three being mirror images of them.

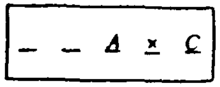
Case 1



Case 2

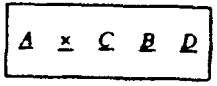


Case 3

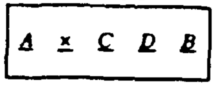


Using the fact that *C* must be seated closer to the center than *D*, we can eliminate Case 3, since *C* is not closer to the center than any available chair. Similarly, since *D* must sit next to *B*, we can rule out Case 2, since no contiguous chairs are available. This leaves us with the following two possibilities:

Case 1.1



Case 1.2



In both of these cases, all of the stated constraints are satisfied, and so we know that neither case can be ruled out. This allows us to answer the questions posed in the puzzle. First, we see that in both cases the empty chair is not in the middle or on either end, as desired. Second, we see that *C* must be seated in the middle. And finally, we see that *A* must be on one end of the row, but that we do not know whether *B* or *D* is on the other end. Either one is possible.

We are convinced that, properly understood, the demonstration above is a valid proof of the conclusions reached, not just a psychological crutch to help us find such *a prof, as the traditional account would have us believe. Moreover, we think that there* are some parts of this proof that are missed in traditional accounts of inference. Notice, for example, that although our reasoning is all of a piece, the three parts of the problem have quite different characteristics when looked at from the traditional perspective. The first question asks us to prove that a specific fact follows from the given information. The second, in contrast, ask us whether a certain sort of information is implicit in the given information, and the answer is *yes*. The third question is of the same sort, but here, we end up showing that something does *not* follow from the given, namely, who is seated on the end opposite *A*. We show this *nonconsequence* result by coming up with models of the given, one of which has *B* on the end, while the other has *D*. Normally, showing that something follows from some assumptions and showing that something does not follow are thought of as dual tasks, the first deduction and the second model building. This dichotomy seems unhelpful in analyzing the reasoning above, since there was no apparent discontinuity between the portions of the reasoning that demonstrated consequence and the portions that demonstrated nonconsequence. This occurence, in one reasoning task, of parts that are usually thought of as duals of one another (deduction and model construction) turns out to be typical of a whole spectrum of reasoning tasks. The desire to account for this dual nature of ordinary reasoning is an important motivation for some of the details of our theory of heterogeneous inference.

An Efficient Method of Representing Spatial Structure in Neural Networks

Geoffrey Hinton

University of Toronto
Department of Computer Science

I have no thoughts on mental Models.

However, I have an idea *about how to represent* spatial structures efficiently in neural networks (as may title indicated). This idea may help to clarify what is meant by "iconic". The scheme I shall present requires a person to adopt a viewpoint in order to access their knowledge of spatial structures.

Of Representations and Models - a View from the Bottom

J.J. Hopfield

Caltech
Crellin Laboratory

The details of the sensory biophysics at the molecular or receptor level determine which stimuli a sensory neuron will respond to, and thus create a low-level *representation of the sensory world.* Higher level neurons construct their own representations on the basis of the lower level representations. In a simple organism, a model might be defined as the way in which an appropriate response or action is calculated from the available sensory stimuli. There is no real separation between representation and the calculational algorithm built on these representations. Both are a consequence of the details of neuroanatomy and neurophysiology. The abstraction of representations is merely the first step in modeling. In simple organisms and simple senses, models are implicit, and not available to inspection by cognitive analysis in that organism. In highly non-linear systems like neurobiology, it is not possible to do a "complete set of experiments", and from direct experimentation fully test a model. *The study of what particular cells are doing* in behaving animals will be essential for an understanding of cognition in simple animals. And it is difficult to believe that higher animals will be easier in this regard

Symbols and Simulators

Allan Jepson

University of Toronto
Department of Computer Science

Intuitively it seems people can use at least two modes of representation, namely iconic and symbolic. In order to study computational systems which can use both these modes we *first need to formally specify* which *inter-modal* transformations are legal. For example, consider transforming the sentence "A bird that cannot fly" to a mental image. Such an image contains an iconic representation *of a bird, and an index pointing to the fact that it cannot fly.* Here the index for "cannot fly" could be that the imaged bird is an ostrich, or the bird might have a visible bandage on a wing, etc. We would not want a mental image of a robin "just *standing* there" to be an allowable iconic representation of the given sentence, since there would be no evidence for why the bird *cannot fly.* A specification of what constitutes a legal transformation might be made using the recent formal theory of perception proposed by Jepson and Richards. In general such a specification would allow simulators (which are iconic in that they represent through an isomorphism of properties and operations) to be formally tied to symbol systems.

There are several potential advantages of having both types of representation within one system. Simulators can efficiently represent concrete cases, which can be manipulated with a small set of fast operators. They could facilitate such tasks as parameter estimation, case-based consistency checks, and likelihood estimation. On the other hand, symbol systems provide a concise framework for expressing and manipulating declarative knowledge. They appear to be the natural choice for controlling which simulations are performed.

Mental Models That Aren't

Mark Johnson

Brown University
Cognitive and Linguistic Sciences

It's only a slight oversimplification to say that much recent work on mental representations and models is based on the analogy between mental processes and computer programs, and mental representations/models and "data structures" that the program constructs. For example, because there is considerable linguistic evidence that hierarchical tree structures are an important component of the mental representation of utterances, much psycholinguistic and computational work on natural language processing concentrates on discovering how computational processes can actually build such tree structures. But even though it is relatively well-known that data structures are only organizational abstractions (data-structure declarations in a computer program are typically "compiled out"; hence many different data structures can result in identical object code; i.e. identical computational processes), most work is formulated as if "the" data structures that constitute mental representations/models can be unambiguously identified. This paper describes recent research in computational linguistics which suggests that the analogy between mental representations/models with data structures, and the ascription of such structures to computational processes, may be more problematic than previously thought.

One of the difficulties in the construction of computational models for Chomsky's "Government and Binding" theory of human language is that the appropriate representations of an utterance at different linguistic levels are determined by a system of tightly interacting constraints. Simple-minded implementations which construct these representations one by one cannot fully exploit the constraints on the first representations they construct imposed by the levels of representation they have not yet constructed: a disasterous short-coming since these reduce an infinite search space to a finite one. This problem can be avoided by determining in advance some or all of the constraints that the later, as yet unconstructed representations will impose on earlier ones. It leads to a constraint-oriented rather than representation-oriented perspective: the construction of one representation requires only the constraints originating from the other representations, not the representations themselves. If *all* constraints originating from some representation can be determined without the representation, then the representation need not be constructed by the computational model. Simple natural language parsers can be described that function in exactly this way, avoiding the explicit construction of several levels of linguistic representation, although provably producing analyses that satisfy all constraints on all linguistic representations (including those unconstructed).

Such computational models pose interesting problems for the "mental model as data structure" analogy discussed above, since arguably the computational model does exploit all of the levels of representations of an utterance (where else could the relevant constraints come from?), even though it does not construct data structures corresponding to each level of representation.

Forward models and Inverse Models

Michael I. Jordan

Massachusetts Institute of Technology
Department of Brain and Cognitive Sciences

When a system interacts in a closed loop with the environment, there are several roles for "mental models" that it is useful to distinguish. One important distinction is that between "forward models" and "inverse models." A forward model of the environment is a model of the state transition function of the environment, indexed by the control sequence. For example, a multi-step forward model is a mapping of the form $x_{n+k} = f(x_n, <u_n, u_{n+1}, \ldots, u_{n+k-1}>)$, where x is the state and u is a control action. An inverse model is a model whose output is the control sequence that is required to effect a particular state transition; that is, a mapping of the form $<u_{n+k-1}, \ldots, u_{n+1}, u_n> = \psi(x_{n+k}, x_n)$. Both kinds of models are useful in systems that interact with the environment. Forward models can be used for filtering and are useful in internal closed loops that reduce the effects of delays. They can be used for offline search. Inverse models can be used to compute feedforward control terms that allow goals to be achieved directly without iterative error-correcting computation. Moreover, forward models and inverse models have complementary roles when one considers how models might be acquired. It can be shown that forward models yield learning operators that allow adaptive changes to be made to inverse models. Inverse models, on the other hand, are useful in obtaining data for the learning of forward models.

What Is a Mental Model?

Philip Johnson-Laird

Princeton University
Department of Psychology

The concept of a mental model is not one that was defined a priori. It emerged from the work of various cognitive scientists as a way of explaining various phenomena. This talk will present one such case history: a unified conception of mental models explains the reasoning of logically-untutored individuals. It accounts for their fundamental competence and systematic errors in propositional reasoning, relational reasoning, quantificational reasoning, and meta-logical reasoning. The theory has been implemented computationally and throws light on the integration of valid deduction and non-montonic reasoning, and the framing of parsimonious conclusions. The view of mental models that is currently emerging from this work appears to be compatible with other concepts of them.

Representational Transformations

Ronald Kaplan

Xerox Palo Alto Research Center
Palo Alto, California

There is a long tradition of language research in which models and results are couched in terms of representations that are variously described as modular, symbolic, rule-based. More recently, there have been new proposals--connectionist, parallel distributed processing, neural network--to simulating linguistic phenomena that scrupulously avoid such discrete representations. These two approaches are often viewed as incompatible and sometimes adversarial. I'll describe some formal results on certain kinds of modular rule systems which permit them to be transformed into less modular and less rule-like representations that can be shown to be behaviorally equivalent - they have exactly the same input/output behavior. The two provably equivalent representations for the same phenomena are not necessarily in conflict--they may simply provide different kinds of scientific insight into the underlying linguistic reality.

Mental Models and Memory Representations

Stephen M. Kosslyn

Harvard University

Mental models are representations in memory that can be operated on in specific ways. In thinking about such representations, it seems useful to distinguish between different forms of memory representations:

Long-term memory (LTM)

There are two classes of long-term memory representations (in "declarative" memory): those that are amodal and those that are modality-specific. Information in long-term memory can be activated, and one form of activation of modality-specific representations causes a pattern of activation in a short-term memory structure. This mechanism is used in attention, top-down processing during object identification, and visual mental imagery.

Short-term memory (STM)

Some short-term memory structures are intermediate-level perceptual structures, and hence are modality-specific. These structures can briefly retain sensory input, and also can be activated on the basis of information stored in modality-specific LTM. A visual mental image is an example of a pattern of activity in a visual structure that is induced on the basis of stored information (recent PET scanning results from a collaboration with Nat Alpert's group at the MGH support this view). Other short-term memory structures are intermediate-level output buffers. Rehearsing a word to oneself makes use of such a buffer. Again, the information in this structure is filled in from information in long-term memory.

If short-term memories occur in perceptual structures, then their notorious capacity limits may be a consequence of the neural substrate: One does not want "smearing" in vision, and so representations do not persist very long after the eyes move. This "fast fade rate" would also affect representations that arise from stored information, and hence effort would be required to maintain information in short-term memory. And only a certain amount of information can be maintained, given a certain amount of processing capacity (these are vague ideas, but I think they can be firmed up considerably).

Working memory

Working memory is a term that includes the activated material in long-term memory and the short-term memory representation. Because of capacity limits in short-term memory, there often may be more information activated in long-term memory than can be used to create short-term memory representations. This information is used to change the STM representation as necessary to help one notice specific relations. Furthermore, there is a dynamic interplay between the current contents of STM and activated information in LTM; information in STM may evoke additional representations in LTM, and vice versa. In my view, mental models can best be understood in terms of such an interplay in working memory.

Specifically, in my view, visual mental images often serve as mental models. One manipulates the imaged object, observing the consequences in order to predict what would happen in the analogous physical case. As needed, additional properties (e.g., color, texture, parts) are added to the image (on the basis of information stored in LTM), while unnecessary aspects are deleted (respecting the capacity limits of STM). Such images are "inspected" using exactly the same mechanisms used in perceptual recognition; once a pattern of activation arises in the intermediate perceptual structures, it is processed downstream in the same way regardless of its origin (from LTM or sensory input).

Internal Representations for Music

Carol L. Krumhansl

Cornell University
Department of Psychology

Most information concerning the nature of internal representations comes from theoretical and experimental analyses of language and vision. Although music is viewed as having commonalities with each of these domains, internal representations for music appear to differ from those in the other two domains in certain respects. With the aim of stimulating discussion about what, in the most general sense, is meant by iconic and symbolic representations, a variety of experimental results on music perception and production will be reviewed. Various results suggest that representations for music are perceptual ("iconic") in nature. For example, listeners can reproduce with a high degree of accuracy expressive variations in performance timing, and recognize extracts of a piece of music in an unfamiliar style they may have heard only once. Or, to take another example, scanning a melody for its pitch contour produces similar reaction times and patterns of errors whether the melody is actually heard or imagined. However, other results suggest musical representations are more abstract or categorical. For example, musicians tend to show categorical perception of pitch and distort rhythms toward simple ratios of durations. Moreover, internal codes for pitch, time, and dynamics are strongly context-dependent; the perceived similarity of chords, for instance, depends on the prevailing tonality. Finally, various complex transformations, such as theme and variations, are appreciated by listeners, and memory errors show melodic and harmonic sequences are coded with respect to stylistic norms. Whether or not these representations are appropriately called "symbolic" is a matter for discussion.

A Symbolic View

Hector Levesque

University of Toronto
Department of Computer Science

One of the main driving forces behind the move towards non-logical, non-propositional, and non-symbolic representations is a computational one. Logic, it would appear, is too hard to compute. *But is this right? If we take the case of mental models for example,* rather than constituting a new type of representation, it appears that to a first approximation, these can be understood as symbolic propositional representations (i.e. sentences), but of a special and important kind. I have called collections of propositional representations "vivid" if they satisfy two restrictions: first, they are semantically complete, in that for any sentence (over a given vocabulary) either it or its negation is implied by what is represented; and second, they are syntactically simple, in that the collection of representations itself has the same overall structure as any logical model of the propositions represented. The main feature of these vivid representations is a computational one: it is possible to very efficiently calculate their logical implications, even for extremely large collections. Moreover, it appears that many quite different forms of robust, effective, domain-independent reasoning can be understood in terms of [illegible]tions or refinements on vivid representations (involving, for example, certain forms [illegible] unsound or incomplete reasoning). What this suggests is that it is not necessary to abandon a logical (*or semantical or truth-functional*) perspective on reasoning for computational reasons, anyway. The logic in logic textbooks may not compute, but the logic needed for knowledge representation could very well.

Perspective Is Free

Willem Levelt

Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands

There are, presumably, various codes in which the mind can communicate with itself. Apart from a much studied spatial mode of representation, there are the kinaesthetic codes that are so clearly displayed by young children, codes for rhythm and melody, systems of representation for culinary objects, for emotions, and so on. We can move from one code to another, dependent on the requirements of the task at hand. There is no reason to suppose that there is a single dedicated system that is there to mediate among them.

However, as soon as any of these information types is to be expressed in language, it must be translated into propositional form. Or rather, it must be propositionalized. Full translation is typically impossible. No description of a face will do full justice to the visual image of that face (as novelists know all too well), and similarly for the other modes of representation. Speakers approach this problem by TAKING PERSPECTIVE. Here is an example:

Subjects were asked to describe the following pattern (or rather, the color terms in the pattern below were colored nodes in the experiment):

```
BLUE -- PINK -- RED -- YELLOW -- GREEN
                 |
                 |
                GRAY
```

One male subject's description ran as follows:

(1) Begin in the middle, a gray node. From there upwards a red node. Then to the left a pink node from the red. Then from pink again to the left a blue node. Then back again to red. Then from red to the right a yellow node. And from yellow again to the right a green node.

One female subject's description went this way:

(2) I start at node gray. Go straight on to red. Go left to pink. Go straight on to blue. Turn around, go back to pink. Go back, uh straight on to red. Straight on to yellow. Straight on to green.

Now consider how the BLUE-PINK part of the pattern was described by these two subjects. The speakers agreed on their choice of PINK as the reference location for BLUE. In fact, none of the 51 subjects reversed this, as in "Pink is to the right of blue". There is nothing in the visual image that dictates this. It is entirely due to the linearization strategy that subjects choose in describing patterns such as these. We asked them to begin at GRAY. A major linearization strategy is to go over the pattern in connected fashion. So PINK will be mentioned before BLUE. If we had asked them to start the description at BLUE, BLUE would have become the reference location for PINK. Perspective taking is, to some extent, task dependent. But there is no way around it. SOME perspective has to be taken.

But the speakers also disagreed in an interesting fashion. The male subject used "to the left" to denote the spatial relation between BLUE and the reference location PINK, whereas the female subject used "straight on" to denote the same relation. How can "straight on" denote the same spatial relation as "to the left"? This is, again, a matter of perspective taking. The first subject describes the pattern in terms of "deictic" perspective, taking himself as the basis for the "gaze tour" he is making. When he moves his gaze to the left from one node to the next, he will use the term "to the left". If he moves his gaze to the right, he uses "to the right". The prove of this gaze moving theory (which I published in 1982, long before Roger Shepard re-invented it) is in denoting the direction away from the subject (or towards the subject). The pattern is flat on the table in front of the subject, but he denotes the direction from GRAY to RED by "upwards". There is nothing "upwards" in the pattern itself, but the gaze move is indeed upwards in going from GRAY to RED. More generally, I found that speakers who take a deictic perspective always use vertical dimension terms for directions in the pattern towards and away from them.

The second subject didn't make a gaze tour, but a body tour, as if she was walking or driving through the pattern (notice her mention of "turning around"). For every new move through the pattern, the previous imaginary body orientation is taken as the basis for directional coordinates. This basis reorients as moves are made. It is not a fixed deictic perspective, as in the previous case. The "straight on" from PINK to BLUE derives from just having driven from RED to PINK, and continuing the same direction of movement. And indeed, in her description this subject uses "straight on" for all cases where the same direction is continued. Notice also that, as it should be in a flat array of streets, there are no "ups" or "downs" for the body-tour subjects. Because the imaginary moving body adapts to the intrinsic line directions in the pattern, one might call the perspective taken by these subjects "intrinsic".

Speakers are free to take deictic or intrinsic perspective in propositionalizing this kind of patterns. But they have preferences (two-thirds deictic, one-third intrinsic). And those who have left-handers among their parents or siblings tend to prefer intrinsic perspective.

The example makes clear that there is no single NECESSARY way of assigning propositional format to a visual image. One would expect such necessity if the visual system's output were itself in propositional form. In that case the perspective inherent in the proposition would simply be forced upon the mind. There is, however, substantial freedom in putting the perceived structure, which is spatially represented, into one or another propositional form. They are all equivalent descriptions of the same perceptual pattern. But they differ in perspective.

This does not mean that perceptual factors play no role in taking perspective. They do. Speakers prefer to express figure/ground relations such that the ground is taken as reference. They tend to say "The cat is in front of the wall" rather than "the wall is behind the cat". What is smaller is preferably located with respect to what is larger. What is moving is preferably located with respect to what is stable. What is contained is preferably located with respect to what contains it. And so on. But none of these are obligatory. Perspective is free.

Relevant Literature

Levelt, W.J.M. (1982). Cognitive styles in the use of spatial dimension terms. In R.J. Jarvella and W. Klein (Eds.), "Speech, place and action: Studies in deixis and related topics". Chichester: John Wiley.

Levelt, W.J.M. (1984). Some perceptual limitations on talking about space. In A. van Doorn, W. van de Grind, and J. Koenderink (Eds.), "Limits of perception: Essays in honour of Maarten A. Bouman". Utrecht: VNU Science Press.

Levelt, W.J.M. (1989) "Speaking: From intention to articulation". Cambridge, Mass.: MIT Press.

Shepard, R.N. and Hurwitz, S. (1984). Upward direction, mental rotation, and discrimination of left and right turns in maps. Cognition, 18, 161-193.

Plausible Explanation-Based Learning

Gerald DeJong

University of Illinois
Computer Science Dept. & the Beckman Institute

I am not certain what an iconic mental model is. It might be defined narrowly and functionally: "That's the kind of mental model in connectionist systems", it might be defined as a generalization of the kind of representation advocated by the imagery side of the once-heated imagery/propositional debates, or it might be broadly defined as all representations that are not syntactic variants of predicate calculus. For the purposes of this discussion an "iconic" mental model will be taken to mean a representation which is a kind of *analog* of the corresponding object in our universe of discourse. A host of implicit relations hold between two iconic representations by virtue of the fact that the relations hold between the actual world objects and our representations are analogs of the real world. Thus, iconic representations can faithfully represent relations which the system's implementor had no knowledge of or intention to capture. We use this definition because A) it seems interesting, B) it is abstract enough to be defended, and C) it allows me to describe my work of the last year and a half as combining symbolic and iconic representations into a unified AI planning system.

Planning in AI has traditionally been viewed as discovering a sequence of actions that achieve a goal. Each action is the instantiation of one of a known set of operators which are defined as functions that map world states to world states. In the real world many changes (including actions) are continuous and cannot be easily or adequately modeled as a *seqence of dicrete mappings*. These changes are more naturally represented as curves or graphs rather than as symbolic propositions. Goals such as accelerating around a curve or performing a barrel role in an airplane involve gradual, continuous, and coordinated manipulation of several controls. The goal quantities are (rather indirect) functions of the control parameters and other world quantities. Planning in such domains consists in discovering the explicit function relating control parameters to goal quantities and inverting it. Then appropriate control values can be computed from known goal values.

My current research is a synthesis of ideas from Explanation-Based Learning, Qualitative Reasoning, and Numerical Analysis. Planning turns out to be very efficient, although the learning phases can be computationally intensive. Briefly, the system starts out with a substantial background knowledge of the world, represented symbolically in a qualitative reasoning formalism. As with many Explanation-Based Learning systems, it observes an expert solving problems. When the expert achieves a goal that is beyond the system's abilities, it constructs a qualitative, plausible explanation of how the expert's control manipulation may have resulted in the goal's achievement. The observation of each control is a graph of its values over time. The explanation is in qualitative terms (greater-than, less-than, increasing, decreasing, etc.) and is too abstract to be directly useful for planning. However, such an explanation identifies the relevant quantities and imposes a smoothness constraint. This permits the construction of a multi-dimensional interpolation function recording the value tuples observed in the expert's example. Through planning in the real world, more value tuples are recorded and the function's shape is refined. It can be shown that the linear-approximated shape converges on the true function's shape. The research, though still preliminary, promises contributions to planning and temporal reasoning, concept learning and intelligent adaptive control.

Modelling and Media

Pat Hayes

Xerox Palo Alto Research Center
Palo Alto, California

1. First, are we talking about external representations - marks on a page, things which are subject to perception - or internal representations, hypothesised as being involved in a cognitive theory of some kind - mental models? This is important because while in the first case the distinction between pictures and writing just seems obvious, is embodied in folklore, and is supported by all sorts of evidence, nevertheless in the second case the perceptual difference no longer applies, and so it is no longer quite so clear what the nature of the distinction is. Even in the first case of external representations, it is easy to find examples which seem to have both iconic and symbolic aspects. (Maps are perhaps the most striking, involving such "symbolic icons" as contour lines.)

2. Focussing on internal representations, computational modelling immediately suggests that what seems to be an iconic representation might itself be represented symbolically, and vice versa. A robot might represent its environment as a two-dimensional array: but the implementation of this array is a network of pointers, and these are symbolic encodings of the indexical relationships between the parts of the array which are carrying meaning. From this lower-level perspective , the representation can be thought of as a symbolic description of the room. Another example is provided by the characters in a piece of text which are considered symbolic, and whose internal encodings as bit-patterns are also symbolic, but these bitpatterns are themselves *represented* iconically as patterns of voltage in the architecture of the machine.

3. The traditional approach to describing the relationships between representations and *representees* is in terms of model theory. (More recently, Barwise and Etchemendy's semantics of situation theory has suggested an alternative, but one also oriented towards symbolic accounts of the information contained in a representation.) The key idea of this way of defining meaning is the relationship of satisfying, or making true, which can hold between something and its description. A symbolic representation partially constrains the possible things it could refer to by claiming that they must satisfy it. This immediately has some consequences for how a representation works: for example, if more information is added to it, then the set of possible satisfiers is made smaller, but this does not mean that they must themselves be larger or more complex. The relationship between an iconic representation and what it describes seems however to be different, in some sense one of homomorphism: the representation is like the representee (in certain respects) but is (usually) smaller or simplified. Thus on a map, the spatial layout of symbols on the page is a direct orthographic projection of the layout of the towns, roads and hills in the terrain being described. Here, adding information to the representation implies added structure in the world. There may be other ways in which a representation can talk about a domain, and a useful effort of this whole discussion would be to try to catalog some of them.

Depiction Theory and Mental Representations

Alan K. Mackworth

Canadian Institute for Advanced Research
Department of Computer Science
University of British Columbia

The relationship between iconic and symbolic representation systems can be understood within a logical theory of depiction (Reiter & Mackworth, 1989). Iconic representations, whether mental or physical, have a range of characteristics best exemplified by maps. A map exists in a medium, the image domain, but it depicts (represents) a scene domain. A map has a frame, a theme and a scale --- all of these exclude certain objects; everything not excluded must be depicted in the map. Conversely, everything appearing in the map must depict a scene object. The depiction relation may lawfully constrain topological and geometric properties between the image and scene domains.

Under the functional approach to knowledge representation the underlying medium ('icons'/'symbols') does not matter per se. What matters is the descriptive and procedural adequacy of the representation system, viewed functionally. What can be asked? What can be told? Is retrieval correct and complete? How efficient are these operations on particular architectures?

Consider three formalizations of the relationship between a symbolic representation and an iconic representation. It can be thought of as the *relationship* of a knowledge representation language to a restricted, less expressive sublanguage with more effective decision procedures. Or it can be that of a set of sentences in a logic to a logical model of those sentences. Or, finally, it could be the view couched in the language of the logical theory of depiction. A logical model of the descriptions of the image and the scene and the image, scene and depiction rules represents a consistent interpretation. By studying concrete issues in the representation of spatial knowledge in maps we can build a framework for answering the questions posed about the adequacy of alternative representation systems.

Mental Models

Drew McDermott

Yale University
Department of Computer Science

A common mental operation is to reason from premises to conclusion. For example, a creature might want to know what would happen in a certain set of circumstances, and one could model this kind of reasoning as inference from the premises *Suppose the circumstances were such-and-such* to the conclusion *The following would happen.* But of course it is of little value merely to verify that something would happen; we usually require that the inference technique answer the general question "What would happen?" An example of this kind of problem is this: Suppose someone left a can of gasoline in your fireplace. What would happen?

It is natural to model this inference in the form of a deduction. We would take the *circumstances to be represented* in as *neutral* a *form* as possible, and then use deductive algorithms to draw the conclusion. This model is especially appealing because the description of the given circumstances may be quite vague and counterfactual. In the example problem, I did not specify the exact location of the gasoline can, and you may not even own a fireplace. But this will not be a bug if the premises are represented as something like *(true(initially,in(can21,fireplace22)))* (with further descriptions of what *can 21* and *fireplace 22* are); and if the reasoning is mediated by rules like

$$\forall\ x,s_0,s_1(\mathtt{true}(s_0,\mathtt{in}(x,y)) \wedge \mathtt{contains}(x,\mathtt{gasoline}) \wedge \mathtt{after}(s_0,s_1)$$
$$\wedge\mathtt{true}(s_1,\mathtt{on\text{-}fire}(y)) \supset \mathtt{true}(s_1+\Delta,\mathtt{explosion}(y))$$

Unfortunately, there are a surprising number of obstacles to getting this to work:

1) Deduction techniques are much better at verifying conclusions than in generating them in the first place.

2) It has proven more difficult than one would expect to formalize seemingly simple forms of inference. Inference over time, in particular, suffers from the *persistence* and *ramification* problems: It is hard to deduce exactly what does and does not change over time.

3) Algorithms for doing deduction tend to take exponential time. They are not good candidates for what people are doing when they solve such problems.

These difficulties have caused researchers to turn to an alternative theory, the theory of *mental models.* This theory posits that when a conclusion is to be drawn from premises, the first step is to construct a detailed representation of a particular situation satisfying the premises. Then the question is answered by inspecting this "model." In some cases, more than one model gets constructed. In the example problem, one might set up a detailed map of a house and fireplace, and locate the gasoline can in a particular point in the fireplace (in the back, on the right). The model is not an inert tableau, but contains machinery for simulating what the real world does. For instance, if a fire is started in the simulated fireplace, the machinery might track the radiation and convection to determine what *temperatures are reached* in different parts of the fireplace. A simulation of gasoline would know to cause it to expand as the temperature goes up, and eventually to explode.

Not all mental modeling is simulation. There are other ways one might inspect a mental model, including counting things ("A red cube is cut into 27 smaller cubes; how many have two red faces?"), and planning ("How can I cause an explosion?"), and language understanding ("A crane fell on the car, and a man was trapped inside." -- Inside what?). But simulation seems like a natural ability to focus on.

The mental-models idea is introspectively attractive, and has much experimental support. (Johnson-Laird 1983) The idea rests on the observation that most interesting conclusions that follow from a single model of the premises will follow from every model. The fate of the gas can does not depend on its exact location, so we can put it anywhere. The advantage of putting it in a particular place is that simulation machinery requires it to be in a particular place.

In spite of its appeal, the theory raises several tough questions:

1) *How are models constructed?* One might suppose that the original purpose of mental models is to tell a creature what will happen in the *present situation.* Construction of a mental model would then *consist of carrying over whatever aspects* of the current situation seemed important. But for answering hypothetical questions, like the one about the gasoline can, we require the ability to set up a model incrementally as new premises come in. Johnson-Laird (1983) proposed that new premises are incorporated into the model until an inconsistency occurs, when the modeler backs up and generates a new model consistent with all the premises so far. McDermott and Davis (1984), in a spatial-reasoning system, attempted to represent the model vaguely enough that new premises would merely cause it to become less vague. Both approaches have difficulties. Of course, there is no evidence that humans can do a good job with a large number of intricate premises. We need a theory of what is feasible.

2) *What is simulated physics?* It is by now a familiar fallacy to assume that mental images automatically incorporate the physics of the objects they represent. In fact, the physics has to be designed and implemented. One idea is that the physics of mental models is just Newtonian physics, and simulation proceeds numerically, one tiny time step after another. (Gelsey and McDermott 1990, Gelsey 1989) That may make sense for computers with perfect knowledge of the geometry of a situation, but seems quite implausible for humans (most of whom don't use Newtonian mechanics). A model like that of Meltzer may be more appropriate: it's numerical, but abstracts away details. For instance, they model strings as a small set of connected elements, *and leave momentum out of the model completely.*

As you broaden the class of simulation methods, you may begin to doubt whether there is really any difference between simulation and forward deduction. Perhaps the crucial distinction between the two is that the mental-models approach involves adding arbitrarily chosen details to the premises to improve the efficiency of inference.

3) *How do you distinguish between relevant and irrelevant features of a model?* Suppose two children are sitting on a see-saw, one at the North end, one at the South end. Suppose one child is fat, the other thin. Which end will go down? The one with the fat child. Is the fat child facing South? We don't know; but to tell this we have to consider two models. Which child is most likely to have the sun in his eyes? Neither. Wait -- the Sun is in the South during the winter. Is it winter or summer? I guess the answer is really, Don't know. (And which hemisphere are we in, anyway?)

The simplest theory here is that we generate several models randomly and ask our questions with respect to each of them. The hope would then be that most interesting conclusions that follow from a randomly chosen sample of models will follow from most models. But that doesn't feel like what's happening in examples like the see-saw problem. It would be a pity if we had to appeal to a powerful deductive theory to explain how the mental models are constructed in the first place.

4) *How much do mental models explain?* Let's go back to the gas-in-the-fireplace example, where one immediately assumes that the relevant sort of episode is a fire. A deductive account would suggest that there's a rule: "The only interesting thing that happens reliably in a fireplace is a fire." A mental-models account would presumably have to posit a family living in the house, and simulate days or weeks of time (or months, if we start in the summer), until the family decides to build a fire. Very implausible.

If these two mechanisms have to coexist, what's the interface between them? Can the mental modeler call the symbolic inferencer at any time? Or does it run autonomously once started?

Direct Representations of Naive Physics

Bernard Meltzer

Artificial Intelligence Laboratory
Ispra, Italy

Classical physics and most contemporary work in qualitative physics uses representations in which objects and relations are represented by names, and the basic structure used is the application of a function name to one or more argument names. Another possibility is to represent objects and relations by other objects and relations, as in a map. Exploratory research on the usefulness of models of the latter kind for representing our intuitive, common sense knowledge of the behavior of the physical world is discussed, the ideas being illustrated for the cases of strings and liquids. These are represented by pixel sets on a visual display, and it turns out that qualitatively correct behavior can be obtained by using small numbers of local interaction constraint rules (four for strings and eight for liquids) of a very general character, such as those of causality, non-copenetrability, continuity, flexibility and liquidity. Limitations of the programs so far developed are discussed. Possible applications are briefly referred to.

Cognitive Artifacts

Donald A. Norman

University of California, San Diego
Department of Cognitive Sciences

A cognitive artifact is an artificial device designed to maintain, display or operate upon information in order to serve a representational function.

Artifacts pervade our lives, our every activity. The speed, power, and intelligence of human beings is dramatically enhanced by the invention of artificial devices, so much so that tool making and usage is one of the defining characteristics of our species. Many artifacts make us stronger or faster, or protect us from the elements or predators, or feed and clothe us. And many artifacts make us smarter, increasing cognitive capabilities and making possible the modern intellectual world.

My interest is in cognitive artifacts, those artificial devices that maintain, display, or operate upon information in order to serve a representational function and that affect human cognitive performance. I discuss three aspects of cognitive artifacts:

I. *Two differing Views of Artifacts:* The *System View* and the *Personal View*, in which I show that from the personal point of view, artifacts do not enhance ability - they change the task, and thereby, from the system's point of view, perform more powerful operations;

II. *Levels of Directness and Engagement:* The relationship between those aspects of artifacts that serve the execution of acts and those that serve the evaluation of environmental states and the resulting feeling of directness of control or engagement: successful artifacts give the user the impression of direct engagement with the task of interest;

III. *Representational Properties of Cognitive Artifacts:* The relationship between the system state and its representation in the artifact. I show the importance of the choice of representation and suggest two hypotheses on the naturalness and complexity of the mapping between the representation used in the artifact and the environment and an appropriateness principle about the choice of representation used by an artifact.

Adapted from Norman, D.A. (in press). Cognitive artifacts. In J.M. Carroll (Ed.), *Theory and Design in Human-Computer Interaction*.

Stanley Peters

Stanford University

To understand how people can give and get information using language, we need to understand how symbolic representations of information --the paradigmatic example being utterances in language -- connect with the ways minds retain information.

Neither iconic nor symbolic 'mental representations' can be eliminated in favor of the other, for reasons Etchemendy cogently presents. Iconic 'models' are very good at representing in one fell swoop many relationships which all hold simultaneously; icons are rather limited, however, at representing alternative possibilities of which at least one holds but not all. Symbolic 'models' are very good for representing such alternative possibilities; they are also very good at representing limiting conditions or unbreachable constraints, a task at which iconic 'models' are dismal. However, symbolic 'models' are deficient for representing large numbers of simultaneous relationships, especially infinite numbers of them.

If minds utilize both iconic and symbolic 'models', both sorts presumably play a role in the causation of behavior. This gives us an additional reason (besides understanding language use) to study how information flows between the two kinds of 'model'. This motivation is especially forceful because behavior is influenced only by explicit mental information, and not by information that is merely implicit in mental models. (N.b., explicit =/= conscious, and implicit =/= unconscious.) Research like Barwise and Etchemendy's provides a *promising opening wedge* for studying the interaction of iconic and symbolic models in minds.

Mental manipulation of iconic and symbolic models consists in neural activity. To understand minds' use of such models, it will help to learn how we can view brain-style computation as manipulation of them. While the models are to be found in our theoretical analyses of neural activity -- in the mind -- we should not expect to find them in the brain. A major task of integrating neuro-computing into cognitive science is in discovering how to view one and the same event as simultaneously (i) the performance of mental operations on mental models and (ii) activity of neuron-like computing elements.

Two Modes of Mental Representation

Steven Pinker

Massachusetts Institute of Technology
Department of Brain and Cognitive Sciences

What kind of categories do human concepts represent? "Classical" categories are defined by necessary and sufficient criteria that determine whether an object is in a category or not *in it (e.g., "grandmother" = "mother of a parent).* " A popular contemporary view is that concepts correspond instead to "prototype categories." Prototype categories lack necessary and sufficient conditions; their members need not be absolutely "in" or "out of" the category but can be members to greater or lesser degrees; their members display family resemblances in a number of characteristic properties rather than uniformly sharing a few defining properties; and they are organized around "prototypical" exemplars (e.g., stereotypes of a typical grandmother).

Alan Prince and I have discovered that this distinction in category structure can be found in an unusual place: the English past tense system. Irregular verbs fall into subclasses of similar-sounding forms (e.g., 'ring/rang', 'sing/sang', 'spring/sprang') that act like prototype categories: they lack necessary and sufficient conditions, have fuzzy boundaries, and a similarity structure often organized around a prototype. Regular verbs do not fall into such subclasses: any sound can be a regular verb, and (putting aside interactions with irregulars) all regular verbs are equally acceptable.

What is behind this parallelism? Prince and I suggest that there are two architectures underlying mental representations of classes: an associative memory, that represents objects in terms of their properties and that generalizes to new objects on the basis of their similarity to old ones, and a rule-based architecture, which defines lawlike formal systems in which entities are represented as symbols and in which global similarity is ignored. Irregular past tense forms are stored as sets of related items in an associative memory; regular past tense forms are generated by a rule of grammar and not stored. Similarly, prototype categories (e.g., of grandmothers, or birds) arise from superimposed exemplars stored in an associative memory; classical categories are the product of formal systems such as rules of kinship, folk science, law, and so on.

Each system is appropriate to different classes of objects in the world. Classes of similar objects that are the product of divergent or convergent evolution (linguistic, biological, or societal) have obscure causal histories but will display family resemblances that are best captured in an associative memory. Systems of entities that (in reality or in a suitable idealization) are governed by laws (natural or social) are best represented in formal systems.

Modeling Phonological Organization

Alan Prince

Brandeis University
Department of Linguistics

Phonological representation (PR) acts as an internal means of communication between various aspects of grammar (word structure, sound structure) and between grammar (knowledge) and behavior (action: articulation/perception). Time flow is discretized in PR as a *sequence* of segmental units; articulatory activity, however, occurs simultaneously in a number of independent streams, which are coordinated via their individual relationship to the segmental units. The segmental units are in addition the terminals of a hierarchical constituent structure -- segments group together into syllables, syllables group (typically as pairs) into 'feet', feet into (phonological) words, words into (phonological) phrases. These hierarchical constituents delimit domains in which phonological processes take place; they may also be recruited to define units out of which words are constructed by word-formation processes.

The articulatory streams and the levels of the hierarchy are shaped by a variety of characteristic conditions on their well-formedness, which can be seen to be widely operative in the world's languages. (A finer grain of analysis would distinguish *constellations of such conditions -- distinct types of organization.*) Such conditions can be imposed with greater or lesser degrees of intrinsic severity. (Example: some languages demand that all syllables contain a vowel; others can use certain consonants in place of the vowel; still others can use any consonant in place of the vowel.) Even so, it is typically the case that not all such conditions can be met simultaneously --the segmental composition of a word may not be perfectly suited for syllabification; the string of syllables may be not be perfectly suited for pairing into feet, and so on. In this case, strategies of looser matching and even representational modification are called upon to provide each word-form with a licit structure. This can be profitably seen as an optimization process, through which the competing claims of diverse constraints are resolved. The arena of this resolution is often the whole language; *it is never done ad hoc on a word-by-word basis;* but it is not uncommon to find cases where different lexical classes in a single language will adopt somewhat different strategies (e.g. nouns vs. verbs).

What Do We Need Nondiscursive Models FOR?

Zenon Pylyshyn

University of Western Ontario

There has always been uneasiness with the idea that all mental representations take the form of sentence-like symbol structures written in some calculus or mentalese. I share this uneasiness for many reasons, including certain processing complexity properties that look like they may come out wrong if we adopt a sentential format. The problem, however, is that no proposal I have ever seen for an alternative to *symbol structures (e.g. to datastructures with an embedding CONSITUENT structure)* has ever come close to being adequate. The assumption that there are pictorial "images", which are written in some medium that has Euclidean properties, is the most common proposal and the most clearly invalid.

The question that needs to be asked is: What criteria are we attempting to meet when we are driven to nonsymbolic or analogue representation assumptions, and can these not be met by symbol structures of a more conventional kind. The issue in the end comes down to what properties we are entitled to assume to be part of the functional architecture and what properties are encoded symbolically.

Some Speculations on Mental Models

Ray Reiter

University of Toronto
Department of Computer Science

While it is by no means clear to me what cognitive scientists mean by a mental model, I can think of one reasonable notion of what it might be, and a few properties that it ought to respect.

Why Mental Models

I take it that one use of a mental model is quick inference. Whatever else they may be good for, mental models should facilitate reasoning about suitable aspects of reality

Mental Models and Unsound Reasoning

I shall make the fundamental assumption that a mental model is indeed a model in the sense that logicians use that word Which means that it must be a model of something, namely a theory T describing some aspect of reality about which we want to reason. So the information of interest are some of the entailments of T. Normally, computing entailments of T is expensive, whereas computing truths in one model of T might be much cheaper. Since T may have lots of models, a true statement in one model need not be an entailment of T. This means that reasoning with a single model of T might be unsound. When invoking a mental model for reasoning, we are prepared to sacrifice soundness for computational efficiency by pretending that truths in the model are entailments of T.

Two Properties of Mental Models

The Whole Truth and Nothing But the Truth (Most of the Time)

Since reasoning with models is unsound (not all truths in the model need be entailments of T), we want to minimize the number of incorrect conclusions we can draw, i.e. we want to minimize the number of *accidental truths* of the model Let's focus on positive such accidental truths. Then the accidental truths for a predicate **P** in a model **M** will be minimized precisely when in no other model of T does **P** have a smaller extension **M** will minimize its accidental truths when this is true for all predicates simultaneously.

As it happens, this notion can be formalized for first order logic; it corresponds to McCarthy's predicate circumscription.

Population Control

A mental model ought not to contain more individuals than necessary. When reasoning about John and Mary, there should be just two individuals in the model.

When T is a universal theory of first order logic, the population of these models is the Herbrand universe of T. For arbitrary first order theories, the correct formalization is given by McCarthy's domain circumscription.

Some Issues

1) Do we really gain efficiency by restricting attention to minimal models? Computing truths in such models can be computationally unappealing (e.g. recent work in deductive databases). On the other hand, see comments below on the logic programming connection.

2) Relations to logic programming. The semantics of some aspects of the Prolog programming language is phrased precisely in terms of mental models as characterized above. In fact, one way of viewing Prolog is as an evaluator of formulas for truth in a minimal model, *not as a theorem prover.* This suggests that mental models need not be represented extensionally for computation purposes. Instead, it may simply be a byproduct of certain kinds of formula evaluators that they are computing truths in such models, without there being any explicit representation of such models available.

3) What are some *criteria* for guaranteeing soundness of this reasoning process? In some very interesting cases this is possible.

4) Relations to nonmonotonic reasoning in AI.

5) How do these ideas relate to Levesque's notion of vivid reasoning?

Seeing Straight

Whitman Richards

Massachusetts Institute of Technology
Department of Brain and Cognitive Sciences

Is that a straight line? To answer this question using a computer vision system we would check our pixel mapped image for the shortest distance between end points, after applying a correction for projection distortions. So where's the mental model for straightness? For the human visual system the task is more complicated. Firstly, we don't know the imaging distortions. Secondly, at the cortex the line will activate a disconnected, irregular, very non-straight array of neural elements. Unlike the computer, straightness in the world does not map simply onto the neural bit map. However, this difficulty can be easily circumvented by giving our visual device a reference for straight in the world, say the carpenter's "straight edge." Now it's easy to test whether the line is straight, simply by checking whether both the line's edge and the straight edge successively activate the same set (or subset) of neurons. Unfortunately, if we misplace our straight edge artifact, we're out of luck so it would be helpful to have an internal criteria for straightness. John Platt (1962) suggested a solution: if we simply move our eyes along a line, then if it is straight, the same neural elements will continue to be active. (Alternately, we could build symmetry into our cortical array and fixate end points - such coincidence schemes can check for a broader class of equivalences.) Platt's solution thus allows us to consider the mental model for straightness to be the presence of a procedure or routine which can carry out the test for straightness, namely an eye movement command plus a specific (coincidence) test applied to the cortical array. Because these operations include the external world in a closed loop, the model will in some sense be rooted in the world, and will not simply be tied to an arbitrary, unconstrained internal routine. Clearly, the I/o tools used in the procedure then can play a significant role in the construction of the representation. Geometrical or spatial concepts like parallel, symmetry, size, extent, occupancy, etc., most easily fit this notion of a mental model which is based on sensory-motor loops that include the external world. Concepts like "green," may not. What external operation grasps "green" like an eyeball follows an edge? Under this view, a mental model for "green" would require the ability to activate or select among our own internal feature maps, thereby losing a perceptual advantage of having the model tied explicitly to the external world. (In language or reasoning, this may be an advantage.)

Iconic-like mental models rooted in the world could provide useful information for indexing or constraining classes of more language-like models. For example, simple problems in geometry are often probed for solution by construction, before choosing a particular set of axioms and lemmas, etc. They also might serve as useful analogies upon which deeper, more abstract internal models are built - where the underlying assumption is that the analogical process is based upon the same class of lawful behaviors (e.g., models for fluid and current flow).

To summarize, as conceived here, the strong notion of a mental model for "x" is a sensory-motor procedure that when run, will test for the presence of "x" in the world (see Mackay, 1978; Ullman, 1984). Analogical models would then be built upon these strong, direct mental models. The procedure itself is the model, not its execution. When executed, the model's embodiment is delivering a premise about world structure (true or false) to an evaluator. The evaluator then attempts to propose a conceptualization of "what's out there," consistent with the models and the current sense data (Jepson and Richards, 1990).

Reflections on the Use of Mental Models in Soar

Paul S. Rosenbloom

University of Southern California
Information Sciences Institute

Mental models have recently been used in the context of the Soar architecture as the basis for a natural language understanding capability, and models of syllogistic and relational reasoning. Models show up in Soar as problem-space states of limited representational power -- no disjunction -- whose objects map one-to-one onto objects in the domain being modeled. Two models are used in natural language understanding: an utterance model, which models the linguistic structure (the syntax); and a situation model, which models the world being described (the semantics). To perform a task such as syllogistic reasoning, task instructions are first read. The situation model derived from this processing is a model of the behavior that should be exhibited. Then the syllogism is read. Here, the situation model is a model of the syllogism. Under direction of the behavior model, possibly valid conclusions are drawn from this situation model.

In this talk I will provide a brief overview of Soar's use of mental models, and reflect on what has been learned -- and what is still unclear -- from this about mental models in general, with a particular focus on their computational implications.

Reasoning

Stanley J. Rosenschein

Teleos Research
Palo Alto, California

During the past several years I have been exploring the subject of representation from an informational point of view, using models of information based on objective correlation. The guiding assumption in this work has been that a theory of mental representation should ultimately be based on an analysis of how the physical or computational states of one system (the agent) are systematically correlated with those of another (the environment) and how these correlations allow agents to track and react effectively to dynamic events. The correlations might be absolute correlations (e.g., whenever the agent is in state s1, the environment is in a state having property P), or they might be statistical (e.g., when the agent is in state s1, the environment tends to be in a state having property P). In either case, however, information is treated as an objective phenomenon.

As simple as this idea sounds, its mathematical embodiment has turned out to be quite useful as an analytic framework for studying issues in perception, representation, planning, and action, and has inspired practical techniques that have been applied to the synthesis of robotic control software.

By providing an objective account of information, the approach might also serve to ground discussions of specific informational encodings and classes of encodings. Information is implicit in the very fact that a correlation exists; details of the *encoding* of that information, however, are of great interest, especially insofar as they affect the computational complexity of the processes that update internal states in a correlation-preserving way and that map internal information states to overt actions. The iconic/symbolic distinction can be seen as one taxonomic cut on encodings--although one might question how productive the intuitive symbolic/iconic distinction will turn out to be and whether it will actually have any meaningful interpretation in a rigorous information-theoretic and computation-theoretic account of representation.

Some Thoughts About Mental Models

David Rumelhart

Stanford University
Department of Psychology

I have written at least two pieces on the idea of mental models in the human information processing system. The first of these appears in a chapter that Don Norman and I wrote which appears in Steven's Handbook of Experimental Psychology (the new edition, 1988). In this chapter we emphasize the importance of mental models as the medium for imagination and reasoning by imagining. The second was an attempt to fit mental models into the framework of connectionist model building. This is a section of a Chapter written jointly with Smolensky, McClelland and Hinton which appears in volume 2 of the PDP book (McClelland & Rumelhart, 1986). In this paper we argue that we should view the human information system as consisting of (at least) two distinct networks with two distinct functions. One network, called the interpretation net, has as its function the interpretation of current inputs and the specification of potential responses. The second network, called the "prediction net", or more simply the mental model, takes as input information about the current situation and a proposed action and produces as output the expected outcome of the action. This network is, then, a "model" of the world. Now, it is possible to remove the connections between the first network and the actual actions, but feed those inputs into the model and determine the predicted or hypothetical outcome of these actions. These predictions can then be fed into the interpretation network and thereby iterated and predictions can be made several steps in the future. This method can be shown to work in certain simple reasoning tasks and in certain learning tasks in which the mental model can be used to interprete the feedback from the environment.

Does Persistence Occur Outside the Mind of the Beholder?

Erik Sandewall

Linkoping University
Linkoping, Sweden

An *intelligent autonomous agent that moves and acts* in the real world must be able to deal with both qualitative and quantitative information about its environment (and about itself). The qualitative knowledge, by conventional wisdom, should use a conceptual structure that is reminiscent to what we find ourselves using, for example in natural language, and should use concepts and constructs such as "objects", "actions", "events", temporally dependent "properties" of objects, and so forth. The formal character of the conceptual structures may be as wff in a suitable logic, or as partial interpretations for logic formulas, or simply as high-level data structures according to taste.

It is equally clear that the quantitative knowledge must be organized according to the principles of classical engineering, with mathematics (especially calculus), physics, and automatic control engineering in successive layers providing both concepts and theory.

The distinction between the qualitative and the quantitative is almost the same as the distinction between *conceptual* and *physical* knowledge, respectively. The nature of the relationship between conceptual and physical knowledge, and the appropriate ways of describing that relationship, are the topics that I address in this presentation.

The difference between those two types of knowledge structures may be seen as a consequence of how we have acquired them. Physical knowledge structures have been developed from observations of the physical world. Conceptual knowledge structures have been developed from observations of observations of the physical world, for example observations of sentences in natural language, which in turn capture somebody's observations of the world.

The interesting issue is whether that distinction is relevant or not. The following two viewpoints are therefore compared:

1) Conceptual and physical knowledge are two different ways of making statements about the world, and should be compared side by side.

2) Physical knowledge describes the world directly; conceptual knowledge uses structures which are natural in the output from *perceptors of* the world.

If the latter viewpoint is adopted, then it becomes very meaningful to relate investigations into the "natural" design of perceptors, to investigations of actually occurring, conceptual knowledge structures.

I discuss some simple scenarios where these two viewpoints may be applicable. The first viewpoint is illustrated by characterizations, using a combination of temporal logic and differential calculus, of a simple "ball and shaft" example with a few variations. It is shown how the traditional, A.I.style persistence can be generalized into a principle of chronological minimization of discontinuities. (This part builds directly on my papers at KR89 and IJCAI89).

The second viewpoint is illustrated with a few simple perceptors, or "dynamic pattern recognizers". It is shown how persistence can then be *interpreted* as the *consequence* of a very natural design decision in a dynamic classifier.

My conclusion for the time being is that both of the above mentioned viewpoints may be fruitful, and in particular that the second viewpoint may contribute considerably to an understanding of the proper relationship between physical and conceptual knowledge.

George Sperling

New York University
Psychology and Neural Sciences

I would be coming to the workshop to learn. My own work hs been concerned with computational theories of visual perceptual processes (models of contrast detection, depth perception, and motion); with levels of representation in visual sensory and visual short-term memory, and with auditory short-term memory; with invariances in visual search and pattern recogniton tasks, with attentional processes, and with American Sign Language. I have never given sufficiently careful thought to the nature of the implied mental representations as opposed to their functional properties because, it seemed to me, that many different representations might share similar properties. It's the rules that determine the game, not the uniforms of the players. But, I would be coming to the workshop to be educated; hopefully it's not too late.

A Connectionist View

Paul Smolensky

University of Colorado, Boulder
Department of Computer Science

To be frank, whatever "position" I have on the relation of connectionist representation to the iconic/symbolic distinction is at this point unarticulated, and I approach the workshop expecting to take seriously the emphasis on "on-line thinking." I don't really have any idea what I'll say in my presentation, but it might involve developing the following line of analysis:

In relating iconic, symbolic, and connectionist representations, it seems useful to regard a particular *representation* as a *mapping from an indexing set to a value set*; the simplest iconic representations, for example, employ an indexing set which is a 2-D space and a value set of real intensity values. Viewing a representation as such a mapping makes it easier to be clear about certain similarities and differences among the three kinds of *representation*; e.g., typical iconic and connectionist representations both employ continuous value sets, in contrast to symbolic representations; on the other hand, symbolic and connectionist representations both employ discrete index sets, while some sort of geometrical structure on the index set seems an important aspect of iconic representations. Whether index and value sets have discrete, continuous, or geometrical structure is, I believe, quite important; e.g., it has a major influence on the type of mathematical operations it is natural to turn to for modeling cognitive processing.

Another kind of distinction that has been central to debates on the relation between connectionist and symbolic processing concerns compositional or constituent structure. Despite the claims of several authors (including myself) I believe there does not currently exist an adequate characterization of this notion, although the above mapping view of representations suggests a new line of attack: in symbolic, but not iconic or connectionist, representations, the value set and the indexing set are (essentially) the same. If an adequately enlightening presentation of the compositionality controvery appears possible, I may attempt to summarize the state of play in this debate.

Phonology as a Meeting Place for the Iconic and the Symbolic

David S. Touretzky

School of Computer Science

Carnegie Mellon University

Recently there has been a flurry of work on connectionist approaches to modeling phonology. I see this as an encouraging development, because phonology is a simple enough domain to be tractable, yet rich enough to raise many of the important issues associated with more complex forms of rule-following behavior. Vowel harmony processes in particular have been receiving widespread attention. Lakoff's theory of "cognitive phonology," based on a proposal of John Goldsmith's, models iterative processes such as vowel harmony as a parallel constraint satisfaction task, perhaps using simulated annealing. Subsequently, Touretzky and Wheeler showed how harmony could be modeled by a perceptual clustering mechanism using just simple feed-forward connectionist circuitry. Other recent approaches to harmony have involved sequential recurrent backprop nets, notably Gasser and Lee's model of Turkish, and Hare's model of Hungarian. These last two models only address certain qualitative aspects of harmony; they do not purport to be general models of phonological behavior.

Phonology may at first appear to be a purely symbolic process, but closer examination reveals several iconic aspects. For example, the syllable structure of words is based on a wave-like (rising and falling) sonority pattern: the peaks are vowels, which form syllable nuclei; the troughs are the coda of the preceding syllable followed by the onset of the next. Another *iconic aspect of phonology has to do with timing constraints*. Port has shown that Japanese speakers make compensatory adjustments to phoneme pronunciation times in order to keep overall word duration proportional to the number of moras, despite the fact that individual moras may be of different lengths. The process has a parallel constraint satisfaction flavor to it, and seems more iconic than symbolic.

Articulatory constraints, which motivate many phonological processes, are yet another source of iconic intrusions into the symbolic domain. Phonology appears to be a place where the iconic and the symbolic meet. Connectionist modeling may help us understand this interface.

Representations In Mental Models

Anne Treisman

University of California, Berkeley
Department of Psychology

My approach to the problem is through the initial perceptual encoding of information. Questions that interest me are how soon and in what ways does the visual information get transformed into a symbolic form and how different (if at all) is the final representation of a visual scene from the representation that underlies imagery and the representation that underlies language and reasoning. Although conscious experience, both in imagery and in perception, appears to have analog qualities, it may depend partly on categorical codes. For example, illusory *conjunctions* of different features suggest that percepts may be synthesized from recombined codes (like color x, size y, orientation z), which generate spatial extents and configurations for a perceived color that differ markedly from those actually presented. Shapes can be generated in different surface media; for example, discontinuties in luminance, color stereoscopic depth, relative motion or texture can all define edges with particular orientations, curvature, etc. Behavioral studies of selective adaptation and of interference in search may reveal whether the coding of spatial properties is carried out on a *common* abstracted representation, pooling information from all these media, or separately within specialized modules coding the different surface properties. Asymmetries of feature coding between *reference* values and deviations may parallel asymmetries in language (marked vs. unmarked dimensions) and in judgments of similarity of abstract concepts. Thus, studies of perception may *reveal properties of functional* organization that recur in other cognitive domains. Finally, together with Daniel Kahneman, I have been collecting evidence relating to the distinction between perceptual *types and* tokens. *We study how* object-specific representations are established and how information is collected, integrated and updated as the objects move and change.

Pictorial and Symbolic Representations In Object Recognition

Shimon Ullman

Massachusetts Institute of Technology

The distinction between "pictorial" and "symbolic" representations in vision is difficult to define precisely. *In the area of object* recognition it seems clear, however, that some representational schemes (e.g., ones that correlate 2-D images) are more pictorial, while others (e.g. schemes that use primitive categories for parts and spatial relations) are *more symbolic in nature*. Theories in A.I. and psychology have tended to be closer to the symbolic end of the spectrum. In this talk, I will describe an approach to object recognition, called "the alignment of pictorial descriptions". I will compare this scheme with more symbolic approaches, and examine the view that the recognition of specific objects is more pictorial, compared with general classification and categorization that may be more symbolic.

Deictic Reference and Mental Models

Bonnie L. Webber

University of Pennsylvania
Department of Computer & Information Science

In discussing symbolic versus iconic representations in *mental models*, my interest is on the roles such models play in defining the *context* in which language is understood -- in particular, their role in providing referents for the definite noun phrases (NPs), pronouns, and names used in a discourse. It is well-known that the referents of such terms can change easily over a discourse, as the referents of *the lock* and *the key* do in the following example -

Example 1:

> Just as I was about to get into my car this morning, I noticed that the tires looked soft. When I tried to open the trunk to get my tire gauge, I found that *the lock* was frozen so *the key* wouldn't work. Luckily I remembered that I had another gauge in the shed out back. But when I got there I found I had forgotten *the key*, so I couldn't get *the lock* open.

It is reasonable to suppose that such changes in reference follow from changes in context. (Over fifteen years ago, Steve Isard suggested that change was context's most computationally interesting feature [1].)

One aspect of this enterprise may be of particular interest, given the theme of this workshop. It involves a characterization of context and context change that can explain the interpretation of *demonstrative* (or *deictic*) expressions in discourse -- in particular, cases as in Examples 2-5 below, where the deictic pronouns and NPs are clearly not interpretated with respect to any shared spatio-temporal environment of the speaker and hearer.

Elsewhere [3,4] I have argued two points: First, there are deictic pronouns in written text that appear to refer to the interpretation of one or more clauses. I have argued that such referents are constrained to the interpretations of discourse segments on the *right frontier* of an evolving embedding structure representing all or part of the previous discourse. Which of several possible interpretations a pronoun will take as its referent will depend on what is predicated of it, as in the different interpretations of *that* in the following two examples, both of which take their interpretations from the same preceding text:

Example 2:

> Segal, however, had his own problems with women: he had been trying to keep his marriage of seven years from falling apart; when *that* became impossible ...

Example 3:

> Segal, however, had his own problems with women: he had been trying to keep his marriage of seven years from falling apart; when *that* became inevitable ...

If the mental model derived from a text is to function as the locus of all its referents, then I would claim that *polymorphism* is a necessary feature of mental models clausal interpretations must be both *structure* and *individual*. Note that these features of ambiguity and polymorphism appear to be no different from the problems in interpreting pointing gestures noted by many years ago Quine [2].

Secondly, I have argued that demonstrative NPs in written text must refer to something that is actually *present* in the mental model. This is in contrast with definite NPs, which can refer to entities simply *associated* with something in the model, as in

Example 4:

a. Some files are superfiles.
b. To screw up someone's directory, look at *the files*.
c. If one of them is a superfile, delete it.

Example 5:

a. Some files are superfiles.
b. To screw up someone's directory, look at *those files*.
c. They will tell you which of his files is absolutely vital to him.

Here most informants have said that they understand the definite NP *the files* in Example 4 as referring to the files in the person's directory, while they understand the demonstrative NP *those files* in Example 5 as referring to the files that are superfiles (the ones explicitly introduced in 4a and 5a). Again, this feature of "being there" in the model that deictic reference requires may be understood either spatially or symbolically.

Finally, note in neither of these aspects of deictic reference in text does the *type* of text matter: deictic expressions work equally well in texts that describe spatio-temporal scenes as in texts that present formal arguments.

References

[1] Isard, S. Changing the Context. In E. Keenan (ed.), *Formal Semantics of Natural Language*, Cambridge: Cambridge Univesity Press, 1975.

[2] Quine, W. The Inscrutability of Reference. In D. Steinberg and L. Jacobovits (eds.), *Semantics: An Interdisciplinary Reader*. Cambridge: Cambridge University Press, 1971. pp.142-154.

[3] Webber, B. Discourse Deixis: Reference to Discourse Segments. *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, Buffalo NY, 1988.

[4] Webber, B. Deictic Reference and Discourse Structure. *Language and Cognitive Processes*, to appear.

4.0 Mental Models Meeting: Evaluation Summary*

Using a questionaire handed out to all participants, the meeting was evaluated in five categories:

1. Quality of meeting
2. Degree of cross-discipline communication
3. Scientific value
4. Potential impact on Future Research
5. Cost effectiveness

1. Quality and Uniqueness of Meeting

The aim was to provide an intense, high-level interaction and exchange not readily available in present meetings or workshops.

	Percentile (Median)	Range
a) In comparison with other workshops or satellite meetings in the fields of Cognitive Science, Linguistics, AI, or Experimental Psychology, in what percentile does this Study-Session fall with regard to the intellectual content and discussion level?	80	40-95
b) Was the intellectual level of the meeting set solely by the stature of the participants, or did the format and topic flow provide an important role?	Particpants 80%	n/a
c) Was there a sense that the participants experienced a meeting environment different from that currently available to them?	75	60-90
d) Were the subfields adequately represented? Which might be strengthened?	OK, except connectionists	

2. Degree of Cross-Discipline Communication

One of the major goals of the proposal for similar interdisciplinary meetings is to foster communication between disciplines that otherwise would not interact. To what degree was this meeting successful in this regard?

	Percentile (Median)	Range
a) What percent of the participants knew one another or had read a paper of the participant?	50	30-90
b) Did discussions occur across disciplines rather than within?	across: 50%	n/a
c) Were there profitable or unexpected interactions between disciplines during the breaks, at meals or "after hours"?	30 (Too few)	n/a
d) Were there indications that some participants were alerted to work or ideas they otherwise would have been ignorant of, or, that after departing, would read a paper that they otherwise would not?	70	10-80
e) Did the speakers make an effort to describe their subjects in a way different from what they would choose to address an audience from their own discipline?	40	30-70
f) Did the speakers respond well to questions from individuals outside their discipline?	50	50-75

3. Scientific Value

A successful interdisciplinary meeting should be able to report a new result or an advance in the field as a whole, as well as advancing each individual's knowledge.

	Percentile (Median)	Range
a) Did the meeting add any knowledge to the field? Did the discussions and exchanges lead to an insight, partial solution, or hint toward a solution not previously available to the scientific community?	60	10-80
b) Was there a consensus among the participants that they each learned something of value they otherwise might have missed?	30	10-50

4. Potential Impact on Future Research

	Percentile (Median)	Range
This issue interacts with Scientific Value. If there is a communal advance in knowledge, then the significance of this advance can provide a crude measure of impact. For an individual, however, the key question is: was a participant's research direction altered by the meeting? If so, was it a major change, a mid-course correction, or simply an inclusion that otherwise might not have taken place?	50 (mostly minor changes in direction)	0-80

5. Cost Effectiveness

Are the benefits of such interdisciplinary meetings worth reallocation of present funds, either those of the researcher or of the supporting agencies, or both?

	Percentile (Median)	Range
a) What percent of currently attended meetings yield the same impact upon a participant's research as this meeting did?	20	5-25
b) What is the estimated lead time for participants whose research was impacted?	6 months - 1 year	

6. Summary

Although there was general agreement that the quality of the meeting was quite high for everyone (top 80%), there were very significant differences among individuals in the impact the meeting had upon their perspectives and research. Similarly, there were marked individual differences in benefitting from the interdisciplinary backgrounds of the participants. Reasons for the failures to better communicate across disciplines are discussed in the body of the report.

*Compilation based on 25% responses to formal survey, plus informal responses.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a REPORT SECURITY CLASSIFICATION
Unclassified

1b RESTRICTIVE MARKINGS

2a SECURITY CLASSIFICATION AUTHORITY

2b DECLASSIFICATION/DOWNGRADING SCHEDULE

3 DISTRIBUTION/AVAILABILITY OF REPORT
Approval for public release:
distribution unlimited

4 PERFORMING ORGANIZATION REPORT NUMBER(S)

5 MONITORING ORGANIZATION REPORT NUMBER(S)
AFOSR-TR- 90 1127

6a NAME OF PERFORMING ORGANIZATION
Mass. Inst. of Tech.

6b OFFICE SYMBOL *(If applicable)*

7a NAME OF MONITORING ORGANIZATION
Air Force Office of Scientific Research N/L

6c ADDRESS *(City, State, and ZIP Code)*
Dept. Brain & Cognitive Sciences
79 Amherst St. E10-120
Cambridge, MA 02139

7b ADDRESS *(City, State, and ZIP Code)*
Bldg. 410
Bolling Air Force Base
Washington, DC 20332-6448

8a NAME OF FUNDING/SPONSORING ORGANIZATION
AFOSR

8b OFFICE SYMBOL *(If applicable)*
NL

9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
AFOSR-
90-0177

8c ADDRESS *(City, State, and ZIP Code)*
Same as 7b

10 SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO	PROJECT NO	TASK NO	WORK UNIT ACCESSION NO
61102F	2313	A4	

11 TITLE *(Include Security Classification)*
Representations in Mental Models

12 PERSONAL AUTHOR(S)
Whitman Richards

13a TYPE OF REPORT
Final Technical

13b TIME COVERED
FROM 3/90 TO 9/90

14 DATE OF REPORT *(Year, Month, Day)*
90 Sept. 14

15 PAGE COUNT
30

16 SUPPLEMENTARY NOTATION

17 COSATI CODES

FIELD	GROUP	SUB-GROUP
05	09	

18 SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
AI, Cognition, Data Structures, Knowledge Representation, Language, Mental Models, Neural Nets, Perception, Reasoning

19 ABSTRACT *(Continue on reverse if necessary and identify by block number)*

On March 12-13 an interdisciplinary group of thirty-five, composed of computer scientists, experimental psychologists, linguists, philosophers and "connectionists" met to share views on representations and their role in mental models. Although at least two books and several papers directly address these issues, the nature of mental models is far from clear. The meeting shed some light on why "understanding" mental models is difficult. Simply put, the reason is that mental processes are described in many different ways and at quite different levels of abstraction, depending upon the researcher. For example, some emphasize the cognitive properties of mental models, whereas others are more concerned with the internal data structures. Still others may stress the logical form and content of the mental process, as contrasted with the actual computational machinery. The diversity of these viewpoints is clear upon reading the abstracts prepared by the participants. Further study is needed to

20 DISTRIBUTION/AVAILABILITY OF ABSTRACT
□ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT □ DTIC USERS

21 ABSTRACT SECURITY CLASSIFICATION
Unclassified

22a NAME OF RESPONSIBLE INDIVIDUAL
Alfred R. Fregly, PhD

22b TELEPHONE *(Include Area Code)*
(202) 767-5021

22c OFFICE SYMBOL
NL

83 APR edition may be used until exhausted
All other editions are obsolete

SECURITY CLASSIFICATION OF THIS PAGE

*U.S. Government Printing Office: [illegible]

examine how these diverse viewpoints fit together into a useful, integrated framework.

The difficulty the group had in sharing a common framework for "What Is a Mental Model" poses a challenge for Cognitive Science. The wide diversity of viewpoints and approaches stems in part from the fact that different cognitive tasks often create different classes of problems which generally require quite different tools and expressions for solution. Thus, each subdiscipline within Cognitive Science has its own, quite different "Mental Model" that provides the framework for study. Again, although dramatic at the meeting, this wide spectrum of approaches is also visible upon perusing the prepared abstracts. This diversity suggests the need for an investment in exploring information flow and inference across representations.

The group included the following:

E. Charniak
E. Davis
G. DeJong
J. Etchemendy
P. Hayes
G. Hinton*
J.J. Hopfield
A. Jepson*
M. Johnson
M. Jordon
P. Johnson-Laird
R. Kaplan
S. Kosslyn
C. Krumhansl
H. Levesque
W. Levelt
A. Mackworth
D. McDermott
B. Meltzer
D. Norman
S. Peters
S. Pinker
A. Prince
Z. Pylyshyn
R. Reiter
W. Richards
P. Rosenbloom*
S. Rosenschein
D. Rumelhart*
E. Sandewall
G. Sperling
P. Smolensky
D. Touretzky*
A. Treisman
S. Ullman
B. Webber

*absent