Lars-Erik Janlert

STUDIES IN KNOWLEDGE REPRESENTATION

Modeling change - the frame problem
Pictures and words
STUDIES IN KNOWLEDGE REPRESENTATION
Modeling change - the frame problem
Pictures and words
Report UMINF-127.85, ISSN 0348-0542

Lars-Erik Janlert
October 1985

Institute of Information Processing
University of Umeå, S-901 87 UMEÅ
Sweden

Akademisk avhandling som med tillstånd av Rektorsämbetet vid Umeå universitet framläggs till offentlig granskning för avläggande av filosofie doktorsexamen onsdagen den 13 november 1985 kl 10.00 i hörsal F, Humanisthuset, Umeå universitet.

ABSTRACT

In two studies, the author attempts to develop a general symbol theoretical approach to knowledge representation.

The first study, Modeling change - the frame problem, critically examines the - so far unsuccessful - attempts to solve the notorious frame problem. By discussing and analyzing a number of related problems - the prediction problem, the revision problem, the qualification problem, and the book-keeping problem - the frame problem is distinguished as the problem of finding a representational form permitting a changing, complex world to be efficiently and adequately represented. This form, it is argued, is dictated by the metaphysics of the problem world, the fundamental form of the symbol system we humans use in rightly characterizing the world.

In the second study, Pictures and words, the symbol theoretical approach is made more explicit. The subject is the distinction between pictorial (non-linguistic, non-propositional, analogical, "direct") representation and verbal (linguistic, propositional) representation, and the further implications of this distinction. The study focuses on pictorial representation, which has received little attention compared to verbal representation. Observations, ideas, and theories in AI, cognitive psychology, and philosophy are critically examined. The general conclusion is that there is as yet no cogent and mature theory of pictorial representation that gives good support to computer applications. The philosophical symbol theory of Nelson Goodman is found to be the most thoroughly developed and most congenial with the aims and methods of AI. Goodman's theory of pictorial representation, however, in effect excludes computers from the use of pictures. In the final chapter, an attempt is made to develop Goodman's analysis of pictures further turning it into a theory useful to AI. The theory outlined builds on Goodman's concept of exemplification. The key idea is that a picture is a model of a description that has the depicted object as its standard model. One consequence is that pictorial and verbal forms of representation are seen less as competing alternatives than as complementary forms of representation mutually supporting and depending on each other.
STUDIES IN KNOWLEDGE REPRESENTATION
MODELING CHANGE - THE FRAME PROBLEM
PICTURES AND WORDS

Lars-Erik Janlert
October 1985
In two studies, the author attempts to develop a general symbol theoretical approach to knowledge representation.

The first study, *Modeling change - the frame problem*, critically examines the so far unsuccessful attempts to solve the notorious frame problem. By discussing and analyzing a number of related problems - the prediction problem, the revision problem, the qualification problem, and the bookkeeping problem - the frame problem is distinguished as the problem of finding a representational form permitting a changing, complex world to be efficiently and adequately represented. This form, it is argued, is dictated by the metaphysics of the problem world, the fundamental form of the symbol system we humans use in rightly characterizing the world.

In the second study, *Pictures and words*, the symbol theoretical approach is made more explicit. The subject is the distinction between pictorial (non-linguistic, non-propositional, analogical, "direct") representation and verbal (linguistic, propositional) representation, and the further implications of this distinction. The study focuses on pictorial representation, which has received little attention compared to verbal representation. Observations, ideas, and theories in AI, cognitive psychology, and philosophy are critically examined. The general conclusion is that there is as yet no cogent and mature theory of pictorial representation that gives good support to computer applications. The philosophical symbol theory of Nelson Goodman is found to be the most thoroughly developed and most congenial with the aims and methods of AI. Goodman's theory of pictorial representation, however, in effect excludes computers from the use of pictures. In the final chapter, an attempt is made to develop Goodman's analysis of pictures further turning it into a theory useful to AI. The theory outlined builds on Goodman's concept of exemplification. The key idea is that a picture is a model of a description that has the depicted object as its standard model. One consequence is that pictorial and verbal forms of representation are seen less as competing alternatives than as complementary forms of representation mutually supporting and depending on each other.
ACKNOWLEDGMENTS

First and foremost, I thank my friend Bo Dahlbom for his constant readiness to read, comment, and discuss what I write. His comments are always interesting, often unexpected, and I value them highly.

I thank my thesis supervisors, Axel Ruhe and Per-Åke Wedin, for their support and encouragement, and their successful efforts to provide me with favorable working conditions.

I am grateful to Martin Edman and Eva Ejerhed for their long and strong support, their good advice and helpful criticism.

I am grateful to Aaron Sloman for his generous and stimulating comments on a draft version of the section about himself in the pictures and words paper.

There are a few persons whose appreciation of my paper on the frame problem gave me the courage necessary to go on with my work: Daniel Dennett, John Haugeland, Patrick Hayes, Zenon Pylyshyn.

Several others gave me valuable reactions on a draft version of the frame paper. Among them: Peter Gärdenfors, Kenneth Kahn, Staffan Löf, Ingemar Widegren.

I have received financial support from: The Faculty of Mathematics and Natural Sciences, University of Umeå; The Swedish Humanities and Social Sciences Research Council (project 'Modeling time'); The Bank of Sweden Tercentenary Foundation (project 'Natural communication with computers').

Many persons have helped me in various practical matters. Thank you!

My family - Gunilla, Magnus and Niklas - has endured this ordeal undaunted. Well done!
# TABLE OF CONTENTS

Acknowledgments........................................iv

INTRODUCTION AND SUMMARY..........................vii
  Knowledge representation............................vii
  Modeling change - the frame problem................ix
  Pictures and words....................................xii
  Concluding remarks..................................xiv

MODELING CHANGE - THE FRAME PROBLEM..............1

Part 1: DEFINING THE PROBLEM..........................3
  Representation and problem solving...................4
  The non-deductive approach..........................6
  The deductive approach................................8
  The original frame problem..........................11
  The general frame problem...........................13
  The prediction problem...............................15
  The general book-keeping problem.....................21
  The qualification problem............................23

Part 2: SURVEY OF SUGGESTED SOLUTIONS..............25
  Frames................................................25
  Causal connection....................................28
  Consistency..........................................30
  STRIPS..............................................32
  PLANNER...............................................37
  Unless...............................................41
  TMS..................................................45
  Some related ideas...................................50

Part 3: DISCUSSION AND CONCLUSIONS..................53
  Metaphysics...........................................53
  The situation concept................................57
  Explicit, Primitive, and Primary.....................58
  Intrinsic representations............................62

Acknowledgments.......................................65
References.............................................67
A cornerstone of artificial intelligence (AI) and cognitive science is the assumption that intelligent systems depend on having and manipulating internal models or representations of their task environment, their problem world (or just world, for short). "Environment" is here to be taken in a broad and abstract sense: a problem world may be rather remote from everyday experiences, it may be a toy world, or an imaginary world. A system which is able to perform some form of action in its problem world, is clearly to some extent part of its own environment, which means that also parts of the system itself must be represented.

Knowledge representation

The study of such representations is the subject of knowledge representation (KR) - a subfield of AI with close connections to cognitive science. There is no general agreement on why it is called "knowledge" representation, and what exactly is to be understood as represented.\(^1\) It would seem that a right representation is knowledge\(^2\) to a competent user, which would mean that "knowledge" in "knowledge representation" is superfluous, and could safely be dropped. But, perhaps it is unwise to let any representation

---

1) See SIGART Newsletter #70, 1980.
2) Or true belief; AI is - so far - not very concerned about this distinction, which has baffled philosophers for so long.
qualify as knowledge representation. Anyway, the present studies were made on the assumption that representation is a relation between symbols and world, and that knowledge about representation in general can help us to understand knowledge representation better, whatever the difference might be.

Although an impressive amount of work is being done in knowledge representation in the form of various powerful description languages and modeling mechanisms, theoretical progress has not really met expectations. There are quite a few specialized microtheories, some remarkably successful, but there are almost no generally accepted unifying principles, and only isolated attempts to develop general theories of knowledge representation encompassing a sizable part of the field.\(^3\)

Of course, it is too soon to judge the feasibility of general theories of knowledge representation. Meanwhile, eclecticism and lack of unifying theoretical approaches, may be taken simply as indications that the fundamental questions of knowledge representation are much more difficult than they were first thought to be. On the positive side, we have several generally applicable techniques, several concepts and questions that appear to be pertinent over the entire field of knowledge representation, and certain universal problems. These are signs that there may be underlying unifying principles; for example, a universal problem suggests that there may be a common solution, or hopefully at least a common way of approaching particular solutions.

\(^3\) Far from being a theoretical breakthrough in artificial intelligence, the current success of so-called "expert systems" is the breakthrough of large scale application of artificial intelligence techniques: the breakthrough of a new style of programming language, a new style of programming methodology, and a new area of computer applications.
I have chosen one general problem - the frame problem - and one general question - the distinction and choice between pictorial and verbal representations - for closer studies; primarily in the hope of shedding some new light on the issues, but also - on the long view - as a way of approaching the underlying principles of knowledge representation.

Next follows brief summaries of my paper on the frame problem and my paper on the pictures and words issue, and some general concluding remarks. Readers who prefer to be kept in suspense should now skip to page 1 (or page 73).

Modeling change - the frame problem

Outline

The paper consists of three parts. Part 1 gives an introduction to the frame problem, its role, origin, and consequences. The frame problem is given a more precise definition and distinguished from several related problems which are also identified and discussed.

Part 2 contains a critical survey of suggested solutions which covers: frames, causal connection, consistency, STRIPS, PLANNER, Unless, TMS, and a few other related ideas.

Part 3 discusses some important points and concepts that have emerged in the preceding parts, and draws conclusions from the results of part 2 as to how the frame problem should be approached.
Summary

The frame problem in artificial intelligence can be described, roughly, as the problem of representing change. In the deductive approach to problem solving in AI, axioms are used to characterize the problem world, and the result of performing a given action in a given situation (state of affairs) can be deductively inferred with the help of axioms describing the effects of various actions. Not only the changes effected must be deducible, but also what doesn't change. What makes this a problem is that so much more is logically possible than what is empirically possible. Almost everything doesn't change, and the axioms necessary to describe what doesn't happen literally drown the axioms describing the empirically relevant effects. The problem-solver becomes bogged down with calculations of non-effects.

This is only one manifestation of a problem which has turned out to be a serious obstacle, not only to the deductive approach to problem solving, but to any kind of system that is to model a complex changing world. The attempts to solve the general frame problem has so far resulted in little more than pushing it around, and rephrasing it in various ways. A closer analysis reveals a conglomerate of closely related but distinct problem, and leads to a more precise definition of the frame problem. This definition also gives an indication of how a solution could be found: the frame problem is a problem of choosing the suitable form of representation, a form which is consonant with the metaphysics, the ontology, of the modeled world. What this means is exemplified in a concluding discussion of explicit/implicit and intrinsicness.
Conclusions

The frame problem is the problem of finding a representational form permitting a changing, complex world to be efficiently and adequately represented. It has not yet been solved. It is a general problem of representation. The frame problem is a problem of modeling, not heuristics; it is a question of form, not content; the rightness of form depends on the problem-world, not only on computational considerations. The frame problem is distinct from the prediction problem, the revision problem, the qualification problem, and the book-keeping problem. The prediction problem concerns the instrumental adequacy of the representation. The revision problem concerns the routine post factum corrections necessary to keep an idealized representation in pace with a real world. The book-keeping problem concerns the realization of a particular representational form. The qualification problem concerns the adaptation of an idealized representation to a real world. The frame problem, in contradistinction to all this, concerns the metaphysical adequacy of the representation, the choice of suitable form of representation. The repeated failures to solve the frame problem is a consequence of a general tendency to overlook the importance of ontological considerations in artificial intelligence. In particular, the McCarthian situation concept appears to be an unsuitable basis for a metaphysics of change.
Pictures and words

Outline

The paper consists of six chapters. Chapter I gives a general introduction to the questions at issue.

Chapter II presents and discusses theoretical work on pictorial representation in artificial intelligence.

Chapter III presents and discusses theoretical work on pictorial representation in cognitive psychology, mainly in connection with the mental imagery debate.

Chapter IV presents and discusses philosophical theories of pictures. Special attention is given to the theory of Nelson Goodman. The chapter also contains a discussion of "analog".

Chapter V presents and discusses experimental studies of pictorial representation in artificial intelligence.

Chapter VI outlines a new theory of pictorial representation which complies with demands raised by computer applications and which relates pictorial representation to verbal representation.

Summary

Comparatively little research has been done on pictorial representation in AI. "Pictorial" should here be taken in a very broad sense, as non-linguistic, non-propositional, analogical, or "direct" (as it is sometimes called in AI) representation. The work that has been done - both theoretical and experimental - is quite encouraging; to name the most important contributions, there is Aaron Sloman as the theoretical pioneer, and H. Gelernter et al. and Brian V. Funt with two very interesting working
applications. Many questions remain unanswered. Even the most basic issues can hardly be said to be fully understood: - what is a pictorial representation? - what is the difference between pictorial and verbal (propositional) representation?

There are other disciplines with an interest in the distinction between pictorial and verbal. In cognitive psychology there is a debate on "mental imagery". In philosophy we find a number of different theories of how pictures represent. But in all this wealth of ideas and theories, it is still difficult to find a view of pictorial representation that is at once cogent, compatible with computer applications, and not too counter-intuitive. Among philosophers, Nelson Goodman's symbol theory is, I think, the most thoroughly developed theory of representation. Moreover his general pragmatic approach accords remarkably well with the aims, methods, and material constraints of artificial intelligence. Goodman's more specific theory of pictorial representation, however, is of little help to knowledge representation since, in effect, computers become excluded from the use of pictures. Building on Goodman's very important concept of exemplification, I have tried to develop his analysis of pictures further, turning it into a theory useful to AI. The theory thus outlined appears capable of development and fruitful in suggesting and understanding different ways in which pictures may be used. I have also tried to bring in some relevant discussions in philosophy of science, which I think has closer parallels to artificial intelligence than is commonly thought.
Conclusions

The distinction between verbal and pictorial representation is relevant to knowledge representation. A new theory of pictorial representation is needed: a theory which is compatible with, and preferably gives useful support for, computer applications, and which takes into consideration relevant results from artificial intelligence, cognitive psychology, and philosophy. Also the parallels between philosophy of science and artificial intelligence should be paid more attention to. A promising idea is to characterize a pictorial symbol as being a model of some description that has the depicted object as its standard model. It appears that, instead of taking pictorial and verbal representations as competing alternatives, a more fruitful approach would be to view them as complementary forms of representation mutually dependent on each other.

Concluding remarks

Of course, these two papers by no means mark the end of the research they report on. So far, the symbol theoretical approach to knowledge representation looks quite promising. Generally, I think that a general symbol theory could serve as a bridge between computer science (in particular AI and knowledge representation), philosophy, and philosophy of science, facilitating the exchange of knowledge between these areas. The organizational and the algorithmic aspects of systems of symbols rather than questions of informational adequacy seem to pose the most difficult problems in knowledge representation - problems which traditionally have received comparatively little attention in philosophy. The computational
and empirical constraints enforced on AI give a perspective on symbols which is largely new to philosophy. On the other hand, it is of course precisely in philosophy and philosophy of science that we find fully developed theories of symbols and representations, and a long tradition of thinking about symbols; moreover in such a philosophical matter as I think knowledge representation fundamentally is, it is necessary to acquire and use a philosophical approach.
Modeling change - the frame problem

(To appear in Zenon W. Pylyshyn, ed., The frame problem and other problems of holism in artificial intelligence, Ablex.)
This is an attempt to elucidate the so-called frame problem in Artificial Intelligence, roughly the problem of representing change. On the basis of an examination of suggested solutions to the frame problem, which makes up the main bulk of this paper, it appears safe to claim that the frame problem has not yet been solved. And, I regret to say, there is nothing like a complete solution to be found in this paper either. What I have to offer is a more definite notion of what the frame problem is, what would constitute a solution to it, and an indication of the general direction in which a solution should be sought.

Part 1 sets the stage for the critical survey in part 2. First, the original frame problem is introduced and put into a historical perspective. The general frame problem is then defined as the problem of finding a representational form permitting a changing, complex world to be efficiently and adequately represented. This definition permits us to single out the frame problem from a cluster of related problems: the prediction problem, the revision problem, the qualification problem, and the general book-keeping problem.

Part 2 contains the survey of solution attempts: frames, causal connection, consistency, STRIPS, PLANNER, Unless, TMS, and a few other related ideas.

Part 3 sums up the examination by discussing some important points and by drawing some general conclusions: McCarthy's situation concept is an inappropriate basis for a metaphysics of change; the frame
problem is the problem to find the metaphysics of the modeled world and make it the metaphysics of the modeling system, which requires the metaphysics to be intrinsically represented.

Representation and problem solving

The frame problem is a general problem of representation, relevant to all types of set-ups where a complex and changing world is to be represented. Historically, the frame problem is intimately linked to the research on problem-solving and planning\(^1\) in AI—a research that has as one of its most basic assumptions that the problem solver should have at its disposal an internal symbolic model or representation of the problem world. By manipulations on this internal model, corresponding to real-world actions, the problem solver can do such things as predict what would happen if a certain action were performed, decide which conditions must be fulfilled to be able to produce a particular situation, compare different hypothetical states of affairs, etc.

This idea of internal representation is brought out clearly in a distinction between \textit{epistemology}\(^2\) and \textit{heuristics}, introduced by John McCarthy and Patrick Hayes [1969], that has been generally adopted in AI under various aliases. The epistemological component of the problem solver models the world, and the heuristic component does the problem solving (the

\(^1\) Here, \textit{problem-solving} is used as a general term, including \textit{planning}—solving problems of achieving a certain state of affairs by actions—as a special case.

\(^2\) Readers with a background in philosophy had better not think of the usual connotations of the term.
talk about components should not be taken too literally).

Unfortunately, these concepts are far from clear and well understood. Basically, epistemology is the penultimate arbiter of plans proposed by the heuristics; it is what it takes to determine if a proposed plan stands a good chance of succeeding. Contrasting with this selective role of the epistemology, there is the heuristics as a generator of plans. As the name implies, we certainly expect it to be a smart generator - something considerably better than exhaustive search. In practice, processes of generation and selection are intricately interwoven, which makes the distinction difficult to apply. Our understanding of the distinction between epistemology and heuristics also involves a range of more or less loosely associated ideas: justification vs discovery, certainty vs uncertainty, deduction vs induction, knowledge vs reasoning, neutral vs goal-directed, reason vs will, etc. Disentangling all this would require a separate study. I will simply proceed on the assumption that we have a general idea of what epistemology and heuristics are, which will do for the present purpose.

It is tempting to speculate on a general trade-off between modeling and heuristics. Some problem-solvers may rely on having a very rich and accurate model of the world, and then they don't have to be very clever, because all they have to do is to consult the model for the answers. But some problem-solvers may rely on their cleverness, what they don't know they can always figure out, and so manage without much explicit knowledge of the world.

3) The ultimate arbiter is, of course, reality.
The research on mechanical problem solving can be roughly classified as belonging to two different research traditions: a deductivist, and a non-deductivist. Although there has been an exchange of ideas and techniques between them, each tradition has preserved its own view on the problem of problem-solving and how it should be attacked; for example, deductivists have a particular concern for modeling (epistemology), whereas non-deductivists have a particular concern for heuristics.

The non-deductive approach

The non-deductivist tradition is the older of the two. It has its roots in a psychologically oriented approach to artificial intelligence, that assumes that the most practicable road to artificial intelligence systems is via the study and imitation of natural intelligence. On this view the intelligent machine simulates the brain at a high level, in terms of thoughts, reasoning, etc.

Human reasoning does not seem to depend very much on deductive inferences, at least not at the level of reasoning which Herbert Simon and Allen Newell have explored and documented in their Human Problem Solving [1972]. The General Problem Solver (GPS) of Newell, Shaw and Simon [1960] uses not logic and deduction, but a means-ends analysis for its reasoning. Here is a much simplified description of how it works. The goal is approached in a series of steps. Each step starts with calculating the difference between what has been achieved so far and the goal. On the basis of the characteristics of this difference, operations that are most likely to have an effect in reducing the difference are tried in

PART 1: DEFINING THE PROBLEM
succession until an operation succeeds in bringing the system closer to its goal. The reasoning that takes place in the process of solving a problem is mainly heuristic: it is by informed guesses, rules of thumb, trial and error, etc, that a solution is reached.

Some systems, although non-deductivist in spirit, depend considerably on deductive methods. A notable example is STRIPS (Fikes and Nilsson [1971]). In STRIPS, reasoning within a given world-state or situation (see below) proceeds deductively, and is sharply distinguished from reasoning between different world-states, which proceeds in a GPS-like manner. These two modes of reasoning — within a given world-state, and between different world-states — I will call synchronic and diachronic respectively. Diachronic reasoning is of course crucial to the system's planning ability, which is why I don't hesitate to place STRIPS within the non-deductivist tradition. In fact, STRIPS was introduced as a system that liberated diachronic reasoning from deductivist constraints (at the time, the deductivist tradition was dominating), and thereby, it was contended, solved the frame problem.

4) The feed-back mechanisms studied by early cybernetics were quantitative and continuous. GPS can be seen as a qualitative and discrete feed-back system.

---

- The non-deductive approach
The deductive approach

Approaching the problem of problem-solving, free from psychological considerations and inspirations - either because of a technological, performance-oriented outlook, or because one has in view to study intelligence per se (one variety of which happens to be instantiated in humans) - logic stands out as a promising means for implementing symbolic models of problem-worlds. Logic is a formal language with well understood syntax and semantics, which should make it suitable for machine application. Furthermore, it comes with a deductive system, a mechanism of inference. The deductive structure, however, could not be put to practical use until the resolution method was formulated (Robinson [1965]), a method that made it feasible to construct logical proofs mechanically. It soon turned out to be possible to use formal logic and mechanical theorem provers to implement question answering systems (synchronic reasoning), and even more general problem solving systems. The ensuing deductively oriented research in AI can boast of a solid and prestigious basis in the long and successful history of deductive ideals in western philosophy and science.

The efficiency of pure theorem provers in solving general problems is often unacceptably low - considerable amounts of more or less domain specific heuristics is needed to make deductively based system work reasonably well. Nevertheless, the deductivist tradition attaches greater weight to the problem-solver's ability to model the world, its epistemological power. Difficulties are met with a general strategy of extending the ontology, to include for example history, intentions, and knowledge, as entities in the model explicit to the system itself. Extensions of logic itself are rare - there seems to
be a general conviction within the deductivist tradi-
tion that a fairly simple logic will do, preferably 
first order predicate logic. (An exception is the 
recent, not wholly successful attempt, to develop a 
non-monotonic logic, see Part 2 below).

A central concept of the deductivist paradigm, which 
is rather popular also with non-deductivists, is the 
situation, "the complete state of affairs at some 
instant of time" (McCarthy [1968:411]). The world is 
conceived as a sequence of situations, determined by 
a set of laws of motion. A description of a situ­
tion consists of a set of sentences of first order 
logic. The deductive system can then be used to 
infer other sentences which also apply to the same 
situation. Explicit knowledge is in this way comple­
mented by implicit knowledge, by means of synchronic 
reasoning. Moreover, physical determinism is 
assumed: "The laws of motion of a system determine 
all future situations from a given situation" [ibid]. 
That is, given the laws of motion, a description of a 
situation, and an action, a description of the new 
situation that would result if the action was exe­
cuted can be deductively inferred. Thus, also 
diachronic reasoning is deductive.

A summary of some important ideas in the original 
deductivist setting follows:

1) The situation concept of McCarthy and the related 
concepts of action and law of motion.

2) The use of first-order logic to represent situa­
tions.

---

5) A caveat: I will frequently write situation 
when I mean a description of the situation, action 
when I mean the action's representation in the 
problem solving system, etc.

- The deductive approach
3) The use of first-order logic to represent laws of motion. Deductive relations between descriptions of situations: physical determinism.

4) Actions are the only entities that have the power to create a new situation. Nothing happens that is not caused by an action.

5) The problem solver is the only agent, all actions are actions of the problem solver.

6) When an action is performed the problem world reaches a new equilibrium fast. A new action is never initiated before the commotion caused by the preceding action has petered out.

7) There is no history in the situations - there is no explicit memory of past actions, no objects or properties designed to represent earlier events.

8) The problem world is in principle completely described. The effects of an action on every detail of the problem world is completely specified.

Several of these points are apparently consequences of a decision to avoid interfering processes - there is to be no explicit treatment of interactions. There are no other agents whose actions may interfere with the problem solver's own actions, and an action is not allowed to interfere with previous actions. Causal chains are not explicitly represented. Since there are no simultaneous events, nor memory of past events, it appears more expedient to deduce all effects of an action at once, skipping intermediate results.

In short: the early deductivists were committed to a simplified 17th century mechanics way of representing the world.
The original frame problem

The deductive research program soon ran into difficulties. In the process of formulating laws of motion it turned out that the relevant laws were drowned in a flood of seemingly irrelevant but necessary pseudolaws. Deducing a new situation requires not only that changes effected by an action are inferable, but also that what does not change can be inferred. And, in fact, very little of what from a purely logical point of view could happen as a result of an action, does happen. Whereas it is logically possible that turning on the light in my living-room will make the walls turn black, halt the inflation, and raise the melting point of lead, nothing of this does in fact happen. Every empirical truth relevant to the modeling task has to be imposed on the logic in the form of proper axioms.

The outcome is a problem-solver sitting in the midst of a sea of axioms representing non-changes, endlessly calculating non-effects. This is the frame problem in its original manifestation; let us look at some details.

A sentence which is true of one situation may be false of another. Somehow each sentence in the model must be tagged with the situation it applies to. One way of doing that is McCarthy's fluent. For example, rains(x) is a fluent such that rains(x)[s] is true if and only if it rains at x in the situation s.

Let us illustrate with an example from the blocks-world - a toy-world containing some blocks on a table and a robot-hand that can do things such as move to different positions, grasp a block, etc. The GRASP-operation could be defined by the following law of motion (using fluents):
\( \forall s \forall x \forall y ((\text{HAND-OVER}(x) \land \text{HAND-FREE} \land \text{BLOCK}(x) \land \text{ON}(x,y))[s] \Rightarrow (\text{HOLDS}(x) \land \text{HAND-OVER}(y))[\text{GRASP}[s]]) \)

That is: If the hand is over \( x \) (in the situation \( s \)) and it is empty (in \( s \)) and \( x \) is a block (in \( s \)) and \( x \) is on \( y \) (in \( s \)); then the operation \( \text{GRASP} \) will result in a situation \( \text{GRASP}[s] \) where the hand is holding \( x \) and is above \( y \). This is all that can be said of the new situation, given this law. Note that \( \text{ON}(x,y) \) is no longer known to be true (neither is it known to be false), and that the hand is no longer known to be free (and neither is it known not to be free). If we want \( x \) to remain a block, that fact must be possible to infer, for example by including \( \text{BLOCK}(x) \) in the consequent of the implication. If we do not want the other blocks to lose their existence or properties, all these facts must also be explicitly included in the laws of motion. Everything that does not change has to be included.

Normally these non-changes are represented by special axioms, so called frame axioms. A frame-axiom for \( \text{GRASP} \) might be:

\( \forall s \forall x (\text{BLOCK}(x)[s] \Rightarrow \text{BLOCK}(x)[\text{GRASP}[s]]) \)

These axioms represent the frame of reference for the action, the background which remains unaffected. (Hence the name frame problem.) Clearly, this way of modeling is not practicable in a slightly more complex problem world; the frame axioms will occupy all space and the deductions of non-changes all time.
The general frame problem

At this point one may begin to suspect that the frame problem is really just a consequence of choosing logic as the representation language. Will the problem not simply disappear if we abandon the strict use of logic? For example, we could try to introduce some kind of convention that things not explicitly changed are taken to be unchanged.

Superficially, this would seem to solve the problem. Unfortunately, only superficially - new problems crop up that are, like the original frame problem, intimately connected with the phenomenon of change. As we shall see in part 2, the attacks on the frame problem has so far resulted in little more than pushing it around, and rephrasing it in various ways. It is now widely accepted that the original frame problem is just one manifestation of a general, and - some would say, profound, problem of dealing with change. It is the frame problem in this general sense I address in this paper. Whether there is a general solution is a much more controversial matter.

The frame problem is not just a minor technical nuisance, but a very real and serious obstacle to the design of any kind of system that is to model a complex and changing world. Anticipating the analysis and discussion that follows, here is my definition of the problem. The general frame problem is the problem of finding a representational form permitting a changing, complex world to be efficiently and adequately represented. There are three points in particular, I wish to draw attention to:

i) The frame problem is a problem of modeling - not heuristics. For one thing, because the problem solver will not always know in advance what it will do next that requires a good understanding
of the present situation also in aspects not directly addressed in its previous problem-solving activities. Second, because the frame problem is real enough, even if we don't set out to make a problem-solving system.

ii) The frame problem is not a question of content, but of form. Logic is perfectly capable of representing the blocks-world, as far as content is concerned. The frame problem is an indication that informationally equivalent systems, systems conveying the same information about the represented world, may yet differ drastically in efficiency. The situation is somewhat analogous with choosing between different programming languages; although they are computationally equivalent in the mathematical sense, we recognize that there are important differences in the ease and efficiency with which they are applied to different types of tasks.

iii) The choice of suitable form depends on the problem-world - computational considerations will not suffice. The frame problem is not strictly a problem of inventing the right data-structures and algorithms from an internal Computer Science perspective, its solution requires that we seriously take into account how the problem-world fundamentally is.

We should expect that there was a frame problem also in the early non-deductivist systems, such as GPS. But the circumstances were not such that the frame problem could crystallize into a problem separable from general book-keeping problems; modeling was not very developed nor clearly distinguished from heuristics, and there was a mathematically inspired tendency to view actions as complete transformations of the world rather than piecemeal manipulations. In
the deductively based systems the frame problem is so much more conspicuous, which makes the deductivist research tradition's ideas and views the natural point of departure for discussing the frame problem.

The frame problem is one of four closely related, and sometimes confused, problems pertaining to the modeling of change. The other three are: the prediction problem (Hayes [1971]), the general book-keeping problem, and the so-called qualification problem (McCarthy [1980]). By showing what the frame problem is not I hope to give it a more distinct demarcation. I will argue that the frame problem, the prediction problem, and the book-keeping problem, concerns respectively, the metaphysical, the instrumental, and the implementational adequacy of the representation. The qualification problem, can be construed as the problem of applying idealized knowledge to the real world, involving the frame problem as well as the prediction problem.

The prediction problem

Sometimes a prediction does not come true; a certain sequence of operations has been calculated to solve a certain problem, but when the plan is executed it somehow fails to accomplish its goal. This may happen even if we disregard the possibility of mistakes due to carelessness, forgetfulness, and the like — errors that a mechanical problem solver is not likely to be prone to. Given the laws of motion and a description of a real-world situation it is still possible that a correct application of the laws to the given description will produce a prediction that does not completely agree with the subsequent real-world situation. In such a case, either the laws or
the description of the given situation must be imper­fect.

In practice, this is an inevitable difficulty. We have to accept that predictions can go wrong, and in fact in non-trivial systems it will happen frequently enough to warrant the introduction of some mechanism of revision whereby the system may automatically correct itself. The revision problem is concerned with the construction of such mechanisms. I will come back to it at the end of this section.

The nature of the prediction problem can be uncovered by discussing some important factors that limit the system's ability to make correct predictions. To start with, the principle of determinism may itself be put in question; as we move to more complex problem-worlds, for example including conversation with humans, one may begin to doubt that there really exists a totally determining set of laws. For the sake of argument however, determinism is assumed in the following.

Let us make a distinction between the system's point of view and the outside observer's point of view. As an extreme case, consider a system that from the out­side point of view totally lacks perception. The system may yet consistently believe that it makes authentic observations of the world and find itself always in complete agreement with its reality, given laws and descriptions that produce descriptions con­stituting what the outside observer considers to be its self-generated perceptions. So this is one way, admittedly weird, of "solving" the prediction prob­lem: by isolating the system from the world, letting it withdraw to a hallucinatory world controlled only by itself. Obviously, we have only theoretical use of such systems.

16

PART 1: DEFINING THE PROBLEM
From the outside observer's point of view there are some limitations in the modeling that have a negative effect on the reliability of predictions. McCarthy [1968:411] says that a representation of a situation is necessarily only partial. The scope of a representation is for practical reasons limited. But perfectly isolated parts of the world are rare, if they exist at all. Anyway, a part of the world that appears isolated to the system, and that includes everything more directly relevant for the problem at hand, will most likely be far too large to represent in toto. Thus, the limit in scope limits the reliability of predictions.

Given a particular language or form of representation, a description expressed in this language remains ambiguous even if it says all that can be truthfully said using this language. At least, there is a wide-spread belief that the world always exceeds any description of it in richness of details and subtlety of nuances. And, of course, if the system has only one concept of red for example, it will not be able to discriminate between different nuances of red, so its description of some red object will be satisfied by several possible situations. From the outside point of view, the system divides the world into equivalence-classes, each class appearing to the system as a unique situation. As far as the prediction problem is concerned, this may seem unimportant; a coarse description can be quite as truthful as a more detailed description, and we have some freedom to choose the level of detail to fit our purpose. However, a coarse-grained representation may have unpleasant side-effects: there will possibly be, from the system's point of view, unobservable, unrepresentable processes at work underground that now and then will surprise the system by producing perceivable effects that the system cannot anticipate. To guarantee accurate predictions at a coarse
level of description, we may be forced to use forms of representations that are too fine-grained and too resource demanding, to be practically feasible.

Incidently, this suggests another "solution" to the prediction problem: making the system so dull and undiscriminating that its predictions run very little risk of being wrong. Such a system is like the perception-free system in that it becomes isolated from the world - as the number of world-states it can differentiate between diminishes it loses contact with reality.

The coarseness of the representation language may be countered by introducing an idealized problem world (there might of course also be other reasons for doing this) which exactly matches the resolving power of the language. In an idealized world we could for example limit the repertoire of red colors to just one. Evidently, as soon as we apply calculated solutions of problems in the idealized world to the real world, we run into a new source of inaccuracy: the discrepancy between the idealized world and the real world. The early research in mechanical problem solving was inclined to use extremely idealized problem worlds, like the blocks-world. In these neat and tidy environments the efforts could be concentrated on pure problem solving. Some of these idealized problem worlds make possible systems that have completely accurate representations (of the idealized world, that is), systems that in fact do not suffer from the prediction problem. This "solution" of the prediction problem is yet another version of the perception-free, hallucinating system.

For various reasons, descriptions are commonly less than total even when considered from the system's point of view. One obvious reason is incomplete perception; hidden views, etc. Another common reason is
that goals are seldom totally specified, partly because the system wants to solve a class of problems, partly for reasons of economy and focus of interest; there may be a specific interest in having block A on top of block B, without caring the least what happens to block D. Notice also that a more fully specified goal runs a greater risk of being impossible to attain. There may also be actions with less than total specificity, (see Part 2: Unless, below) reducing the degree of specification. With these observations we have approached the last important limiting factor that will be discussed here:

The best map is not the terrain itself. The more minutely we represent the world, the more complex the representation, till at last it is too complex to be useful. As an extreme, consider the case where the more exact method of representation is so slow as to produce the correct prediction at a time when the predicted situation already has occurred. Painstaking precautions to avoid mistakes are not always worthwhile. This is a point where trying to cope with the frame problem appears to be at odds with solving the prediction problem. Most probably, there are such conflicts, but the problem of mastering complexity is a more general concern than the frame problem. Even in systems modelling static worlds we may have to simplify in order to make problem-solving and question-answering reasonably efficient.

In all these cases, imperfect prediction calls for readjustment and revision of the model after failures. The task of finding good strategies for revision, might be called the revision problem. I think it is important to keep a distinction between revision and debugging. Debugging is something that takes place between the heuristics and the epistemology, when the epistemology isn't quite satisfied with
what the heuristics has come up with. Debugging implies an investigation into how the system has come to have a faulty belief.

Revision, on the other hand, is the concern of some basic support system for the model. Primarily, it is just a way to get rid of conflicting facts as fast and painless as possible; readjusting to new, unexpected, circumstances, with a minimum effort and a minimum loss of information. The purpose of revision is thus not to make appropriate "corrections" so that future predictions will be more accurate. The discussion above points out the limits of such refinements; besides, adjusting to the unexpected, and learning from it, are not in general the same.

It should be clear at this point that the prediction problem and the frame problem are distinct; recall that it was in the very same early problem solving systems that had actually eliminated the prediction problem by using idealized problem-worlds, that the frame problem was first discovered. To sum up: the prediction problem is the problem of making the representation instrumentally adequate, it is mainly the problem of representing the physics of the modeled world. It does explicitly not concern the form or the internal workings of the representation. That, precisely, is what the frame problem is about: to choose a form of representation that makes for efficient administration, a form in which the physics is efficiently expressed. In particular I will suggest that the suitable form is determined by the

6) Of course, the term "debug" reflects only one type of idea about how to exploit discarded plans; a good plan is a bad plan minus a number of errors. Incidentally, the term debugging as used for developing general plans or acquiring general laws, suggests an inductive or cumulative view: the laws are essentially correct, they just need yet some small adjustments to work perfect.
metaphysics of the modeled world, so we might say that the frame problem is mainly the problem of representing the metaphysics of the modeled world.

The general book-keeping problem

The book-keeping problem is the problem to keep track of the various plans and hypothetical situations that the system, partly in parallel, makes use of during the planning process. One of the major problems is to simultaneously keep in storage several different descriptions of situations. There is also the problem of relating the descriptions to one another, for example that situation B is the calculated result of performing X in situation A. Various context mechanisms have been used to avoid duplications of information and to structure the set of situations.

On the face of it, the frame problem looks like a special case of a general, and rather fuzzy, book-keeping problem. To be sure, they are both concerned with efficient model administration, but actually they have radically different characters. Most of all, it is a difference of approach. The book-keeping problem implies an essentially technical view on the problem of representing change; an instrumentalist approach that opens up for patching and hacking. By now, it is rather evident that there is more needed than a bag of tricks to speed up and save space, a tactic somewhat reminiscent of a novice programmer trying hard to optimize the code, instead of trying to find a better algorithm. A radical solution to book-keeping problems requires a penetration in depth to the underlying causes of poor book-keeping.
The frame problem, viewed as first and foremost a problem of representation - not administration, in particular the problem of representing the metaphysics, is one such in-depth approach. It invites to realism and theorizing, which I believe will be more fruitful in the development of more efficient methods for book-keeping than direct attacks on the book-keeping problems.

From this point of view, the book-keeping problem is less a research problem than a symptom of deeper problems, one of which is the frame problem. But there is a genuine problem half hidden in the all-embracing worries about efficiency, which is: assuming that we have found out what the representation should look like - how do we implement it? We have no reason to expect that there is a predestined perfect harmony between existing software and hardware, and the representational forms demanded (by the metaphysics, say).

I don't know if "the book-keeping problem" is the most accurate heading for this, but I think that the scientifically and technically most interesting side of book-keeping is the question of implementational adequacy, the question of how the adequately designed representational system is to be implemented using current hardware and software technology. Or, if that can't be done, how we are to develop the required programming languages, computer architectures, hardware components, etc. Even if the instrumental sufficiency, the representational power, of our most cherished forms or languages of representation that are machine implemented, is not called in question, it just may be too hasty to assume that the demands on form and organization, raised by (for example) the frame problem, can be met with current hardware and software technology.
The qualification problem

A complete representation of all conditions that must be satisfied to guarantee a successful performance of an action puts an unreasonable, maybe unlimited, amount of qualifications on the laws defining the action. Here is an example from McCarthy [1980]. The successful use of a rowboat to cross a river, requires that the oars and rowlocks be present and unbroken, and that they fit each other. Many other qualifications can be added: the boat must be sufficiently watertight, it must have sufficient carrying capacity, there should be no logs floating in the river that could upset the boat, and so on - we have no trouble in extending the list indefinitely as it seems. The rules for using the boat, rapidly becomes hopelessly complex to use.

This may seem to be the prediction problem seen from a different angle: how can we possibly account for all the premises that are needed to make safe predictions? We generally cannot. There is no fail-proof closed formula for crossing a real river by a real row-boat. Basically, this is why we need a model to "run" proposed plans on. As far as the qualification problem is a problem it seems to be a problem of heuristics.

Yet, the qualification problem appears to bear on the frame problem as well. In fact, the qualification problem seems to involve epistemologically relevant decisions as to what properties of an object or an action are taken to be essential, which is a metaphysical question if any. Is it part of the essence of a row-boat that rowlocks be present? That the rowlocks be unbroken? Is it part of the essence of the MOVE-operation that what is to be moved is moveable? Etc.
I think that the qualification problem is largely just another way of expressing the problem of applying an idealized representation to the real world. In a common-sense type of idealized world there might for example be a law that says, without qualifications, that one can use a rowboat to cross a river. The epistemologically relevant part of the qualification problem is to apply that knowledge in a real situation: what qualifies as a rowboat (oars, watertight, etc)? what qualifies as a river? etc. Since one would expect an idealized world to differ from the real world both in its physics and its metaphysics, it would seem to explain why both the prediction problem and the frame problem are reflected in the qualification problem.
 Frames

An event will normally affect the world only in a very limited aspect. Change of location does not affect colors. Painting does not change the location of things. Applying insecticides does not (at least, should not) affect the health of other animals than insects. If it were possible to separate the world into different aspects, so that the effects of a certain type of event are limited to a certain aspect associated with that particular type of event, it would follow that the world remains unchanged in every other aspect. This is the idea of frames, briefly discussed in McCarthy & Hayes [1969] and Hayes [1971, 1973].

A frame separates the description of the world into blocks, each corresponding to a different aspect of the world, so that an event belonging to a certain block affects only facts within the same block.\(^1\) A new inference-rule is introduced: If \( F \) is a formula true in situation \( s \), the event \( e \) has the property of belonging to block \( j \), and all non-logical symbols crucially occurring in \( F \) belong to other blocks than \( j \), then \( F \) remains true of the situation that results

\(^1\) The intention of Hayes [1971] and McCarthy & Hayes [1969] seems to be that the blocks are to be disjunct, but Hayes [1973] does not postulate this, and it is doubtful whether it is possible in a common-sense world to have disjunct blocks.
if the event e takes place in situation s. There are some choices as to what exactly is to be meant by **crucially occur**, the important thing being that it remains a purely syntactically defined relation.

To carry out these ideas it will be necessary to cut the description of a situation into small pieces, where each piece belongs to a particular block. The reason is that we do not want an aggregation of information (e.g. a sentence) to be put in doubt in its entirety, just because one of its parts belongs to the block that includes the current event. Block membership has to be decidable without knowledge of the context, i.e. the membership of a particular fact of a particular block must be independent of the other facts. This leads to a separation into blocks that is based on which predicates, functions, constants, etc, that occurs in the description.

The weak point in the frame idea is that the separation into blocks will have to be so coarse that it will be of little help. Even if we know that painting only affects the color block, we are still left with the considerable task of deciding exactly which of our facts involving color remain valid. If we stick to natural actions like painting, moving, and so on, the number of blocks has to be small, consequently the average block has to be large. This follows from the inherent generality of natural actions. Any finer classification into many, smaller, blocks would only be disrupted by actions cutting across block boundaries. On the other hand, if we try to make a finer classification by reducing the natural actions into small atomic actions that affect the world in fewer aspects, we are confronted with three problems: (1) natural actions have no fixed decomposition into atomic actions, there are, for example, innumerable ways of moving something; (2) instead of affecting one big block, the original action will now
affect many small blocks, so the amount of work in deciding which facts have been changed is not necessarily reduced; (3) it is doubtful whether the blocks in such a classification would have even the remotest affinity with natural aspects and natural common-sense properties.  

As an example of two blocks in a coarse but safe classification, Hayes mentions location and color. But consider a chameleon: moving it will frequently incur a change in color. What affects what is apparently highly dependent on the circumstances, the particular environment of the events. As pointed out in McCarthy & Hayes [1969], it is not possible to make a classification into blocks that will apply always, in any environment. It is not even possible for a relatively limited part of the real world. This observation naturally leads to the idea of having many frames, each tailored to a particular kind of environment or activity. Thus specialized the frame method would perform much better, allowing more reliable, as well as finer, classifications. We have now come quite close to the frame concept of Minsky [1974], and the idea of prototype situations and environments. The price we have to pay is a new set of difficult problems. Among them is the problem of how and when to switch frames, and in this process, how to reshuffle information from the old frame into the blocks of the new frame. These problems remain largely unsolved. Add to this the disquieting question of how severe restrictions the general use of prototypes puts on the ability to model the world.  

2) But why is it important to have natural actions, natural objects, natural properties, and so on? Are we really restricted to common-sense metaphysics? I shall attempt an answer in part 3.
There is a certain similarity between the method of frames, considered as an exploitation of the idea that the effects of an event are propagated property-wise — for example, painting affects a definite property of things, viz. their color — and the method of causal connection (Hayes [1971]), which can be viewed as an attempt to formalize the idea that effects are propagated object-wise — painting affects things in the vicinity, movement within a room affects things in that room, etc.

A formal relation called causal connection is introduced that relates objects to each other, and also actions to objects, in a situation. Informally, that A is causally connected to B, means that a change of some property of object B may cause a change of some property of object A. This relation is, according to Hayes, reflexive and transitive. If, in a given situation, it can be proved that an action and an object are not causally related, it is a valid inference that that object is not changed by the action.

The method of frames and the method of causal connection have a common strategy for the frame problem: giving general rules for classifying the different parts of a description into a class of facts that remain unchanged, and a class of facts that are candidates for being retracted or revised. In both methods, the strategy, even when successful, leaves us with a local frame problem: to decide what is, and what is not, changed within the class of doubtful facts. Notice that the causal connection relation depends on the situation. Contrary to the methods of frames where class membership is static, the method of causal connection allows dynamically varying membership, actions may establish new causal
connections and remove old connections. This also allows for a flexible way of adding to the problem world, introducing additional mediate effects of the actions, without having to redefine existing actions; the immediate effects could stay the same.\(^3\)

We recall from the previous section that in the method of frames the local frame problem is in fact not very local. The method of causal connection has a local frame problem that is probably even worse: the objects which are causally connected to the current action may change in every respect, not only in the property in virtue of which the causal connection holds. If, for example, a change of position of object A may cause a change of temperature in object B, the problem is to decide whether the temperature of B actually changes, plus the problem to decide whether its color changes, its weight, electric potential, etc, changes.

The usefulness of Hayes's proposal depends on whether it is in general possible to limit the causal connections of an action, indirect connections included, to a small number of objects. Unfortunately, there are causal connections (in Hayes's sense) that are almost ubiquitous. One example: whenever two objects are in contact, or very close to each other, it is in general true that a change of position of one of the objects may cause a change of position in the other. The transitivity of the causal connections makes it plausible that most of the objects in an ordinary living-room, say, are causally connected to each other. Furthermore, the causal connection relation does not discriminate between different types of

\(^3\) It seems reasonable that this is a condition that any solution to the frame problem must satisfy: that (conservative) additions to the problem world does not require a complete recompilation of the model.

- Causal Connection 29
causes so that missing links in the causal chains corresponding to, for example, change of temperature as a cause, will by transitivity probably be bridged by some other cause or causes.

For a particular problem it might in some cases be possible to keep down the number of causal connections by using a specially tailored representation. For a particular problem it might in some cases be possible to keep down the number of causal connections by using a specially tailored representation. Of course, such ad hoc methods have very little to do with a general solution to the frame problem.

Consistency

The world is essentially inert. Things are generally inert, their properties are changed only exceptionally. Therefore: assume that each object retains its properties as long as this assumption does not lead to a contradiction. This view underlies the consistency based methods for the frame problem. Surprisingly, the principle of inertness is not the motive stated in the introduction of the method (Hayes [1973]). Instead, the proposal is motivated by fears that the comprehensive axiom systems demanded by non-trivial problems will, in practice, contain inconsistencies and so be useless.

4) The fact that Hayes's illustrating example, the monkey and the bananas, works at all, is completely dependent on that kind of move. For example, the function location is introduced as a loose horizontal coordinate measure together with an axiom to the effect that two objects are causally connected only if they have the same location. The axiom is misleadingly interpreted by Hayes as "there is no causality at a distance", which sounds like a well-known and respectable mechanistic principle - there can be no immediate action at a distance - but in fact the axiom forbids any causal chain extending beyond the location of the action. The result is a universe of completely isolated cells, each corresponding to a different location.
The method in short: Given a situation and an event we may, using the laws of motion, immediately infer certain changes giving us a partial description of the new situation. This partial description is completed by transferring the old description piece by piece under constant surveyance of the consistency of the new description being created, stopping when inconsistency arises. The task is to determine a maximal consistent description that extends the explicitly inferred partial description into the description of the old situation.

There are, as remarked by Hayes, two serious theoretical problems. First, consistency is not decidable for first order theories. Second, there will generally be many maximal consistent extensions - which one should we choose, and why? The first problem is very serious to a deductivist since the deductivist ideal demands that inferences always are logically valid. Besides, it would seem to dispose of Hayes's original motivation for his proposal. Abandoning these ideals, ways could possibly be found to solve the problem.  

The second problem seems more tractable. In connection with his work on counterfactual conditionals, Nicholas Rescher [1964] has developed a method of classifying sentences in modal categories as a basis for choosing maximal consistent subsets. A modal categorization can roughly be described as a sequence of sets of sentences (the sets obeying certain conditions of closedness) where every set includes the preceding set, arranged according to how fundamental the sentences in each set are. A maximal consistent subset is built by adding sentences starting with the

5) But Hayes remains a strict deductivist and he can see no other way out than to drastically constrain the expressive power of the language to make it decidable.
most fundamental category and proceeding down the sequence. Depending on the criteria of fundamentality, different categorizations with different properties, are obtained. Hayes suggests as a suitable criterion the theory of causal ordering proposed by Herbert Simon (see Simon & Rescher [1966]).

Given that one could somehow be reconciled with the problem of decidability, we are still very far from a solution to the frame problem. The ideas of subset construction outlined above must be developed into a usable method using specific categorizations. But the problem of elaborating a suitable theory of categories of fundamentality—indeed a metaphysical problem—is on the whole just a reformulation of the frame problem in a different vocabulary, and Simon's theory of causal ordering gives nothing but hints. Even if we had that theory it is still difficult to believe that the consistency method would be reasonably efficient. It is too explicitly occupied with managing non-changes.

The basic idea of inertness, however, is attractive. It has survived and reappears in several later proposals. One difficulty lies in formulating this principle in a suitable procedural form; the metaphysical principle of inertness is too vague to guide the design of an efficient procedure.

STRIPS

In STRIPS (Fikes & Nilsson [1971]) there is for the first time a clear separation of synchronic and diachronic reasoning. The synchronic reasoning works deductively, the diachronic reasoning uses GPS-like methods. Precisely through not using deductive methods for diachronic reasoning, the frame problem
is claimed to have been overcome. The main principle is that all that is not explicitly changed by an action, remains unchanged. Although based on the same principle of inertness as the consistency method, the superior computing power of the STRIPS principle is apparent: there is no need to consider in any way that which is not explicitly changed.

The following description is limited to the modeling mechanism of STRIPS. An action is represented by: (1) a set of preconditions; (2) a delete list; (3) an add list. An action may be applied in a certain situation if the preconditions are satisfied. The result is that everything in the description that matches the delete list is removed, and the items in the add list are added to form a description of the new situation.

An example: the definition of an operator push(k,m,n) (with semantics: object k is moved from m to n by the robot):

Preconditions: \( \text{ATR}(m) \land \text{AT}(k,m) \)
Delete: \( \text{ATR}(m), \text{AT}(k,m) \)
Add: \( \text{ATR}(n), \text{AT}(k,n) \)

where k, m, n, are parameters to be instantiated before applying the operator. (Semantics: \( \text{ATR}(x) \), the robot is at \( x \); \( \text{AT}(x,y) \), object \( x \) is at \( y \)). The definition may be supplemented by synchronic laws, like:

\[ \forall u \forall x \forall y \ (\text{AT}(u,x) \land (x \neq y) \supset -\text{AT}(u,y)) \]

(Semantics: If an object is at one place, it is not at another).

All diachronic knowledge is contained in operator definitions. As we have noted, the use of logic and deductive methods may at first give the impression that STRIPS belongs to the deductivist tradition.

- STRIPS 33
There are also additional features that are typically deductivistic, such as a reductionist tendency. But diachronic reasoning is actually GPS-like.

The add and delete lists cannot in general include everything to be changed, since very detailed lists cannot be very general and it should be possible for the effects of an action to vary with the situation. The step taken to compensate for this limitation is to define a set of primitive predicates to which all other predicates can be reduced. Only primitive predicates are allowed in the delete and add lists. To keep non-primitive information consistent with the primitive information expressed in terms of primitive predicates, various methods have been suggested. A very simple solution is to let all non-primitive information be implicit; it will then have to be inferred from the primitive information every time it is requested. Another method is to link each explicit piece of non-primitive information to the particular pieces of primitive information that was used to infer it; whenever a piece of primitive information is changed, the links leading from it are used to find and erase non-primitive information based on it. A more sophisticated technique is dependency-directed back-tracking (see TMS, below).

There are actually three distinctions involved here that should be kept apart. One is the distinction between explicit and implicit information. Another is the distinction between primitive and non-primitive information. The third is the metaphysical distinction between primary and secondary properties of the world (Galileo, Locke). Any particular separation of the information into primitive and non-primitive, is a property of the system of representation. Even in a strictly logical representation there is a wide range of choice between different axiom-systems. The primary - secondary
dichotomy on the other hand, has to do with how the world is (or how it is viewed). From a strictly internal point of view there is freedom of choice in deciding what information is to be primitive. But to be useful, the choice of primitives should be grounded on a primary-secondary categorization that is fundamental for the problem world and the system's relation to it. Generally, the metaphysics of the problem world should agree with the internal organization of the representation: the primitive-non-primitive dichotomy corresponding to a primary-secondary dichotomy. (These ideas are developed further in part 3.)

Actually, in STRIPS, the use of primitive predicates is introduced as simply a technical device facilitating book-keeping, and consequently the choice of primitives is not discussed. The ontological assumptions behind the actual choice (of, for example, AT, ATR) remain hidden; the basic questions of what is required to carry out a primary-secondary separation, or whether it is at all possible, are not explicitly considered.

The following example from Hayes [1971] illustrates a serious problem:

![Diagram of cup and saucer](image)

If the saucer is moved by push(s,m,n) the cup will move with it (under certain assumptions). Yet, the cup and the saucer are two distinct objects, and if we move the cup, the saucer will not move. Push(s,m,n) has the effect of erasing AT(s,m) and adding AT(s,n), but to be a correct representation it
should have the effect of erasing AT(c,m) and adding AT(c,n) as well. This cannot be done within the STRIPS formalism since the push operator cannot presume that there is a cup on the saucer.

Actually, the cup and saucer example demonstrates a difficulty that is not peculiar to STRIPS, but is pertinent to all solution attempts that rely on the situation concept: apart from the main effects that can always be expected from an event (in this example the change of place of the saucer) there are certain effects - let us call them parallel effects - that depend on the context and cannot be described as caused by the main effects (in this case the change of place of the cup). What happens can not be described as a causal chain since there is no situation (in McCarthy's sense) which is a state of transition where the saucer has changed place but the cup has not.

There would be no trouble if the parallel effects could be handled by synchronic reasoning, in other words, if information about parallel effects could be kept as implicit information. In the example, the position of the cup would be implicit - derivable with the help of some primitive relation ON(c,s), say (which is not changed by push(c,x,y)), and AT(s,x). But that won't do, because it would mean that some instances of the AT-relation were primitive and some not, and even that what is considered primitive will depend on the situation: when, for example, the cup is put on the table its position would suddenly become primary.

STRIPS's format for representing effects as unconditional, almost completely context independent, may be appropriate for a Laplacean universe filled with bouncing atoms. For problem worlds more of a common-sense type, this kind of metaphysics is
PLANNER

PLANNER (Hewitt [1972]) is a programming language designed as an implementation language for problem solving systems. PLANNER employs a database similar to the one used in STRIPS. A program consists of procedures called theorems. Despite the term "theorem" and a number of language constructs with names borrowed from logic (OR, AND, EXISTS, etc) PLANNER does not work by logical deductions.

There are three main types of theorems: antecedent, consequent, and erasing. Associated with each theorem is a pattern which is a sentence in the database language that may also contain free variables. Inside the body of the theorem, ASSERT and ERASE can be used to add or delete from the database.

Here is an example of a consequent theorem:

(CONSEQUENT (BLACK ?X)
 (GOAL (RAVEN ?X))
 (OR (GOAL (FRENCH ?X))
  (GOAL (ENGLISH ?X))))

This theorem represents that all French and English ravens are black, or rather, to show that X is black, try to show that X is a raven, and if it succeeds, try to show that it is French, or if that fails, try to show that it is English - theorems are normally construed as inference procedures. The language

6) The notation is somewhat modified.
7) The information that the object is French, need not be explicitly in the database. With the GOAL mechanism it suffices if it can be inferred using other consequent theorems.
contains no general mechanism of inference, so the programmer has to provide the system with a theorem for each (elementary) inference that should be possible. This particular theorem about ravens, for example, cannot be used to make any other inference based on the same underlying fact that all French and English ravens are black. If we wish that X is black should be inferred from X is a French raven, we have to construct another theorem, for example:

(ANTECEDENT (RAVEN ?X))
(GOAL (FRENCH ?X))
(ASSERT (BLACK ?X)))

Contrary to the consequent theorem above, which derives implicit information without changing the database, this antecedent theorem explicitly adds the derived information to the database. Whenever something of the form (RAVEN ...) is added to the database this theorem will be alerted, and if it succeeds in showing that (FRENCH ...) it will add (BLACK ...) to the database.

Thus, a fact or law, and its intended use, are inseparable in the representation as theorems - PLANNER implies proceduralism, a movement within the non-deductivist tradition which advocates the amalgamating of heuristics and epistemology into procedures. Heuristics swallows epistemology. The contrary point of view, declarativism, has a more conservative attitude to logic; just as in natural language, the information that all ravens are black should be represented in a way neutral to several different types of application. Applying different general mechanisms of inference it is possible to use that declarative information in a variety of ways.

Nor is there in PLANNER any formal distinction between synchronic reasoning, as in the examples above, and diachronic reasoning, like for example:
(CONSEQUENT (MOVE ?OBJ ?LOC)
(EXISTS (?X ?Y)
  (GOAL (HOLDS ?OBJ))
  (GOAL (ATR ?X))
  (GOAL (AT ?OBJ ?Y))
  (ERASE (ATR ?X))
  (ERASE (AT ?OBJ ?Y))
  (ASSERT (ATR ?LOC))
  (ASSERT (AT ?OBJ ?LOC))))} corresponds to STRIPS:
- preconditions
- delete list
- add list

An interesting point about PLANNER is that antecedent theorems together with the third type of theorems, erasing theorems, provide tools useful in managing parallel effects. The cup and saucer example can be handled with the following two pairs of theorems:

(ANTECEDENT (AT ?OBJ ?LOC)
  (EXISTS (?X ?Y)
   (GOAL (AT ?X ?Y))
   (GOAL (BENEATH ?Y ?LOC))
   (ASSERT (SUPPORTS ?X ?OBJ))))
(ERASING (AT ?OBJ ?LOC)
  (EXISTS (?X)
   (GOAL (SUPPORTS ?X ?OBJ)))
  (ERASE (SUPPORTS ?X ?OBJ)))))

The first pair of theorems administrates the SUPPORTS relation. The following pair administrates the AT relation:

(ANTECEDENT (AT ?OBJ ?LOC)
  (EXISTS (?X ?Y)
   (GOAL (SUPPORTS ?OBJ ?X))
   (GOAL (BENEATH ?LOC ?Y))
   (ASSERT (AT ?X ?Y))))
(ERASING (AT ?OBJ ?LOC)
  (EXISTS (?X ?Y)
   (GOAL (SUPPORTS ?OBJ ?X))
   (ERASE (AT ?LOC ?Y))))

As soon as an object's place is changed (by erasing (AT object old-place) and asserting (AT object new-place)), the first erasing theorem, responding to the erasure, will try to find out if there is an object

8 As it happens, the antecedent theorems are poorly organized: for example, it is inefficient to search for (AT ?X ?Y) before (SUPPORTX ?OBJ ?X).
that supports the object to be moved - if so, the SUPPORT relation is erased. Similarly, the second erasing theorem finds out if there is some object supported by the object to be moved - if so, its old place is erased. The antecedent theorems will respond to the addition of (AT object new-place) - the first by finding out if the new place is on top of some object, in which case a new SUPPORT relation will be added - the second by finding out if there was an object on top of the moved object, if so, its new place is added (the last theorem invocation is thus directly involved in effecting the desired parallel effect).

Of course, PLANNER is not (and has not been claimed to be) a solution to the frame problem; it is a programming language, and as such, leaves most of the really hard problems to the programmer: making sure that erasures keep pace with additions, preventing theorems from running amok jamming the database in a combinatorial explosion of additions, making sure that theorems are effectively organized and do not lead to fatal contradictions, etc. Even the outlined method for managing parallel effects is not without problems - it remains essential to keep down the number of parallel effects or find ways of avoiding to handle them explicitly, which mainly is a question of organizing the system's view of the world properly, that is, a metaphysical question. For example, if a robot represents positions of things in a self-centered coordinate system, a simple change of location will have immense parallel effects.

Then, of course, there are the general declarativist and deductivist objections to the proceduralism in

---

9) See Moore [1980] for a discussion of the limitations of the procedural deduction used in PLANNER-like systems.
PLANNER. Declarative information does not set bounds to the ways it can be exploited, some of which may not be thought of at the time when the information is acquired. In contrast, procedural information demands a completely new representation for each new application. Also, there is no clear connection between different procedures operating with the same underlying declarative information, greatly increasing the risk of inconsistencies and obscuring an overall view. Sandewall [1972b] thinks that the integration of epistemology and heuristics will simply make the construction task for non-trivial problem worlds too big; the separation of modeling and heuristics seems to be a very natural and suitable division into subtasks.

Unless

Erik Sandewall's proposal [1972b] is formulated within a system (Sandewall [1972a]) that in several respects deviates from the original deductivistic paradigm. Predicate logic is not used in the usual straightforward way of letting properties and relations in the problem world be represented by predicates. Instead, there is a limited set of predicates that takes as arguments both objects, properties and situations. For example, IS(banana,yellow,s) represents the fact that the banana is yellow in situation s.10 Other examples: EXISTS-IN(banana,s) represents that the banana exists in s; IS2(banana,yellow) that the banana is yellow in all situations where it exists. A more definitive break with the old paradigm is to allow actions to be part

10) The notation and syntax is somewhat modified.
of the description of a situation; for example, \( \text{INN}(\text{push}(\text{banana}, x, y), s) \) representing that the banana is pushed from \( x \) to \( y \) in the situation \( s \). This makes it possible for the system to reason about actions and events. For example, we could have an axiom

\[
\text{INN}(\text{push}(o, x, y), s) \land \text{INN}(\text{support}(o, o'), s) \Rightarrow \text{INN}(\text{push}(o', x, y), s)
\]

Applied to the cup and saucer example, the axiom will have the effect that when the saucer is pushed, the cup is simultaneously pushed. Sandewall's approach to the problems of parallel effects is thus to trade the parallel effects for parallel actions, the effects of which are deduced in the same way as for any ordinary action.

Sandewall's approach to the frame problem in general, is based on the same principle of inertness that figures in Hayes's consistency-based method and in STRIPS. This principle is formulated as an inference rule which says (informally) that: if an object has a certain property in the old situation, and it can not be proved to have lost that property in the new situation, then the object still has the property in the new situation. Formally:

\[
\text{IS}(o, p, s) \quad \text{UNLESS}(\text{ENDS}(o, p, \text{Succ}(s, a)))
\]

\[
\text{IS}(o, p, \text{Succ}(s, a))
\]

where \( \text{Succ}(s, a) \) is the new situation that results if action \( a \) is performed in situation \( s \). We have here not only objects, properties, relations and actions in a situation, but also endings (and there is still more).
The operator UNLESS has the extraordinary property that UNLESS(X) is proved exactly in those cases where X can not be proved. With this operator we have abandoned the monotonicity of ordinary logic:

If A and B are axiom systems with A ⊆ B, then Th(A) ⊆ Th(B), where Th(S) = the set of sentences that can be proved from S.

That is, if an axiom system A is extended to a more comprehensive system B, all the old theorems will remain theorems in the new system. If non-monotonic inferences, like the one with UNLESS above, is introduced, this property is lost. One example: Given an axiom system with A and UNLESS(C)B as the only proper axioms, B will be a theorem since we can not prove C:

A, UNLESS(C) ⊆ B ⊆ B ;

but if the axiom C is added, B is no longer a theorem:

B, UNLESS(C) ⊆ B, C ⊏ B .

Some problems with non-monotonicity will be discussed in the next section.

The representation of an action can be more or less specific in describing the effects of the action, it can be of different degrees of (let us call it) specificity. The actions of the blocks-world have a very high degree of specificity, we might even say total specificity: given a description of a situation and an action, the action will never result in a reduced information content in the new description, it is as detailed as the old one. The total specificity of the blocks-world is a realization of an ideal

11) There is a similar device in PLANNER: THNOT.
of total representation, total knowledge of the problem world, that is part of the early deductivist paradigm.

In less trivial problem worlds it will be neither possible nor desirable to avoid actions of specificity less than total. Therefore, it seems that Sandewall's proposal is open to the following type of objection: suppose we have an action throw a stone at the window, and a window consisting of several small and unbroken panes, and suppose we have arranged that exactly one of the panes will break when the action is performed, but the action has such a low degree of specificity that there is no way of predicting (proving) which particular one it will be. Since we cannot prove that any particular pane ends being in one piece, we may, using the inference rule above, conclude that all panes remain unbroken in the new situation.

However, the predicate ENDS does not have the semantics that its name suggests. This is not clear from (Sandewall [1972b]), but Sandewall [1972a] informs us that ENDS(o,p,s) does not imply that -IS(o,p,s). The meaning of ENDS is thus closer to may end than to ends. It is now a simple matter to let the throwing result in a situation where each window-pane may have ended to be unbroken.

It is clear then, that the predicate ENDS really just serves to mark something as a possible candidate for change; we are left with a local frame problem, once again. More important - we have no guidance for the use of ENDS. What ENDS-effects should an action have? There are few formal restrictions in Sandewall's system, hardly any axioms giving rules for the ENDS predicate.12 With lavish use of ENDS-

12) There is a small attempt in that direction in Sandewall [1972a] where it is postulated that if
effects associated with each action, a robust and general system could be built, at the cost of a very large local frame problem. On the other hand, with a sufficiently clever organization we could probably make do with quite sparse use of ENDS-effects, thus keeping the local frame problem small. What is this organization like? Well, there's the frame problem again, in a slightly different formulation.

TMS

Truth maintenance systems are systems constructed mainly to aid in revising a knowledge base when conflicts arise, which make them highly interesting in connection with the revision problem. But truth maintenance systems do have some relevance to the frame problem as well, so I will describe the system of Jon Doyle [1978, 1979].

The two datastructures of the system are: nodes representing beliefs, and justifications, representing reasons for beliefs. A node may be in or out, meaning that the system has, or has not, the corresponding belief. That the system does not have a certain belief does not imply that it believes in its negation. Also, there is in principle nothing to stop the system from simultaneously believing a proposition and its negation; in other words, a node representing $p$, and a node representing $\neg p$, may both be in at the same time.

ENDS is true of a certain property $p$ of an object, then ENDS is true of every property of the object that is a subproperty of $p$.

13) The name is rather misleading. Consistency maintenance would be better, but even that is quite an exaggeration.
A justification consists (in the most common case) of a pair of lists: an inlist and an outlist. A node is in if and only if it has some justification (there might be several for the same node) the inlist of which contains only nodes that are in, and the outlist of which contains only nodes that are out. Thus, non-monotonic reasoning is a fundamental design characteristic of the system. An assumption is a node the justifications of which all have non-empty outlists.

An example will illustrate these concepts:

Justifications:

| N1) The weather will be fine | inlist  | outlist |
| N2) Be optimistic about the weather | < (...) , (...) > in |
| N3) It is going to rain | < (N4, N5) , () > out |
| N4) It may be assumed that the weather forecast is correct | < (...) , (...) > in |
| N5) The weather forecast says rain | no justification, out |

Starting in this situation, if N5 is given some justification that makes it in, N3 will become in and consequently the assumption N1 out. Of course, TMS takes no notice of what the nodes happen to stand for, I have added sentences to the nodenames just to give a hint at how the system could be used.

TMS is an attempt to overcome three related problems which all, according to Doyle, have to do with non-monotonicity. They are: (1) the revision problem; (2) the frame problem; and (3) the problem of common-sense reasoning. The contribution of TMS to the solution of these problems is belief-revision: the system can
i) automatically adjust what is to be in and what is to be out when a new justification for a node is added - the addition may cause the node to be in, which in its turn may have the effect that other nodes, the justification of which rests on that node, has to be changed from out to in or vice versa, and so on;

ii) when justification for a contradiction is added, automatically perform a belief-revision so as to make the contradiction node out (if possible). A contradiction is a special type of node, the justification of which consists of a list of nodes representing beliefs that together constitutes a contradiction.

The technique used by TMS in trying to resolve a contradiction is called **dependency-directed backtracking** (Stallman & Sussman [1977]): the system tracks backwards to the fundamental assumptions that are in at the highest possible level and which directly or indirectly give support to the contradiction, chooses one of these assumptions as a culprit and makes it out by making one of the nodes in its outlist in (justified by the contradiction). The choice of culprit, or way of making the culprit out may later prove to be wrong, a new contradiction will then show up and the process is iterated, other ways of making the culprit out, and other culprits are tried in turn proceeding to assumptions at lower levels until no contradiction shows up (or all possibilities are exhausted and the system has to give up).

Non-monotonic reasoning leads to some logical peculiarities (also discussed in Sandewall [1972b]). Consider the following example from Doyle [1979]:

(F) X+Y=4
(G) X=1
(H) Y=3

some support  \(<(J),()\)  in
\(<(K),()\)  out

- TMS
where it is assumed that \( J \) is in and \( K \) out so that \( G \) is in and \( H \) out. It is also assumed that \( F \) is in for some reason not dependent on \( G \), \( H \), \( J \) or \( K \). If we add to \( H \) the justification \( \langle (F,G),() \rangle \), a safe decision as we are doing just ordinary arithmetics, \( H \) will become in. If \( J \) for some reason becomes out (hence also \( G \)) but \( K \) becomes in instead (hence also \( H \)), we may likewise add to \( G \) the justification \( \langle (F,H),() \rangle \) on sound mathematical grounds, \( G \) will again be in. If now \( K \) becomes out, so that both \( J \) and \( K \) are out, we are left with a pair of self-supporting nodes: \( G \) is in because \( H \) is in, and \( H \) is in because \( G \) is in! And yet by now \( X \) might be 0 and \( Y \) be 4.

Doyle identifies two other types of circularity. There is one type of circularity where at least one of the involved nodes has to be out, a sort of multi-stable arrangement. And there is a paradoxical type of circularity where no combination of in- and out-states is consistent. TMS allows neither of these types of circularities.

No doubt, systems like TMS have an important role to play in the solution of the revision problem. It should be clear from the description of the belief-revision mechanism, however, that the really hard organizational problems are left to the programmer: organizing the information in levels, engineering contradictions, and so on. TMS is not in itself a mechanism for revision of the type called for by the prediction problem, rather it is an implementational tool for such revision systems.

Let us return to the frame problem itself. Belief-changes in connection with actions or events are of an essentially non-monotonic nature, Doyle claims; at one moment the robot believes that the vase is on the table, in the next moment, as a consequence of a move operation, it believes that the vase is on the floor.
and - voilà - the old belief that the vase is on the table has disappeared. We must not forget, however, that there already is a solution to the problem of vanishing beliefs, which is to introduce situations, so that a belief is always connected to, or restricted to, a situation. Non-monotonicity in the sense of vanishing beliefs can in fact be modeled by ordinary monotonic logic.

There seems to be a conflation between certain modes of reasoning and certain technical peculiarities of ordinary logic. It should be clear that monotonicity and non-monotonicity are properties of formal systems, not properties of reasoning, nor of the world. Overlooking this fact results in confusion. It is a trivial fact about non-static problem-worlds, that representations of the world at one moment will have to differ from representations of the world at other moments; provided we agree at all with the possibility of isolating representations of the world at one moment from the total representation. (Notice that this does not imply that the representation itself has to change with time.) Where is the non-monotonicity? Only in the eye of the beholder; non-monotonicity is an artefact of logic.

Then what makes many of the proposed non-monotonic methods of reasoning interesting and new is not that they have a non-monotonic appearance when formulated in the language of logic, but other things, such as their use of knowledge about knowledge or lack of knowledge. It is unfortunate that the term non-monotonic has come to stand for these important but poorly understood modes of reasoning. By this misnomer, questions as disparate as reasoning from lack of knowledge, and the problem of vanishing beliefs is brought under a common heading, a heading which can easily lead one to believe that these are all just technical problems of logic.
Although I think we have now covered the main themes in the work on the frame problem, there remains quite a few research ideas and results that seem to have some relation to this problem. Some of that research is performed outside of AI. I will quickly go through a few of the contributions in this category, without going into details.

Non-monotonic logic, Circumscription, and Default reasoning, are three different attempts to breathe new life into the deductivist tradition by extending logic to allow for non-monotonic inferences.

**Non-monotonic logic** (McDermott & Doyle [1980], McDermott [1982]) is the most general approach. It is an attempt to develop a radically new deductive system permitting non-monotonic inferences of a general type. A complete and rigorous formulation of such a logic does not yet exist; for example, the model theory is incomplete and it is doubtful whether it can be completed (Davis [1980]). These doubts seems to be shared by McDermott himself: "...[non-monotonic logic] is not that well understood, and what is well understood about it is not all that encouraging." (McDermott [1981:23]).

**Circumscription** (McCarthy [1980]) and **Default reasoning** (Reiter [1980]) are more specific about the type of non-monotonic inference. Circumscription is based on the idea of jumping to conclusions: "...the objects that can be shown to have a certain property P by reasoning from certain facts A are all the objects that satisfy P". Default reasoning is based on the idea of defaults, a set of rules to use when information is lacking.
Dynamic logic (Harel [1979]) is a rather conservative extension of ordinary, static, logic to handle processes conveniently. It is thus explicitly concerned with time and change. The research is mainly motivated by applications to program verification and program synthesis, but applications to logic of action have been considered as well. Dynamic logic is solidly founded on the situation concept - despite some notational conveniences and a more elegant treatment compared to the situation calculus, there is nothing really new with respect to the frame problem.

On the whole, there is less to learn from temporal logic, than might be expected. One reason is that philosophers and logicians working in this field have paid little attention to common-sense time, naive time. Here, I will only consider a contribution by McDermott [1981b], that has no obvious relation to classical works on temporal logic, such as Prior [1967]. Again, we find the situation concept at bottom, but the concepts of action and event are less restricted than in the original deductivist paradigm. For example, events are not identified with state changes, and are allowed to take time. The frame problem is addressed somewhat en passant with the hope that this temporal logic will make it possible to reason about facts from the side, inferring that they are true for whole stretches of time. Although several new concepts are introduced, and although many interesting observations are made (e.g. "when we see a boulder, our eyes are telling our data base about a persistence, not about an instantaneous fact"), it is at this stage difficult to get an overall picture of McDermott's theory, and in particular to appreciate its relevance to the frame problem.
Metaphysics is a key concept in my attempt to analyze the frame problem - actually, I think, a concept essential for a proper understanding of a wide range of representational issues in AI. Metaphysics is sometimes thought to be a rather obscure topic, and there are metaphysical issues that I believe are totally irrelevant in this context - materialism vs idealism is a typical example - so it is perhaps in order that I make myself a bit more clear.

Some examples of metaphysical questions relevant to the frame problem are: What are the fundamental entities? What are the fundamental presuppositions? What basic categories are there? What basic principles? To call these questions metaphysical is to indicate that they are related to empirical evidence only in a very indirect way. It is not important to the arguments and conclusions of this paper whether metaphysics is taken in the classical absolute sense (Aristotle, Descartes, Kant) or in the modern relative sense. In the classical sense, metaphysics is about the way the world fundamentally is, or the way it is necessarily apprehended. In this sense there can be only one true metaphysics. In the modern sense metaphysics concerns fundamental presuppositions, conceptual frameworks in general. Several different, to some degree interchangeable or complementary metaphysical systems are conceivable. These
may differ in efficiency, fruitfulness, or heuristic power, and may be more or less specialized to different aspects of the world.

Whether the metaphysics is believed to reside in the world or in the mind, it will have to be deliberately put into the problem solving system. Of course, every system will have some metaphysics. The point is that it matters which, and that it will not be the right one by chance, or by some force of nature, or by some process in which the designer of the system unwittingly transfers his own metaphysics. All these possibilities are contradicted by experience from the frustrated efforts to solve the frame problem reported in part 2, above.

To avoid confusion, let common-sense refer to the common-sense view of common-sense. The metaphysics actually used by human beings in a common-sense world, I will refer to as the operative (human) metaphysics. Many of the metaphysical systems that have been advanced by philosophers do not appeal to common-sense. (Some do, but they still end up very different from our common-sense view.) On the contrary, the typical metaphysics is very different from common-sense.

Why is AI then confined to common-sense? Would perhaps a more sophisticated, less naive, metaphysics solve some problems? I think that very few researchers within AI has pondered much over these matters, but one can think of various reasons for sticking to common-sense: For example: (1) our operative metaphysics is the true metaphysics; (2) since it is difficult to construct a suitable metaphysics, let us use the one that is already there, and which we all know (now of course, our operative metaphysics is not as unambiguous and easily accessible as that). A more powerful argument is the argument from communication:

54 PART 3: DISCUSSION AND CONCLUSIONS
problem solving systems and human beings will not be able to communicate if they do not have a common way of categorizing the world. If a robot is to be of any significance to a human being, they have to coexist in the same conceptual world. The robot will have tremendous difficulties in picking up a chair if thing is not among its fundamental categories; or rather, if it cannot easily decompose thing into those fundamental entities. Awkward translation processes will hamper its activities, its observations and reasoning. To understand a command, for example, it will have to unload the command from the user's metaphysical vehicle and then reload the information on its own metaphysical vehicle.

Thus, it appears that, at least as long as we busy ourselves with common-sense problem worlds, it is not advisable to deviate much from common-sense. We must be careful, however, not to make the mistake of assuming that our operative metaphysics is trivial and lies open to inspection for anyone who cares to look for it. Evidence to the contrary come from philosophy, psychology, linguistics, and the frustrations of AI. What goes under the name of "common-sense" in AI is mostly a curious mixture of Aristotelian and Newtonian mechanics, not strikingly common-sensical, and not to be uncritically accepted as the operative metaphysics. Comparing common-sense with science we would expect that the operative metaphysics is much more sophisticated and complex than the metaphysics of science since, contrary to popular belief, the world of common-sense is enormously more

---

1) A rather isolated example of a more serious effort within AI to lay bare a part of common-sense physics and metaphysics are the works by Patrick Hayes on naive physics [1978a,b], although he does not make what I consider to be the crucial distinction between physics and metaphysics.
complex than the world(s) of science. A major reason for the success of science is probably its simplified metaphysics. It may be that the following quote from Popper [1969:137], has some relevance also to AI:

"There is a widespread belief, somewhat remotely due, I think, to the influence of Francis Bacon, that one should study the problems of the theory of knowledge in connection with our knowledge of an orange rather than our knowledge of the cosmos. I dissent from this belief, ..."

The irritatingly slow progress in the research on common-sense reasoning could be an indication that we are prematurely attacking the really hard problems before we have mastered the more tractable ones.

AI has, on the whole, overlooked the importance of metaphysical deliberations, which has resulted in a proliferation of unpremeditated or ad hoc metaphysical assumptions; metaphysical questions can be avoided, metaphysical commitments not. I think that finding and implementing the adequate metaphysics is a promising approach to the frame problem, as well as an important key to successful modeling in AI in general. Of course, it is far from obvious what an "adequate metaphysics" is, philosophers have searched and quarreled for thousands of years. Rather it is my belief that a better awareness of metaphysical issues, and more conscious efforts to explore and represent adequate metaphysical systems, are important elements of a viable research strategy - a necessary counterbalance to the prevailing tendency of letting work on representational issues be governed by which technical opportunities are seen.

To me, it appears that "experimental ontology" is an apt description of much of what goes on in Knowledge Representation in AI. I don't claim that philosophy

\[2\] This should stir the philosophers, unless they think that philosophy should stick to unsolvable
has the answers, but I think that any significant progress will hinge on our ability to acquire a philosophical perspective.

The situation concept

From the critical examination in part 2, it appears plausible that the situation concept is an unsuitable foundation for a metaphysics of change. We have found that STRIPS, PLANNER, and TMS, all of which make no explicit use of situations, have in common that they need not consider things that do not change. Unchanged facts simply remain. In contrast, every deductivist system has to process every fact, including those that do not change. This should not be confused with the fact that logic is permissive, not imperative (Minsky [1974]), which is something different. The permissiveness of logic makes it necessary to detach the heuristic component of deductivist systems from the logic itself, but this does not affect the modeling.

The situation concept goes back to 17th century physics. But in current physics there is nothing left of its former metaphysical status; although it is still sometimes used as a convenient computational device, it is not granted any real significance, being outdated by the theory of relativity and by quantum theory. Neither is there any evidence that the situation concept is part of our operative metaphysics. If one believes (as I do) that the operative metaphysics is at least partially reflected in natural language, linguistics offers a wealth of evidence for contrary hypotheses about our fundamental problems.
operative concepts of time and change: rather than latitudinal, absolute entities, one finds longitudinal, relative entities.  

What still might seem to speak in favor of the situation concept is its simplicity. But this is a devious simplicity. What is in a moment? Time derivatives, dispositions, past facts, past possibilities, even future facts and future possibilities? There seems to be no limit to what must be loaded into a single situation when trying to make more accurate and clever problem solvers mastering ever more complex environments. We risk ending up with a concept of a situation that encompasses literally all there is and has been and will be and a little more. 

But the most important flaw in the situation concept is that it simply does not seem possible to give definitions useful in common-sense type of worlds, of concepts like action, event, change, etc, in terms of the situation concept. This is true even for almost trivial problem worlds. The long record of failures to solve the frame problem bears witness to this.

Explicit, Primitive, and Primary

I will exemplify my view on the role of metaphysics in relation to the frame problem by discussing three pairs of concepts - explicit-implicit, primitive-non-primitive, and primary-secondary - and their relationship.

3) The entirely different situation concept of Barwise and Perry [1981] is one example of an approach to time and change that lies much closer to common-sense.
The terms explicit and implicit have a clear meaning when applied to formal systems. Axioms are explicit, and everything else that can be inferred from the axioms using the rules of inference is implicit. This is satisfactory from a logical point of view, but those concerned with the algorithmic dimension of formal systems, either performing or engineering inferences, have a need for distinguishing between information actually inferred, and information that is in principle derivable, but is not at hand, either because it has in fact not been inferred, or it was inferred but the result was subsequently dropped or deleted.

Let us reserve explicit-implicit for distinguishing between information that is actually at hand (the axioms as given), and information that is not (but can be inferred from information at hand); then we shall take primitive-non-primitive to stand for what remains of the original logical explicit-implicit distinction, which is a distinction between independent foundational information (the axioms as a source), and dependent superstructural information, derivable from the foundational (primitive) information. Both distinctions are the offspring of the notion of derivability, and clearly, where both distinctions apply, implicit information must be non-primitive, and primitive information must be explicit.

Within the deductivist tradition, committed to formal systems, these distinctions are obviously applicable. Outside of the deductivist tradition we find that in

3) Here, I don't take primitive to imply that it is impossible to derive primitive information from other primitive information (nor, of course, from non-primitive information). In the end, primitive information is primitive in virtue of being designated as such.
STRIPS the facts are formulated using a formal system, and even if we depart from the STRIPS-like systems of the non-deductivist tradition but retain a discernible database, the explicit-implicit distinction still applies: sentences in the database are explicit, sentences inferable from the explicit information using the methods of the system are implicit. A policy for administrating how the contents of a model is distributed into explicit and implicit, usually rests on some specific notion of primitive information, meaning that also the primitive-non-primitive distinction, is applicable.

Whether information is explicit or implicit, primitive or non-primitive, are formal questions, decidable without reference to the modeled world outside the system. A given body of information can be axiomatized in many different ways, if it is at all axiomatizable. There is a considerable freedom of choice in deciding what should be explicitly and what should be implicitly represented, what should be taken as primitive and what as non-primitive.

Keeping information implicit may in certain circumstances have the advantage, beside the obvious one of saving space, of facilitating the maneuverability and robustness of the model. As an extreme, consider a completely explicit representation (if such a thing is possible): this will be a body of information with a very rich underlying structure, a great number of relations between different parts of the representation. When we wish to make a change in the model, to reflect a possible change of the represented world, we have to make adjustments in many different parts.

4) Yet, there remains a certain relativity in applying the distinction. Perhaps it would be better to introduce a concept of degree of implicitness, defined, for example, as the number of elementary operations needed to produce the information.
of the model; the model will be cumbersome to handle, adjustments cannot be done efficiently. There is also a great risk of introducing inconsistencies; the high degree of interdependence between different parts makes it very probable that an adjustment that does not completely cover all relevant parts of the model will result in an inconsistent model.

The other extreme of using a minimally explicit representation, a form of representation close to the mathematician's ideal of independent axioms, amounts to an "orthogonalization" of the description in, ideally, completely independent parts; a change of one part can never lead to a conflict with other parts. Thus, to be robust, the model should be approximately orthogonal. However, to ensure a maneuverable model, it is also necessary that the choice of "base vectors" corresponds to the structure of the problem world in such a way that the relevant changes and actions can be easily analyzed into components that correspond to the base vectors spanning the description space. Only when this is the case is the efficiency and reliability of the type of adjustments that actually may occur, guaranteed.

Explicit, implicit, primitive, non-primitive, are all properties that apply to the representation. Primary and secondary, on the other hand, are properties that relate to the world. Secondary phenomena are in some sense less real, they are epiphenomena in principle completely reducible to primary phenomena. Conditions of the world described by secondary information are in principle reducible to conditions described by primary information. This means that given a specific notion of primary properties we have

5) Here, the vector space metaphor breaks down, since we can not expect that the description-space is anything like linear.
a basis for an internal notion of primitive information, corresponding to dependency relations that can be exploited to represent secondary information implicitly. Which information should be stored explicitly is thus a question that can not be adequately answered without considering the world and its metaphysics.

Of course, the whole idea of a primary-secondary structure, i.e. reductionism, can be questioned. Is it possible or even desirable? The important thing about implicitness is not that it saves space. Organization and access is the vital point. Space is practically unlimited (or will soon be), time is not. Explicitness is in the end self-defeating: the more we have at hand, the greater trouble we have in quickly locating a particular datum, that is, the less at hand each single datum becomes. Something we might call the immediately-at-hand-ness of the model is strictly limited, and we must learn to use it wisely.

Intrinsic representation

The explicit situation concept of the deductivists exemplifies an ambition to keep metaphysics, rudimentary as it may be, accessible to the system. This is not a consequence of some deeper considerations about the role of metaphysics - it is just part of a general deductivist aim to keep as much as possible as contents of the model, directly accessible to the system, and in fact, metaphysics is frequently mixed up with physics. What I said about the situation concept seems to imply that we should abandon the kind of physicalistic modeling favored by the deductivists. That is not to say that non-deductivist
models are phenomenologically inspired - what may appear as a tendency to subjectivism could also be an effect of trying to implement the system's metaphysics as a capacity rather than as explicit knowledge. Ideally, the metaphysics is built into the system, so that it becomes embodied in the form, or medium, of the representation. The system simply obeys it, without being aware of it. The metaphysics is then intrinsically represented. In contrast, the deductivist models are highly extrinsic: logic, the building blocks of the model, has few natural isomorphies with the world. It is much more general; the specific form of the actual world has to be externally imposed on it.

An example (adapted from Palmer [1978]) will help to make the difference between intrinsic and extrinsic representation more clear; note that the example has nothing to do with metaphysics. The represented world consists of three objects of differing lengths. The only things we care to represent are the identities of the objects and the relation is longer than. In the first representation there are three objects each representing a particular object of the world. The is longer than-relation is represented by the relation is arrow-connected to that holds between representing objects connected by an arrow. The second representation also uses three objects to represent the objects of the world, but the relation is longer than is here represented by the relation is wider than. The first representation is extrinsic with regard to the is longer than relation, the second representation is intrinsic. For example, is longer than is a transitive relation, which is wider than also is but is arrow-connected to is not; in this case the transitivity has to be impressed on it and externally maintained. An elaborate and clarifying discussion of these concepts can be found in Palmer [1978].

- Intrinsic representation
The greater generality of the extrinsic representation makes it more resource demanding and more cumbersome to handle. It is also more error-prone since it has more possibilities to disagree with the world. The advantages of intrinsic representations are: swift operation, moderate requirements on explicit information, and security - some impossible world states can simply not be modeled.

It should be emphasized that the motive for making the metaphysics explicit to the designer is more compelling than ever. If, as I claim, the metaphysics has a central role in efficient modeling, and should be intrinsically represented; then we are faced with the double task of finding or constructing a suitable metaphysics, and implementing it intrinsically. In practice there will be a dialectic interplay between these two tasks. We need to be philosophers and computer scientists alike.

6) I don't argue that the physics should be intrinsically represented. There is little hope that the (common-sense) physics of the world in all its variations and aspects can be intrinsically represented (Pylyshyn [1980]). Metaphysics, on the other hand, is a relatively small, unified (and unifying), thing.
Acknowledgements

The major part of this paper was written while I was a research associate on the project Modeling Time at the Departments of Linguistics and Information Processing, University of Umeå, supported by the Swedish Humanities and Social Sciences Research Council (HSFR), 79/80 - 81/82.

Bo Dahlbom has criticized, encouraged, and helped me in various ways through several versions of this paper. I have also received valuable support and comments from Eva Ejerhed, Axel Ruhe, Martin Edman, Peter Gärdenfors, Kenneth Kahn, Staffan Löf, Zenon Pylyshyn, Ingemar Widegren, and others.
REFERENCES


Hayes, P. J. (1978a) The naive physics manifesto, Institut pour les etudes semantiques et cognitives, Universite de Geneve, Fondazione Dalle Molle, Working papers No. 34.


REFERENCES


Sandewall, E. (1972a) PCF-2, a first-order calculus for expressing conceptual information, Computer Science Department, Uppsala Universitet, Uppsala.

Cause and counterfactual, Philosophy of Science, vol 33, 323-340.

Stallman, R. M. & Sussman, G. J. (1977)  
Forward-reasoning and dependency-directed backtracking in a system for computer-aided circuit analysis, Artificial Intelligence, vol 9, 135-196.

Waldinger, Richard (1977)  
Achieving several goals simultaneously, in Machine Intelligence 8, ed. by E. W. Elcock and D. Michie, Edinburgh, 94-136.

REFERENCES
une pipe

ceci n'est pas une pipe!

une pipe, même!

Pictures and words
In the task of building computer representations for a part of some real or hypothetical world we are faced with several design choices. One rather momentous choice would seem to be that between **verbal** and **pictorial** forms of representation. These notions are rather vague and intuitive. One purpose of this paper is to make them more clear. Another purpose is to find out if the choice between verbal and pictorial really is important, and if it is, how we should choose. Preliminarily, **verbal** will be used as a very general label for a category of representations alternatively called **linguistic**, **formal**, **propositional**, **descriptive**, etc. Similarly, **pictorial** stands for a category of representations that could also be characterized as **iconic**, **non-verbal**, **analogical**, **imagistic**, etc.

An illustration

A concrete triangle - a triangular thing - could be represented by a description:

---

1) I draw no sharp distinction between word and concept, sentence and proposition. In so far as these notions are cogent in this context, I will take them all as verbal symbols, possibly at different levels of abstraction, but on an equal footing as symbols.
Page dimensions: 337.7x641.5
[Image 0x0 to 338x642]

Flat thing with three straight edges; first edge is 3 units long, second is 4 units, and third 5 units

or by a picture:

![Figure 1.1](image)

Offhand, two points are rather obvious:

(1) there are certain pictorial qualities of physical objects - such as geometry, topology, color, etc - where a pictorial representation has advantages over verbal representations. In a pictorial representation it is apparently easy to find out what is in the neighbourhood of something, if two things touch, if one thing is inside the other, or on top of it, or above it, etc. This can be a chore using a verbal representation based on coordinates, say.

(2) there are certain operations or transformations affecting depicted qualities in a systematic way that are potentially more awkward to handle in a verbal representation: transportation, rotation, changing size, etc. It is, for example, non-trivial to find a path for moving an object from one place to another avoiding obstacles, if we have a coordinate-based representation.

The whole question of verbal versus pictorial representation would then perhaps seem to have the obvious answer: represent pictorial qualities pictorially, and let the rest be verbally represented; and that's that.
But then what exactly are "pictorial qualities"? Trying to evade a definite answer by saying that these are the qualities which are preferably pictorially represented will only transform the problem into the question what a picture is. Neither of these questions have an obvious answer, in particular not in the present context of computer application. And isn't this approach anyway too narrow? - what about diagrams, and in general pictorial representations that do not picture things that can be (directly) seen or even have spatial or other pictorial properties? - structure formulas in chemistry, population density charts, statistical diagrams, state space diagrams, etc?

On the other hand, verbal representations of (what ought to count as) pictorial qualities appear quite as indispensable. Different geometrical figures have names: rectangle, triangle, cube, sphere, etc; measurements, curvatures, slopes, etc. are often represented by numerals; physicists don't use color samples to refer to different wavelengths of light, probably because color samples do not fit into formulas, but why do physicists represent optics as formulas? etc. And think of art; we (or the critics, at least) are obviously not quite satisfied with just pictorial representations of a painting, say - but have a need also for descriptions, that is, verbal representations of its various qualities.

In practice, there is then a rather complex relationship between the type of things or qualities represented and the preferred form of representation, and also an interesting interplay between verbal and pictorial forms of representation. Let us consider some concrete differences between the verbal representation and the pictorial representation of our triangle example, to get a general idea of the sort of questions that arise in the context of...
choosing and using representational form.

(3) In the verbal representation we have to walk through the entire representation and make some calculations to decide whether a given corner is acute, orthogonal, or obtuse; in the pictorial representation it is enough to take a look at that corner. (Of course, with a different verbal representation it could have been otherwise.)

(4) A verbal representation of the form given can fail to represent a possible triangle by describing sides that violate the triangle inequality; whereas a pictorial representation can not fail like that. Or - to put it more cautiously and perhaps more precisely - it's more expensive to make sure that the verbal representation is a possible representation of a triangle than to verify that the pictorial representation is. (Actually, there are two questions involved here: - is it a picture? - is it a picture of something possible?)

(5) Is the color of the triangle significant in the pictorial representation? The lines are slightly wobbly - is this just an insignificant distortion, or should we pay closer attention to it? This sort of question doesn't arise in the verbal representation; we have no trouble deciding which features that count, and a shaky writing doesn't worry us (cp. also (9) and (10) below).

(6) The lengths of the sides are explicit in the verbal representation; in the pictorial representation we have to measure them. Coming to think of it - is there anything explicit about a picture? Or - is the trick here that a "length" is a kind of verbal representation so
that we have a bias towards verbal representa-
tions? Had we simply asked for something "representing the length", an alternative answer would have been to point to that linesegment in the picture.  

(7) The area of the triangle has to be calculated from the verbal representation; but using the pictorial representation we can measure this one too: for example by cutting it out and weighing it, or by painting it to see how much paint it takes. A perhaps rash generalization: we can pull out the information of a pictorial representation by various measuring procedures (or, in some particularly simple cases, observations); whereas the information in a verbal representation is either explicit or must be calculated from explicit information. But is there a clear distinction between calculating and measuring or observing?  

(8) Whereas the verbal representation seems to have a prescribed order in which it must be read off, there seems to be no similar restriction on the pictorial representation. The verbal representation is discursive, and the pictorial representation non-discursive. (But it's not very clear what that means - something about how the meaning of the whole is built up from the meaning of the parts.)  

It's not obvious exactly how to make the verbal representation of the triangle; in fact one has the impression that there is a richer range of possibilities in making a description than in making a

---

2) Think e.g. of the full-scale drawings used by the medieval cathedral builders; the stone blocks were simply put on top of the drawing to mark where they should be cut.
depiction. For instance, we could have chosen to say right from the start that it was a right-angled triangle:

*a right-angled triangular thing with sides 3, 4 and 5*

With this particular choice of verbal representation, we notice another difference:

(9) A small change in the triangle represented will correspond to a small change in the pictorial representation but possibly to a dramatic change in the verbal representation, namely if the right angle is transformed into an acute or obtuse angle.

There are also possibilities of generalizing the verbal representation in various ways, an extreme case being:

*a (concrete) triangle*

(10) The verbal representation is now completely general, it applies to any triangle, similarly we could have descriptions like "a right-angled triangle" that applies to any and all right-angled triangles. But it is very hard to imagine a pictorial representation of a general triangle (or even a typical triangle), or a picture representing any right-angled triangle, etc.

On the other hand, we can think of very different ways of making a pictorial representation of the triangle; we can choose different techniques, different perspectives, pick out or emphasize different aspects of it - all of which seems to have a parallel in verbal representation. But do we have quite the same freedom as in verbal representations?
Improbable, but not unthinkable, "square" could be a verbal representation of the triangle (in some language). But is it conceivable that the following is a pictorial representation of it?

![Figure 1:2](image)

Hardly. The verbal representation appears to be more conventional than the pictorial representation.

What, if anything, of these differences can be attributed to the difference between verbal and pictorial? In general: (i) do pictorial and verbal representations differ systematically in any important respects? (ii) are those differences related to cognitive performance? Intuitively, it is perfectly clear (to me) that the answer to both questions is yes. If that intuition should withstand closer scrutiny it is certainly a fact of great interest for any theory of knowledge representation, and for the practical design of representations and systems of representation.

The examples given are, to be sure, extreme cases (or at least pretend to be). One would expect that many representations should be put somewhere in the middle of a scale ranging from "purely pictorial" to "purely verbal". If the majority of interesting and useful representations should turn out to have that kind of doubleness or undecidedness - which I suspect they do - it is still a good idea to investigate the extremes. First, because knowledge about ideal cases often gives important insights into non-idealized
cases. Second, because there is hope that representations can be picked apart, or analyzed piecemeal using the ideal concepts: certain details of a representation represent pictorially, other details represent verbally.

Pictures in computers

Turning to the realities of computers and computer applications, we meet a rather heavy bias towards verbal representation. Considering the apparently verbal nature of current computer hardware - can it really accommodate pictorial representations? Part of one's hesitation originates in a common prejudice that pictures must be non-digital and continuous. But surely ordinary color prints, and pictures on a video screen are pictures! Actually, there are computers where the content of a video screen is a bit-by-bit hardware mapping of a part of the computer's memory; and indeed there are also a few special-purpose processors for picture processing.

Notice that it is not enough to have a dot-matrix (say) stored in the computer memory. Before we can claim to have a pictorial representation there must also be operations that treat that part of memory as a picture; for example, operations for accessing contiguous dots, recognizing various picture elements and features, operations for rotation, translation, scaling, etc. In other words: if a picture isn't used as a picture then it is not a picture in any interesting sense. Being pictorial depends on use. Even if a rectangular grid of transistors in a computer memory should be seen as a picture by us given that the states of the transistors were made visible, it doesn't mean that it is a picture to the computer.
Conversely, what is properly called a hardware pictorial representation in a computer memory is not from our (normal) point of view a pictorial representation at all; even if we could see it, the physical arrangement (in contradistinction to the access structure, the logical arrangement) of the memory cells most likely doesn't reflect the physical arrangement of the corresponding parts of what is depicted.

As far as hardware is concerned, it seems that pictorial representations, even if uncomfortable, still are far from being precluded. So, what has software made of this? Apparently nothing. Raising our eyes from hardware to software, we see but words. It is true that data sometimes enters or leaves in a pictorial form; for example, as signals, or diagrams. But programs are usually very quick at transforming incoming pictures to the appropriate internal form, which is almost invariably verbal, meaning that pictorial representations are confined to rather shallow, "raw" data, input and output interfaces.

And if we look to the programs themselves - if ever there was a verbal representation, programs certainly are! It is conceivable that a program could be a verbal description used only as a means for generating pictures in the computer. But, in general, this is not so. What finally exists and takes place within the machine does not appear to be pictorial at all (except in a few special applications).

Of course, even if pictorial representations are sparse and hard to implement, that's no excuse for underrating their usefulness and importance. But is this first impression correct? Do pictorial representation really play such a minor role in computer applications? Taking a less narrow view of what pictorial representation is, we will find that

- Pictures in computers
pictures are in fact quite common in computers. Partly it is a question of levels, partly a question of stressing the analogy aspect of pictures.

First, the question of levels; a representation that is obviously verbal at one level, may yet be aptly characterized as pictorial at a higher level. For example, a dot-matrix could be transformed into a set of statements: "red at (1,2)", "blue at (3,4)", "blue at (3,5)", etc. Obviously a clear-cut case of a verbal representation at this level. But if this description is embedded within a set of picture-oriented access-methods and operations consistent with the implied pictorial qualities of the description - then at this higher level it would appear to have turned into a pictorial representation, and anyone who didn't look deeper would presumably never know "the truth".

Conversely a verbal representation may turn out to be based on a pictorial representation at a lower level. Consider individual replicas of the letter "A" - are they not little pictures of one another or some archetypical "A"? We can imagine a system that stores information about the environment with the help of a little scale model viewed through a TV-camera, but has an interface to this information that turns out descriptions like "(CENTER-AT BOX14441 34.21 441.11)".

Secondly, there is the analogy aspect. If we, for example, represent a bus-queue with an array in the obvious manner, the order of persons in the queue is not described, but pictured: the sequence of person-designators in the array mirrors the sequence of persons in the queue. A moment's reflection on this trivial example leads to the conclusion that if we accept this as pictorial, then there are lots of pictorial representations, at least small and partial
ones, in just about any computer application. One particularly common example is our habit of depicting a time order as a spatial order.

**A note on "representation"**

Sometimes, "representation" is used to refer to the concrete entities (data structures, processes, states, or whatever) that are used to represent, without involving a precise notion of what features are significant, how the symbols can be accessed and operated on, etc - which is then referred to as the "use" of the representation. I think this usage invites confusion and misunderstanding.3

Alternatively, we may take "representation" to include both the representing entities and their use. **Elementary use**, we must add, to prevent the concept of representation from taking over completely: for example, one purpose of having a concept of representation is to make a distinction from reasoning. Now, what we mean by "elementary" is to some degree a matter of choice, it is relative to what view we wish to take of a representation. In an hierarchically organized AI-system we can choose between different possible organizational levels where to make a horizontal cut and study the relation of what is above - the applications (non-elementary operations) - to what is below - the representation (including "elementary" operations).

It is in the latter, operational, sense "representation" is used in this paper. The concrete entities involved in a representation, I will refer to as

3) For example, it makes an expression such as "pictorial representation" dubious, since the pictorial character of a symbol is (at least) a product of the symbolic object and the operations which define how it is accessed and manipulated. Compare the above examples.
symbolic objects or symbolic entities. "Representation" is ambiguous also in a different sense: it may signify a system for representation, it may signify a particular individual symbol in a system for representation, and it may signify the state or relation of representing; but this is on the whole a rather harmless ambiguity, sometimes even convenient, so I have decided to keep to common usage on this point. "Pictorial representation" and "verbal representation", of course, inherit the same ambiguity. A "picture" is a symbol involved in pictorial representation (the more exact meaning of which will, hopefully, become clearer as we go on); a "depiction" is either a pictorial representation or an act of creating a pictorial representation, a term I will sometimes use by way of variation.

It will be convenient, and improve clarity and precision, to supplement this concept of (concrete) representation with the concept of virtual representation, by analogy with the virtual machine concept. A virtual representation abstracts from everything but the fundamental questions of access and manipulation of the symbols in the representation. It enables us to concentrate on the elementary tests and questions that can be put to the representation and the elementary manipulations that can be made on it, disregarding how it is implemented. Questions of implementation can be dealt with separately.

A virtual representation is defined by a set of tests (predicates), questions (functions), operations, and a set of constraints on the results. Some of these

4) Speaking of "predicates" and "questions" in this definition must not be taken to mean that the raw interface to a representation is verbal, it is just a way of describing what is going on. Generally, it is extremely difficult to avoid a verbal bias in formulations like these. Note also, that it may be non-trivial to determine the exact virtual representation in a concrete representation.
constraints may be **timing restrictions**; for example, in some "pictorial" representation we may not wish to allow that the time for accessing two different "points" in succession is proportional to the square of their "distance".  

---

**Pictures in knowledge representation**

There have been few purposeful attempts at a more substantial exploitation of pictorial representations for other purposes than picture-processing and interfaces to pictorial input and output; the two most important examples are Gelernter (1959) and Funt (1976). To be sure, there's something to learn also from AI-applications using pictures in more surreptitious ways; interestingly, some recent approaches to machine learning, can be seen as involving what is essentially a pictorial mode of representing.

Theoretical work by researchers and philosophers of AI directly adressing the pictorial/verbal issue, is also meager, even if many topics in knowledge representation touch on the distinction between verbal and pictorial.

Just outside AI we find a parallel discussion in cognitive psychology. In contrast to AI, there is a vivid debate and a considerable amount of experimental research going on. This debate may not be entirely relevant for knowledge representation, but it does play a role both as a kind of general frame of reference, and as a source of inspiration, stimulating inquiries into the foundations of

---

5) This description of a virtual representation fits in quite well with the notion of an abstract datatype; one difference is that I allow for timing constraints. Of course, AI has more ambitious and more specialized plans for its representations than computer science for its repertoire of fundamental data types.
representation in general.

In short, not very much has been done. On the other hand there is a number of questions and problems about verbal and pictorial representations in AI that seem well worth a closer study. Most of them can be grouped under two general headings:

First, there is what we may call the **characterization problem**: how are the different forms of representation distinguished, internally? What makes pictures different from words as representations? What is the essence of a verbal representation as distinct from the essence of a pictorial representation?

Second, we have the (let us call it) **performance problem**: how is performance related to the form of representation used? How do AI systems operating with pictures differ from a systems operating with words? Are there, from an external point of view, really important differences between these forms of representation, and if so - what are they, and how can they best be exploited?

The relation between these two problems is rather complex - we need the characterization of words and pictures to be able to investigate the performance problem, but clearly much of the evidence for having made a fruitful distinction will come from performance considerations, so really the two problems are mutually dependent. Notice also that the answers are relative to how much we choose to include in the "representation" and the "system" using it, respectively, in the formulation of the performance problem. Of course, the performance level should be distinct from the representational level, so that it is reasonable to talk about their relation to each other.
The general plan for this paper is first to explore and analyze various theories and ideas about pictorial representation and how it is related to verbal representation, in pursuit of interesting questions, insights and suggestions that appear fruitful, but also noting defective and false proposals on the way; and subsequently to outline a theory that takes as much as possible of the knowledge gained in this way into consideration. It is natural to concentrate on pictorial representation, considering the predominance of verbal representation, both in computer practice and in theories of representation.

In the explorative part, I will refrain from pushing any particular thesis. Rather, I will try to find faults with everything that comes in sight. Four different areas are considered. Chapter II considers the explicit discussions of the pictorial/verbal issue in AI. Chapter III covers some relevant theoretical work in psychology and cognitive science, mainly connected with the so-called "mental imagery" debate. Chapter IV presents and discusses various philosophical theories of pictures, in particular the theory of Nelson Goodman. Chapter V takes up some experiments with pictorial representations in AI.

Chapter VI contains the constructive part, an outline of a theory of pictorial representation and how it is related to verbal representation, mainly inspired by Goodman's theory.
Theoretical work on pictorial representation in artificial intelligence, is sparse. One reason may be the firmly rooted beliefs that "pictorial" has to do with "continuous", which seems difficult to reconcile with the digital nature of current computer technology. Such misconceptions have been vigorously fought by Aaron Sloman, the pioneer theoretician of analogical representation. A different kind of skepticism is expressed by Daniel Bobrow:

"Having representational dichotomies such as analogical versus propositional requires, I think, that we make overly sharp distinctions." (Bobrow 1975:31)

Instead,

"I believe that such debate is best viewed by considering claims made along separate dimensions." (1975:32)

Examples of such dimensions are access, mapping process, inference, matching, etc (ibid.). Now, the point with a dichotomy is often not precision but comprehensiveness and insight. The purpose of discussing a distinction between verbal and pictorial is certainly not to be able to classify any representation as either verbal or pictorial, but rather to explore the verbal - pictorial "dimension", on the assumption that this is an interesting and important aspect in understanding and designing representations. (Which remains to be seen.)
Aaron Sloman's discussion of the pictorial/verbal issue (1971, 1975, 1978, 1985) is the fullest and most ambitious in the AI-related literature. His main theme is that verbal (logical) methods of representation and reasoning are not necessarily always the best ones to use in AI systems. Pictorial representation and reasoning is perfectly possible and valid, and has properties that definitely sets it apart from verbal representation and reasoning; properties which may give important advantages in some circumstances.

The core of Sloman's theory is the distinction between analogical and Fregean (or applicative). According to Sloman, what he calls Fregean is sometimes referred to as symbolic, linguistic, formal, propositional, or verbal; other terms that at least partly overlap with his use of analogical are: iconic, non-verbal, intuitive, and pictorial. Logical and mathematical notation come nearest to being purely Fregean representations, whereas many aspects of natural languages fail to conform with Fregean characteristics. We should also keep in mind that "the Fregean/analogical distinction does not exhaust the variety of important kinds of symbolising relations" (1978:167). On the whole, it seems reasonable to view Sloman's analogical-Fregean distinction as a partial explication of the pictorial-verbal distinction, even though natural language is not Fregean.1

1) Certainly, there is no need to assume that natural languages, the most advanced and complex forms of verbal representations we know of, are at all representative for the entire class of verbal representations.
Fregean and analogical

Some important characteristics of a Fregean, or applicative, representation are:

(a) it is applicative, which according to Sloman means that what a complex symbol represents is a function of what the parts of the symbol represent;

(b) a complex symbol can be interpreted as a procedure (constituting the "sense" of the symbol) for identifying what the symbol represents (the "reference"), (moreover the symbol is an analogical representation of this procedure);

(c) a complex symbol has parts that do not represent parts of what the symbol represents;

(d) the structure of a complex symbol gives no information about the structure of what it represents.

On the other hand, we have the contrary characteristics of an analogical representation:

(a') it is generally not applicative;

(b') a complex symbol can not as directly as a complex Fregean symbol be interpreted as a procedure for identifying what it represents;

(c') the categorematic parts (= parts representing on their own) of a complex symbol always represent parts of what the symbol as a whole represents;

2) Based on Sloman 1978. I am grateful for his comments on an earlier version of this section.
(d') the structure of a symbol gives information about the structure of what the symbol represents;3

It is evident that the analysis is largely based on the notion of a complex, structured, symbol, having "parts" (which I will call subsymbols). Sloman doesn't elaborate on this - I will simply assume that a subsymbol satisfies the following conditions: i) it is a (proper) part of the complex symbol; ii) it is a symbol in the same symbol system as the complex symbol; iii) it contributes to the meaning of the complex symbol. (Possibly, there are other requirements.) Hopefully, some of the ambiguity of the first requirement is dissolved by the other two requirements.

Notice that a subsymbol may be complex, and that it may be syncategorematic, that is, not represent anything by itself, only in combination with other symbols. Of special interest are the (let us call them) maximal subsymbols of a given symbol X; a subsymbol of X is maximal if it is not a subsymbol of any subsymbol of X. Since the discussion rarely involves other than maximal subsymbols, I will shorten "maximal subsymbol" to "subsymbol" in the following.

There are some additional features of analogical representations that Sloman draws attention to:

(e') analogical representations are generally less likely to represent "impossible objects" (be uninterpretable);

3) Some of the properties (a) - (d') are not explicitly stated by Sloman, just hinted at by the way he puts a property of the one type of representation in contrast to the other type of representation.
the mapping between symbols and objects is (likely to be) "smooth": small changes in the representation are likely to correspond to small changes in what is represented;

constraints in a problem situation are easily expressed by constraints on the kind of syntactical transformations permitted;

analogical representations are often economical, both statically and dynamically;

Presumably these properties are not to be taken as criterial; rather they seem to be connected to what I called the performance problem in the Introduction.

Applicativity

I will compare and discuss each of these points, beginning with (a) and (a') which concerns applicativity - apparently the core of the distinction between Fregean and analogical. Notice that the definition of applicativity excludes both referential opacity, and context dependency. So this introduces rather severe restrictions, which are not met, for example, by large parts of natural language.

Applicativity entails compositionality, that much is clear; what each complex symbol represents is a function of what each of its (maximal) subsymbols represent in isolation and - the syntactic relations between the subsymbols. It makes a difference if it is a function of the maximal subsymbols or if it is a function of all subsymbols, non-maximal included. In the latter case we have a context-free but possibly non-compositional representation. An example:
Interchanging "1" and "2" preserves the reference of "(1+2)" but changes the reference of the symbol as a whole from "9" to "18".

Incidently, here is an example of the delicacy of the notion of a subsymbol: should instances of syntactic relations perhaps be counted as subsymbols? The crux is whether they pass as "parts" of the complex symbol - ethereal as they may seem. Ontological scruples aside, I can see no serious formal obstacle, so - for convenience - I will let "subsymbol" include syntactic relations in the next few paragraphs.

Now, according to Sloman, applicativity means more than compositionality: it means that a complex symbol consists of a function-symbol and a number of argument-symbols, such that applying the function represented by the function-symbol to what is represented by the argument-symbols, we get as result the representation of the complex symbol as a whole.

As I understand it, however, with some effort any compositional representation can be viewed as an applicative representation. The only real difference is that the function involved in the definition of compositional is implicit - being always the same there is no need to represent it explicitly. Or - if we really must have a "function-symbol" - pick any subsymbol and designate it the "function-symbol".4

4) Clearly, if what the complex symbol represents is $f(a_0, a_1, a_2, ..., a_N)$, where $f$ is the global composition function and $a_0$, $a_1$, ..., $a_N$ are the subsymbols and syntactic relations, then we may pick $a_0$, say, and view it as a function defined as:
Preferably, some instance of a relation is chosen, since it will probably not lead to quite as violent a clash with intuition. For example, a picture of a glass on a table might be analyzed as 
\[=on=(=glass=,=table=),\]
where "=glass=" and "=table=" denote the subpictures, and "=on=" their syntactic relationship.

For comparance - surely both representations below are applicative?

\[2 \ast \ast 3\]

If the procedure seems arbitrary it is mostly because it is described in general terms. Besides, it follows from essentially the same argument, that any applicative representation permits of alternative interpretations in terms of function plus arguments decompositions.\(^5\)

To claim that pictures are generally not applicative, is then - in effect - to deny that what a picture represents is determined by what its subpictures represent and their syntactic relations to each other. Or so it would seem. In a later paper (1985) it appears that the chief obstacle to applicativity in pictures would be

\[a0(a1,a2,\ldots,aN) = f(a0,a1,a2,\ldots,aN).\]

5) For example, "the wife of Tom" interpreted by Sloman as the-wife-of(Tom), could alternatively be analyzed as Tom(the-wife-of), where "the wife of" represents a certain set of ordered pairs of women and men, and "Tom" represents a function which applied to an empty argument picks out Tom, but applied to a set of ordered pairs picks out the first element in any pair that has Tom as its second element. "1+2" may be analyzed not only as +(1,2), but also as 1(+,2) (as in object-oriented programming).

- Aaron Sloman
"the principle of substitutivity: in an applicative formalism, symbols and well-formed combinations of symbols will be divided into syntactic types, such that in a complex well-formed formula, any well-formed sub-formula can be replaced by another (simple or complex) of the same type and the result will still be well-formed. Moreover this will not affect the significance of any other symbol in the formula." (Sloman 1985) 6

This principle is generally violated by pictures -

"E.g. in a picture of animals in a field, there is no way that a picture of an elephant can be replaced by a picture of a giraffe, without it also making a difference to how much of the background is depicted." (ibid.)

Do we have here an additional necessary condition for applicativity in Sloman's sense, a condition that is violated by pictures? Not necessarily so. What reason is there to assume that pictures of animals belong to a single syntactic class? Perhaps it is more reasonable to say that (sub)pictures with the same outline form a syntactic class? As Sloman implies, we expect subsymbols of one and the same syntactic class to be substitutable for each other (preserving well-formedness) - this could actually be taken as a definition of what a syntactic class is.

Then what his example shows is that giraffe-pictures and elephant-pictures in general do not belong to the same syntactic class. 7 Any attempt to remove the

6) This "principle of substitutivity" should not be confused with the more common notion of substitutions preserving truth and/or meaning.

7) An exception would be if they were depicted posturing from some carefully chosen angle, so that their outlines came out the same.
giraffe and put in the elephant would ruin the well-formedness - there would be holes in the picture. 8

From the above quote it appears, however, that Sloman wishes to put the blame on semantic problems, not syntactic. It is possible to read "without it also making a difference to how much of the background is depicted" as implying that a picture of something represents only (some of) its visible parts. 9 The further implications of such a reading are far-reaching. It would mean that not in the whole history of art is there a single portrait of a person, it's all pictures of people's husks; it would mean, that a passport photography of Tom doesn't depict Tom, nor Tom's head, but only the outermost layer of Tom's face. And, of course, if there was a fly sitting on his nose, the picture wouldn't even represent as much as that.

Surely, this is somewhat counterintuitive. So, what if we remove the elephant outlined against the empty blue sky and put in a giraffe instead, filling in with blue as required - doesn't the sky-part of the picture still represent the empty blue sky? The sceptic might answer - yes, but it also represents that "this part of the sky is empty, that part of the sky is empty, ..." and these lists will be slightly different for the different pictures. Still, most pictures are not made with the purpose of hiding important aspects, so in many circumstances it is

8) The objection that the new subpicture might cover the old completely and so leave a well-formed picture as a result, does not hold since syntactically equal symbols should be mutually substitutable. Besides, it is doubtful whether we should even consider substitutions that make syntactic changes to surrounding symbols.

9) Incidentally, notice that many pictures have a kind of background that doesn't represent anything at all, it is there only as a necessary contrast, just as a "3" needs a contrasting background to exist.

- Aaron Sloman
reasonable to claim that if the sky-part of the picture represents empty sky then it represents that also the invisible parts of the sky are empty.

Sloman also presents direct evidence that pictures may be highly context-dependent. In a perspective drawing he demonstrates how lines of the same length in the scene may be depicted by lines of different lengths in the picture, and how lines of different lengths in the scene may be depicted by lines of the same length in the picture (1978:165-166).

Hopefully, the following is an essentially Fregean expression:

A ditch, 5 decimeters wide, 5 kilometers long, and half a meter deep.

Here, too, equal tokens are involved in the representation of different lengths, and unequal tokens are involved in the representation of equal lengths. It seems that the role played by "5" in this example, corresponds to the role of the lines of equal length in Sloman's example. Just as "5" doesn't denote a length, yet contributes to the representation of a length, couldn't we argue that the line doesn't represent a length, but contributes to its representation? It can be argued that both occurrences of "5" denote, and moreover, denote the same thing, namely the number 5. But it can also be argued that the lines of equal length in the pictures also denote - something equally abstract - namely "projected length", and moreover, the same projected length.

Procedural interpretation and Fregean semantics

Let us now turn to (b) and (b'), which concerns the possibility of interpreting a complex symbol as a procedure for identifying what is represented. For example, "the brother of the wife of Tom", can be
interpreted as a procedure for finding that person: find whatever is denoted by Tom, apply the the-wife-of function to that argument, and then apply the the-brother-of function to the result, finally producing the answer.

Now, what would prevent us from interpreting a picture of a glass on a table beside a sofa as a procedure for finding a particular glass by first finding whatever is represented by the subpicture of the sofa, then finding the table beside it (via the beside-of function represented in the picture), and then the glass on top of it? (Cp. "the glass on the table beside the sofa"). Nothing. The problem is rather that there is a plethora of possible procedural interpretations. What is to be picked out? - the glass, the table, the sofa, the glass-on-the-table, etc? How should it be picked out? - should the table be picked out via the sofa, or via the glass, etc?

On the other hand, neither are procedural interpretations of verbal representations always unique. If Jim is the brother of the wife of Tom, "the brother of the wife of Tom" can be used to pick out Jim - surely - but it can also be used in a number of alternative ways: to point out a complex property of Jim, to point out a certain relationship between Jim and Tom, to point out a particular instance of brotherhood, etc. When it is used to pick out Jim, we can still conceive different procedures based on the same symbol; for example, we might start with brothers, select those that have a sister who is married, and finally pick out the one whose sister is married to Tom.10 Not even mathematical notations are exempt

10) Not a remarkably efficient procedure, I admit, it would have worked better on an example such as "the three-legged son of a farmer".

- Aaron Sloman 99
from multiple procedural interpretations: "2+2" can be used to refer to 4, to point out that something has the property of being 2+2, to point out a certain property of +, etc. But, admittedly there are certain standard interpretations.

There are then, at least some weak parallels between verbal and pictorial representations, with respect to procedural interpretations. A case in point is the choice between referring and non-referring interpretations common in verbal representations. "The wearer of a wide-brimmed hat" in "Garbo is the wearer of a wide-brimmed hat", could be taken either as a procedure for identifying Garbo, or as an attribution of a certain property to Garbo; similarly a picture of Garbo wearing a hat could be used to pick her out - using the hat to identify her - or to point out that she's wearing this particular hat. But given more context, the verbal version is normally disambiguated, moreover we can choose to formulate a nonambiguous verbal description (either referential or attributive). In verbal representation, ambiguity is a matter of choice. In pictorial representation not.

(c) and (c') is about whether subsymbols represent parts of what the whole picture represents, or not. The parts spoken of in (c) are not just the syn-categorematic parts, but - what makes it interesting - also parts that represent all by themselves; like "Tom" in "the wife of Tom". What about pictures? Can a picture have a part that is meaningful by itself (represents something), and which contributes to the meaning of the whole, while at the same time what the picture as a whole represents does not include what that part represents? Well, it depends on how we interpret "represent".

11) I'm not thinking here of possible verbal elements in a picture.
Compare "the brother of the wife of Tom" with a picture of Churchill with cigar and hat in front of 10 Downing Street. Couldn't we argue that just as "Tom" (in a referential use of the expression) is part of the background that helps us to identify Jim, just so are the subpictures of the cigar and the hat and the well-known facade parts of a background that helps us to identify the person as Churchill? So the cigar, hat and the house are not, on this reading, represented in the picture as a whole.

Summing up the discussion of (b), (b'), (c), and (c'), it seems that the main differences are that pictures allow of a wider range of interpretations and uses, they are less definitely tied to any particular interpretation, and they cannot by far, as forcefully and distinctly, recommend a particular interpretation.\(^\text{12}\)

Structure and analogical

(d) and (d') makes a distinction between the role of structure in a Fregean symbol and an analogical symbol. It goes without saying, that the structure of a Fregean symbol normally contributes to the representation, and it's a commonplace that the structural, or formal, information of a complex Fregean symbol is an important part of its information content. It would seem odd to claim that this part of the information never could be about the structure of what is represented; for example, doesn't the structure of "2*3" contribute to inform us about the structure of 6? - neither of "2", "*", nor "3", can alone inform

\(^{12}\) However, Sloman's preference for a Fregean type of semantics for Fregean representations may have magnified the differences between verbal and pictorial representations in these respects. Perhaps, a situational semantics in the vein of Barwise & Perry (1983) would play down some of the differences.

- Aaron Sloman
us about the prime factorization of 6. Or, take any constructor function for a data structure. Perhaps we should understand (d) then as referring to some closer and more specific relationship between structures; the point could be that the structure of a Fregean symbol doesn't mirror (somehow) the structure of what it represents, contrary to pictorial symbols; however, I will defer the discussion of structure and similarity of structure to chapter IV.

In (e') it is claimed that analogical symbols are less likely to represent "impossible objects," it's easy to say "a round square" but not so easy to draw one. Let us call a representation which necessarily lacks denotation an impossible representation. An impossible picture is then - not a picture which cannot exist - but a picture which depicts something which cannot exist.

This claim is, although intuitively plausible, not very precise. One possible interpretation would be that given any system for analogical representation and any system for Fregean representation for the same domain, it would generally be the case that the proportion of representations within the analogical system that are impossible, would be smaller than the proportion of representations within the Fregean system that are impossible. But how do we count

13) For example btree(x,y,z) which takes a node x, and two binary trees y and z as arguments, and gives as result the binary tree with root x, left subtree y, and right subtree z; the structure of "btree(a,btree(b,∅,∅),btree(c,btree(d,∅,∅),∅))" tells us all about the structure of the represented binary tree.

14) Sloman (1978:171) writes "lacks a denotation", which he obviously doesn't mean; one very important purpose of knowledge representation is precisely to represent hypothetical - planned but not yet existent - objects, situations, sequences of actions, etc (many of which will never come to existence).
representations? There are some problems: i) how are different symbols within a representational system individuated? - the answer is probably critical for the outcome, yet there may be a wide range of possible and equally arbitrary answers; ii) what's in a system? - much has been said about this problem for natural languages, but there is as yet no general agreement about what is ungrammatical, and what is just meaningless; even less clear is how we are to distinguish between a picture that is impossible, and a non-picture?

I think that the basic intuition that pictures are less likely to be impossible is sound. But we need to make more precise what this means.

Contrary to Fregean symbols, it is claimed in \((f')\), analogical symbols have symbolic mappings which have the property that small changes in the representation are likely to correspond to small changes in what is represented. Sloman is careful to point out that - contrary to common presuppositions - nothing prevents analogical representations from being discrete (as opposed to continuous). These issues will be further discussed in chapter V.

In \((g')\) and \((h')\), finally, it is claimed that with analogical symbols it is easy to express constraints in the represented world, and the analogical representation is often the more economical to use. These points are closely related to \((e')\) and \((f')\). I think we can sum up \((e')\), \((f')\), \((g')\), and \((h')\) by saying that using analogical representations it is sometimes possible to bring it about that what is syntactically possible and efficient corresponds rather closely to what is semantically possible and efficient.

- Aaron Sloman
Patrick Hayes (1974) thinks that the distinction between "linguistic" (verbal) and "direct" (pictorial) representation is important, but surprisingly difficult to make precise.

One important step towards a clarification of these issues, Hayes points out, is the notion of different levels of representation. Only with respect to a particular level of representation can the verbal-pictorial distinction be applied, since a pictorial level of representation may be implemented by a verbal level of representation, or vice versa.

**Direct and strongly direct**

More precisely, there are, according to Hayes, media in which one can create different configurations of marks. A representational language is defined by a set of primitive symbols (marks) and a set of grammatical rules for creating new, complex, symbols (configurations). Models for this language are defined by giving meanings for the primitives and semantic rules defining the meaning of a complex symbol in terms of the meanings of its parts. The difference between verbal and pictorial representations, lies in that for a pictorial representation

"each medium-defined relation used in constructing configurations corresponds to a similar relation in the meanings, and that the representation is a structural homomorph of the reality with respect to these relations." (Hayes 1974)

Whereas Sloman's analogical/Fregean distinction is mainly based on the absence or presence, respectively, of compositionality, Hayes's distinction
rests on structural similarity, and he doesn't even allow for the possibility of non-compositional "direct", pictorial, representations. Behind the insistence on compositionality, one is led to suspect, lies a firm conviction that only representational systems with a precisely defined Tarskian type of model semantics, are really worth considering for applications in AI.

However, Hayes also defines a strongly direct representation in a way that would seem to exclude a Tarskian semantics - it is a representation where

"a medium-defined relation holds between subconfigurations if and only if the corresponding relation holds in the world between the entities denoted by the subconfigurations." (ibid.)

This raises issues of exhaustivness and plasticity. In a map - Hayes's example of a strongly direct representation - we can't have the representation of the distance between city A and city B without also having the representation of the distances of city A to the other cities represented on the map. In this respect, the map is exhaustive. Normally, Hayes remarks, maps are exhaustive even in a stronger sense: we expect that all cities in the area covered by the map are represented in the map; if a city is missing, the map is plainly wrong, not just incomplete. Let us call such a representation super-strongly direct.

Many maps are not super-strongly direct. A map may represent only cities with more than 100,000 inhabitants - or - should we say that it is super-strongly direct with respect, not to cities, but to cities-bigger-than-100,000? Doesn't this open up a possibility to interpret any map as super-strongly direct, with a suitable definition of what is really represented? - even an extremely non-uniform case

- Patrick Hayes
such as the typical tv-news commentator's map (with a few major cities as points of reference, plus one or two cities that are the current focus of interest)?

A major problem with Hayes's type of approach to the verbal-pictorial distinction is to give precision to the notion of structural similarity, and to be able to distinguish "the syntactic relations which are displayed directly by the symbol-configurations of the language". These matters will be discussed in chapter IV.

Adding to a representation

According to Hayes, strongly direct representations are less plastic than linguistic and weakly direct representations: information cannot be added piece-meal in the way typical for linguistic representation. Addition involves alteration.

Clearly, (some) additions to the map are necessarily made in "chunks"; we cannot add the distance between city A and city B without adding other distances simultaneously if other cities are part of the map; however, linguistic representations are not exempt from similar phenomena. Adding that 1+2=3 to a description of numbers cannot be done without simultaneously adding that 2+1=3, if the commutativity of addition is part of the representation. Actually, a closely related point about the interconnectedness of logic has been made repeatedly by Hayes (e.g. 1977, 1978). Perhaps, some would like to make a point about the difference between explicit and implicit information here, but I think that's a slightly different issue. We could choose to add that 1+2=3 by way of explicitly adding "1+(1+1)=3", assuming that we already have representations for the fact that 1+1=2 and for some substitution rule.
Also, it would seem that we can easily add a city to the map without altering the representations of the other cities and their distance-relations (and probably quite a few other things). Now, it must be admitted that the kind of globally determined relations rendered possible by a super-strongly direct representation, such as "the nearest neighbouring city of city A", are liable to change, but notice we don't have to actually do anything to bring it about: the "alterations" are made automatically, instantaneously, and for free, moreover they are guaranteed to be relevant and correct (under suitable assumptions).

If this is something linguistic representations can't parallel, it is a point in favor of pictorial representations. Recent discussions of non-monotonicity have highlighted that ordinary monotonic logic lacks this ability to change - once a theorem always a theorem, and no amount of axioms added can alter it. The attempts to develop a non-monotonic logic have so far not been very successful, it is also evident that model theories for non-monotonic logics will have to deviate strongly from the usual Tarskian semantics.

So far, I have tried to undermine Hayes's argument for a difference between verbal and pictorial additions, in effect reducing it to be valid for super-strongly direct representations. Yet, I think there is a difference between verbal and pictorial additions. If we were to add to the map the distance between city A and city B, we would - as we noted above - be forced to add other information simultaneously, for example other distances. Let us say that an i-chunk is a minimal chunk of information that


16) That is, no chunk included in this chunk will do.
includes the information i and is representable in a given representational system. Thus if \( i = \) the distance between city A and city B, an \( i \)-chunk will include other distance information as well. It appears that, generally in verbal representations, any \( i \) determines a unique \( i \)-chunk (relative to the total representation at hand). But in pictorial representations, some \( i \) are included in several \( i \)-chunks. In the example, \( i \) is clearly contained in several \( i \)-chunks, since \( i \) isn't enough to fix the position of city A, and fixed it must be to form an addable chunk.

Differently put, sometimes adding to a pictorial representation forces a choice between several possible ways of doing it, that will give different results in what is being represented. One consequence of this is that a pictorial \( i \)-chunk is always at least as big as the corresponding verbal \( i \)-chunk, and often it is bigger. A further implication is that a pictorial representation cannot grow as smoothly as a verbal, rather it grows in some sort of "quantum jumps".

Of course, this phenomenon is very much the same as the old problem of drawing a "triangle": one is forced to choose particular angles, sides, and orientation. The common interpretation is that pictures are not as abstract as words; one cannot depict an abstract triangle, only particular triangles. But the example above shows that it's not just a question of level of abstraction: adding a distance to the map doesn't force us to be less abstract, just to specify some other distances, that is, information at the same level of abstraction. Besides, the difficulty

17) Dretske makes a similar point in his distinction between information in digital form and information in analog form (1981:137).
of representing abstractions pictorially is not universal. There seems to be no big problem in drawing a general "circle".

Apparently, Hayes's point about additions raises a number of questions. Let me spell out a few more:

1) What does it really mean to make an "addition"? One plausible requirement is that the syntactic increment is a syntactically well-formed unit of a certain minimum level. We are probably not willing to let "blue" or "the father of John" pass as legitimate additions in a logically based representation, or allow "it is false that" to be "added" as a prefix to some existing sentence in the representation; sentences seem to be the right level for ordinary logic. Is there also a requirement that the old representation is syntactically left intact? If so, exactly what does it mean? - we have seen some of the difficulties in the discussion of Sloman's principle of substitutivity. And obviously, Hayes would argue that an addition must never take away something, so that no super-strongly direct representation can be added to, just altered; any "addition" will at least remove the previously represented exhaustiveness of the represented entities.

If we don't subscribe to Hayes's austere definition of pictorial (direct) representation, it would seem that the above requirement for syntactically well-formed units poses a problem. What are the syntactic units in the map? For example, what is the least unit that includes a city? It should include the symbol for a city, plus, and here comes the trouble, something that will fix its position relative to the map. But what kind of thing is that? Presumably one or several instances of a relation (or relations), but how are these instances to exist without a "substrate"? For less regimented forms of pictures, the
problem seems formidable, or rather, one suspects that there is a wide range of possible "solutions", all equally arbitrary, and equally unsatisfying.

2) There is a tendency to equate information and representation. But it is obvious that not everything that a symbol gives us information about is represented; a hand-written message "Closed for lunch" may reveal much about the author (especially to a graphologist), something about grammar, something about what type of pen was used, and so on. Or, using one of the examples above, adding "1+2=3" may involve adding the information that 2+1=3, but certainly the addition doesn't represent such a thing.

Representation, not information, ought to be the central concept of knowledge representation. And yet information appears to be the more convenient concept when speaking of additions. For example, defining the i-chunks in terms of representation rather than in terms of what information they contribute, makes the problem of syntactically well-formed units for pictures, acute: we can't have something represented without something representing it, meaning we must be able to say what it is.

The efficiency of pictures

Hayes takes the possibility of mixing levels of different modes of representation as an argument against Sloman's idea that pictorial representations are occasionally more efficient. It is an error to speak of a certain analogical representation being more efficient than a Fregean, since the analogical representation "may be embedded in a medium which is itself represented in a Fregean way in some other medium", and "any discussion of efficiency must take into account the computational properties of the
medium" (Hayes 1974). The argument is slightly confusing. Is the point here that an analogical representation can be implemented as a Fregean, thus in a sense being a Fregean representation after all? Probably not, but then why emphasize that it is an analogical representation implemented as a Fregean, if the point simply is that the efficiency of a certain level of representation depends on the efficiency of its implementation, analogical, Fregean, or whatever?

I think that Hayes's attack on Sloman is unwarranted. Obviously, it couldn't be that efficiency only depends on the efficiency of the next lower level, because that would mean that the efficiency of a program is just a matter of the efficiency of the machine hardware, which is patently false. Obviously, the structural organization has very much to do with efficiency, which is one of the most salient insights of computer science. Then, why is the choice between pictorial and verbal not a legitimate and important question of efficient structural organization?

On the other hand, of course Hayes is right in that the computational properties of the medium have important consequences for the efficiency of a representation. Generally, to what degree does a certain mode of representation "shine through" to higher levels of representation? There seems to be no obvious answer to that question. We might also put the question in the negative: what are the possibilities of hiding lower level features? Some attempts to discuss these

---

18) Besides, if Hayes seriously believes that pictorial representations can give no important efficiency advantages, then why bother with non-linguistic representations at all? Perhaps, the point Hayes really wishes to make is that there are only linguistic representations.

- Patrick Hayes
questions have been made in cognitive psychology, and I will return to them in chapter III.
Mental imagery is a hotly debated topic in cognitive psychology. The question is whether mental images are a reality in any more concrete sense than as a subjective experience. Several experimental results seem to indicate that people have mental pictures that can be rotated and scanned at measurable speeds, can be enlarged and reduced (smaller images being more difficult to "see" than larger ones), etc; all "viewed" by an inner "mind's eye" that has a determinate "angle of vision" (e.g. Shepard & Metzler 1971, Kosslyn 1978).

On one side are the pictorialists claiming with support from these experiments that mental images derive from pictorial representations. On the other side are descriptionalists maintaining that, in spite of appearances, imagery derives from verbal representation, something like natural language, or predicate logic.

I have no intention to take sides in this disagreement; I don't think it is directly relevant to AI whether people do or do not have pictures in their heads. What is relevant though is what difference it would make, and, for that matter, what the difference between verbal and pictorial representations really amounts to; which are two of the most important issues in the debate on imagery.

Pictorialist views on images range from the view of an image as some sort of "copy" of a scene, emphasizing the continuity and density of images (e.g. Paivio 1971, Kosslyn & Pomerantz 1977) - to the view of an image as an analog, with a structure-preserving
mapping, or some sort of "isomorphism", between object and representation (e.g. Shepard 1975).\footnote{Zenon Pylyshyn's critique (particularly Pylyshyn 1973), has had great influence in urging the pictorialists to make an initially vague picture metaphor more precise, defining more clearly what pictorial representation amounts to. (Whether they have succeeded is a moot question.) On the whole, I have found the most interesting reflections and observations (not to say that they are the most correct) on pictorial representation among thinkers not devoted to pictorialism, which explains the descriptionalist bias of this section.}

Daniel C. Dennett

Dennett (1969) is a straightforward denouncement of pictorialism. Dennett (1978) modifies this view to a kind of "agnosticism", but above all clarifies the ontological question. There are two main approaches to mental imagery: i) the phenomenological approach; and ii) the scientific approach.

The phenomenological approach is to investigate the intentional objects that can be reconstructed from beliefs about images. (Clearly not our concern here). The scientific approach is to investigate the immediate causes of beliefs about having mental images and to find out if these real entities are like pictures or like words (and Dennett still thinks the evidence favors the descriptionalist view). The case between pictorialists and descriptionalists is thus empirically decidable, and not as many have thought a philosophical question.

\footnote{1) Kolers & Smythe 1979, give a survey.}
Dennett (1969) proposes as an "acid test for images" that they must resemble what they represent; "something hard, square and black just cannot be an image of something soft, round and white" (1969:133). Thus, Dennett — along with other contributors to the imagery debate — assumes a traditional copy or resemblance view of pictures, on which his attack on pictorialism is based. Later on, in chapter IV, we shall see how this view of depiction has suffered severe adversities — but let us here examine some observations Dennett makes about pictures.

Consider a picture of a tall man with a wooden leg. Whereas a verbal representation can leave it at that — "a tall man with a wooden leg" — according to Dennett, the picture must either represent the man as wearing a hat, or as not wearing a hat, or else there must be "something positive ... drawn in where the hat should be, obscuring that area" (1969:135).

What exactly is the point Dennett is trying to make? One possible interpretation is that pictures, contrary to descriptions, doesn't permit us to select freely which details of a larger whole (that is represented) to represent explicitly and what to leave out (and what to be explicitly vague about, and what to explicitly leave out). More precisely, pictures have a uniform level of detail (in that which is being represented), a rather obvious consequence of a copy theory of pictures. If we have decided to represent the man as having a wooden leg, we can not possibly escape from representing the upper part of his head with sufficient detail to commit ourselves

What if we took a close-up photograph of that soft, round and white thing, using an old-fashioned camera with glass-plates, wouldn't we get an image (the negative plate) that actually would be "hard, square and black"?

- Daniel C. Dennett
to the question of headgear (or obscuring object).

Well, how about figure III:1?

![Figure III:1](image)

One might perhaps try to dismiss this sort of picture as being fraught with non-pictorial (verbal) elements. Anyway, it isn't exactly self-evident which features are at the same "level of detail". Compare the discussion above of the exhaustiveness of "super-strongly direct" representations.

The Tiger's Stripes

An alternative, or perhaps complementary, way of interpreting Dennett's point, would be simply that pictures do necessarily go into details, they are necessarily specific. A picture of a tiger - another of Dennett's examples - must reveal a definite number of stripes. A description need not even if it does mention the stripes, for example, "numerous stripes".

Why is that so? On the copy view of depiction, it is reasonable that a copy of something must copy at least its gross details. From a general point of view, uncommitted to any particular theory of pictures, the example suggests that in a pictorial representation, representing a complex whole is possible only by representing (some of) its parts. (Cp the discussion of (c') in chapter II: Sloman.)
A picture of a man becomes a picture of a man only because it is composed of a certain arrangement of subpictures: head, arms, etc. It is these things taken together that makes it a picture of a man. We couldn't have a man-picture without man-features.

Not so with words. We simply say "a man", and that's that - no arms, etc.

Compare this example with the abstract triangle problem. We can pictorially represent a triangle only by pictorially representing each of its sides and how they are related to each other (angles, etc). Yet, it is probably an error to generally equate "specific" with "going into details" - "Winston Churchill" is a very specific verbal representations which doesn't go into details at all. Maybe we should say that a picture manages to be specific only by going into details?

Then again, one may contend the factual issue. As Fodor (1975) points out: it's not true that a picture of a striped tiger has to be determinate about the number of stripes. The main cause of indeterminacy, Fodor thinks, is blurring. A photograph of a tiger can show stripes but at the same time be slightly out of focus so that individual stripes can not be counted. And there are other problems, as for example effects of perspective: lampposts vanishing into the horizon - how many are there?

Under some modes of depiction we're not specific about the number of stripes, as indeed for some modes of pictorial representation (e.g. a 2-dimensional projection of a triangle in a 3-dimensional room) we're not specific about a triangle's angles. But it would seem that for any given mode of pictorial

3) Goodman wouldn't agree with that, as we shall see further on.

- Daniel C. Dennett
representation we do not have quite the freedom of (some) verbal representations to choose between being specific or general, and referring to certain details or not.

Jerry A. Fodor

Fodor (1975) thinks that, although pictures cannot be a general vehicle of thought and reasoning, they do probably play a role - empirical studies will have to settle this question.

Vehicles of truth

Fodor's first main point is that pictures are insufficiently abstract to be vehicles of truth. Any picture of a thing will display many properties; which of these are truly significant, that is, belong to the truth conditions of the picture? There is no way to tell. For example, "John is fat" can't be replaced by a picture of John with a bulging tummy, because it might mean that John is pregnant, etc. "For every reason we might have for calling a picture true, there will be a corresponding reason for calling it false." (1975:181) And to this comes the problem of distinguishing a simple reference from a statement-like representation (a vehicle of truth); how do we know that the picture doesn't simply represent John?

I think there are two different kinds of ambiguity involved here: i) what is significant?; and ii) what does it signify? Assuming that the John with a bulging tummy picture is a vehicle of truth, there is first the problem of identifying which, of all the different parts and features in the picture correspond to the truth conditions. If the picture is to correspond to "John is fat", it's something
like the bulging tummy feature, the John-looking features, and their arrangement so that the fatness is pictured as belonging to John.

This ambiguity in what is significant, goes even deeper than example (5) in the introduction indicates (does a wobble in the picture signify a wobble in the object, or just the artist's trembling hand?). Because, it may very well be that John looks middle-aged in the picture and that John is in fact middle-aged, and yet this middle-aged feature of the picture should be no part of the truth-condition, because all the picture should be committed to is that John is fat.  

The second problem is that, given which parts and features of the picture are significant for it's truth, several interpretations of these are possible. What difference could there be between pictures for "John is fat" and "John is pregnant"? One might argue that a very detailed picture would reveal a difference between fatness and pregnancy, similarly for Fodor's second example (borrowed from Wittgenstein) - a picture of a man walking up a hill forward, or, is he sliding down backwards? - possibly a good picture would reveal some difference in the tension of various muscles. This sort of argument isn't very convincing however, and after all it's not a big surprise that pictures are at a disadvantage at differentiating between two things that look the same.

Note that only the first type of ambiguity, can be blamed on pictures being insufficiently abstract. The second type of ambiguity, results from pictures

4) Verbal representations are not completely devoid of similar problems. For example, using a floating-point notation for some measurements how many digits are really significant, that is, are involved in the truth condition?
being too abstract. Fodor will have it that the John with a bulging tummy picture fails to differentiate between "John is fat" and "John is pregnant" because it isn't abstract enough: it hasn't (as "John is fat" has) abstracted from "the fact that fat men do look the way that pregnant men would look" (1975:181).

But, presumably, the picture has abstracted from any property of John that doesn't somehow show in his appearance. Assuming then that all properties that could have served to distinguish pregnancy from fatness has been removed in this process of abstraction, the result is a picture that is too abstract to distinguish "John is fat" from "John is pregnant". Were we to proceed from this point, abstracting also from visual properties (consequently from "looking-alikeness") there would be nothing left.

As we move to a broader definition of pictorial, and bring in the depiction of things and properties that cannot normally be seen, and non-pictorial qualities, do we not then have a free hand to create, for any two sentences (attributes, etc) a pictorial representation that differentiates between them? Maybe so, yet I think that in any given (reasonably rich) pictorial symbol system there will be ambiguities of this second type, so that within the system, differentiability has a limit. The reason is that a pictorial symbol system seems to abstract from certain properties uniformly. Bronze sculptures abstract from colors throughout. Photographs always abstract from smells (among other things); there cannot be two photographs that distinguish between "John smells of garlic" and "John smells of whisky". In a given system of pictorial representation, there is thus a basic level of abstraction which no picture in the system escapes, and consequently we are not as free to choose abstractions, or mix different modes or levels of abstraction, as in the most powerful
verbal systems for representation (such as natural language).

But isn't there still a possibility of interpreting pictures as vehicles of truth? Not vehicles for the simple truths of ordinary sentences, but - from the verbal point of view - very complex truths corresponding to complex sentences such as "John is fat or John is pregnant or ...". Certainly, this wouldn't satisfy Fodor: his discussion is about whether pictures could replace words. Clearly, they can't. Yet, the idea of interpreting a picture as somehow equivalent to a very complex description or proposition has appealed to many. "One picture says more than a thousand words" as the old Chinese saying has it. Only that a picture doesn't exactly say anything, and that "more than a thousand" is an understatement. Just because a picture doesn't say anything (it isn't verbal) someone has to speak for it, of it, and generally there seems to be no end to what can be said. Would the "truth-conditions" of a picture be finitely specifiable in words?^5

Vehicles of reference

Fodor's second main point is about reference: as vehicles of reference, pictures are not fundamentally different from verbal representations; the means by which images refer are essentially the same as the means by which verbal symbols refer to what they denote. This thesis is based on two premises: i) not every instance of a verbal sign that can be used to refer does in fact refer; and ii) resemblance is not a sufficient condition for pictorial reference. The conclusion seems a bit too strong, coming from these seemingly innocuous premises, but let us examine the

^5) And is there any single language where all that can be said can be said?
Not every utterance of "John" does refer to John; it doesn't refer if one just happens to produce that sound with nothing in mind (says Fodor); "...no utterance of 'John' counts as a reference to John unless it was at least produced with the intention of making a reference." (1975:182). Now, compare this with pictorial representations. Forming a mental image of figure III:2 (Fodor's example),

![Figure III:2](image)

then, "if thinking is forming an image of a thing, and if images refer to whatever they resemble, then you must have been thinking of a cube viewed from one of its corners." (ibid.). But surely most of us wouldn't have thought of a cube without a context that prepared us for this interpretation. We may conclude that pictures as well as words must be taken in certain ways in order to refer. But from this there's still quite a step to the claim that there is no important difference in the ways pictures and words are used to refer.

What is more interesting, is that Fodor seems to have missed that his first point about pictures not being vehicles of truth really has little to do with truth and truth-vehicles (statementlike representations). I wonder if the traditional division between sentences and attributes has not lead him astray. There is actually the same problem with attributes, names, descriptions, etc: ambiguity in what is significant and ambiguity in what it signifies. And the puzzling
cube picture is an exact parallel to the problem with the John is fat picture.  

Truth, to put it drastically, doesn't really interest AI. There is in (applied) AI generally no sharp division between propositionlike representations and (for example) attributelike representations. It seems that the concept of truth simply does not emerge naturally from AI's central task of creating models and representational systems. Instead of the question whether a representation is "true" or not one finds that the central question is (unprejudiced by linguistics and analytic philosophy): does it apply or not? This notion of applicability removes some venerable partitions inherited from the study of language, classifying representations into names, attributes, descriptions, statements, etc, so it is in a sense a wider, less precise concept than truth. But it is in other ways more refined. A representation can apply in varying degree, and in various ways. How a representation applies is a very important question, that doesn't arise naturally from the concept of truth. More about this in chapter VI.

Resemblance

As we saw, Fodor objects to Dennett's contention that a tiger-picture must be determinate about the number of stripes. More precisely: "It simply isn't true that a picture of a striped tiger must be determinate under such descriptions as 'has n stripes'" (1975:189). This passage reveals that Fodor assumes basically the same copy or resemblance view of pictorial representation as Dennett. From someone less committed to such a view one would have expected:

6) So this should have been his second point, and then it would have been really just one point: that pictures are haunted by this kind of ambiguity.
'represents n stripes', or at least, 'has n stripe-representing subpictures'. The argument is after all about what the picture represents. As it happens, there's nothing odd in representing stripes by stripes; a different example, lampposts, say, would have disclosed the hidden ambiguity: obviously there are no ordinary lampposts in the picture. And if we do not subscribe to a resemblance or copy view on pictures, it's far from evident that stripes in the tiger must be represented by stripes in the picture, or even that 3 stripes in the picture must represent 3 stripes in the tiger, etc.

Fodor interprets Dennett as claiming that a picture has to be determinate under any visual description, a thesis which Fodor has shown to be false by these examples. But he thinks that a much weaker thesis holds: a picture must be determinate under some visual description. For example, a newspaper picture is always determinate under a gray-matrix description. But if we accept that pictures are symbols and not just objects (something adherents of a copy theory of pictures tend to forget) this is surely trivial; any visual symbol, be it pictorial, verbal, or whatever, must be determinate under some visual description, namely - at the minimum - a description that identifies it as being that particular symbol.

Generally it seems that the copy approach to pictorial representation invites confusion between the syntactic properties of a picture (the picture as an object) and its semantic properties.

Although Fodor sets out from a traditional copy view on depiction, at some places he is clearly heading

7) "By stipulation, a picture is determinate under a description iff the statement that the picture satisfies the description has a determinate truth value." (1975:189n)
away from this view: in the discussion of pictures as vehicles of reference, where he points out that resemblance isn't sufficient for reference (although he lets the reader assume that it is necessary), and in a concluding discussion of what Fodor calls images under descriptions. If a picture is coupled with a description, we get something which has the nondiscursiveness of a picture, yet shares with a description that it needn't look much like what it represents; "they convey some information discursively and some information pictorially" (1975:190). Maps with legends is one type of example. Fodor envisages a psychological faculty to construct mental images from descriptions; these descriptions "can function to determine what the images are images of, and how their properties are to be interpreted." (1975: ). One consequence, Fodor says, is that the idiosyncratic character of images may be irrelevant to their cognitive role.

"Suppose your image of a triangle is scalene and mine is isosceles. This needn't matter to how we use the images to reason about triangles so long as we agree on how the images are to be taken; e.g. so long as we agree that they are to represent any closed three-sided figure whose sides are straight lines." (1975:192)

I don't agree. Of course, it matters. Not every property of the chosen representing triangle is shared by all triangles. That it happens to be isosceles doesn't mean that all triangles are. Examples of less obvious misuse are those spurious geometrical proofs that look convincing by a cunning choice of illustration. And certainly, this is not how Fodor intends how the representing triangle should be used

8) Which means that "the information it conveys is displayed rather than described" (1975:190)?
to reason about triangles in general. But what does he mean? In what sense does an isosceles triangle represent all triangles other than in the sense that we have stipulated it to so represent? If it's only an arbitrary convention - then in what sense is this different from representing the triangles by a word ("triangle")? Fodor fails to bring light on this important - I would even say, decisive - point.

Zenon W. Pylyshyn

Pylyshyn (1973, 1975, 1978, 1980, 1981a, 1981b), known as a descriptionalist, thinks that the most important issue in the imagery debate is whether certain aspects (among them the experimental results referred to above) of imagery ought to be viewed as intrinsic properties of a medium or mechanism, or whether they should be understood in terms of processes operating on symbolic encodings. This distinction between medium and symbolic operations is not simply an analogue to the hardware - software distinction in computer science (although it seems to have been precisely that, originally). Rather than an analogue to hardware, the medium is to be understood as a virtual machine.

Thus, the issue is whether the pictorial nature of mental images should be attributed to properties of the architecture of the human virtual machine - or - to processes operating on that machine which are in principle alterable by knowledge: cognitively penetrable. Based on this analysis, Pylyshyn has given support to the descriptionalist thesis, by showing that many of the more spectacular properties of mental imagery, such as "mental rotation" and "mental scanning", are in fact cognitively penetrable.
Whereas the distinction between virtual machine and software operating on it, is a pragmatic one in computer science—normally we can discern several virtual machines nested in each other in a single computer system—in psychology, the distinction is, according to Pylyshyn, principled: there is a "unique human virtual machine—defined in terms of the lowest level of function that must be explained by appeal to its having representations (goals and beliefs)." (1981a:74).

What has he then said about pictures? Explicitly, not very much. But if something is to count as a picture in any cognitively interesting sense, it must in Pylyshyn's view be implemented in the cognitive virtual machine (not in the software operating on it) as an analogue representation; that is, in an analogue medium, which is a medium that

"...incorporates a whole system of lawfully connected properties or intrinsic constraints ... and that it is precisely this set of properties and relations that determines how objects represented in that medium will behave." (1981b:20)

Also, it must be operated by analogue processes, which are processes

"...whose behavior must be characterized in terms of intrinsic lawful relations among properties of a particular physical instantiation of a process, rather than in terms of rules and representations (or algorithms)." (ibid:19)

Transferring the result of applying the virtual machine metaphor to psychology, back to computer science (or rather, artificial intelligence), we do not get back exactly to where we started from. Pylyshyn

- Zenon W. Pylyshyn
seems inclined to equate "computational" with "cognitive"; his virtual machine has a physical/biological, non-computational, flavor. Yet, the very point with the (original) virtual machine concept, is that a virtual machine can be computationally implemented, and I see no compelling reason to assume that sub-cognitive processes are non-computational or that being cognitively impenetrable implies being physical. And certainly, the "cognitive virtual machine" of an AI-system must be designed as a computational system (whether instantiated in hardware or software, is another matter); the single most fundamental assumption of computer science, is after all that problems can be solved by computations (and if they can't, they are not problems of computer science).

And, if pictures and pictorial operations can be implemented by data structures and computations in the virtual machine, surely they can be so implemented at some higher level of representation.

Although Pylyshyn has taken the virtual machine concept into the service of the computational view of cognition, he apparently doesn't apply the type of separation of what and how which this concept exemplifies, to questions of representation such as the issue of pictorial versus verbal representation. It seems that the how of a picture is in Pylyshyn's view invariably non-computational. As far as his notion of an analogue representation is concerned, it is

9) Incidently, I think it might be convenient to supplement Dennetts three fundamental "stances" (Dennett 1978) - the physical stance, the design stance, and the intentional stance - with a fourth: the computational stance. It would fall somewhere between the design stance and the intentional stance. The computational stance is beginning to have some success also outside the traditional application areas (computer science, artificial intelligence, cognitive psychology); in the theory of human vision (Marr et al.), and in microbiology.
quite similar to Palmer's concept of an intrinsic representation, which will be discussed below.

Stephen E. Palmer

Palmer (1975, 1978) believes, rightly I think, that to make systematic progress on problems of cognitive representation, we must start with the most basic questions about representation in general. He ends up declaring that pictorial representations and verbal representations are fundamentally opposed but—this is a little surprising—no, not in ways relevant to cognitive psychology. (His own analysis does however give some support to the opposite view.)

Palmer (1978) defines a representational system as two related but separate worlds: the representing world and the represented world.

![Diagram of representational system](image)

Figure 111:3. "Examples of representation. The represented world in each case consists of the objects shown in A. For each representing world, the correspondence of objects is indicated by the letters beneath them, and the correspondence of relations is indicated by the expression beneath each world." (Palmer 1978)

- Stephen E. Palmer
A complete specification must state: (1) what the represented world is; (2) what the representing world is; (3) what aspects of the representing world are being modeled; (4) what aspects of the representing world are doing the modeling; and (5) what are the correspondences between the two worlds. Figure III:3 gives some simple illustrations (reproduced from Palmer 1978).

Closure

Palmer is emphatic about relations in a representations being operationally defined - not to be confused with any apparent relation that may be present.

"... the only information contained in a representation is that for which operations are defined ... knowing it is theoretically there is not sufficient for it to be considered part of the represented information ... the implicit ... information is not there at all except for the computer programmed to extract it, and for that computer, it is there even though it does not seem to be." (1978:265f)

I agree that representations should be operationally defined, but what Palmer writes about the operations of a representation needs some comment. I will assume that "information" should here be read as "what is represented"10 - clearly Palmer's phrase "represented information" must be a slip, since normally we don't expect a representation to represent information, rather it conveys or contains information by representing a state of affairs, etc. Now, the quoted passages could possibly be read as also asserting closure of what is represented.

10) Cp the discussion in chapter II: Patrick Hayes.
But then we have fallen into a trap; something we might call the closed knowledge fallacy. The philosopher's model of knowledge or representation, is typically closed under logical inference: any thing which may be inferred from the representation is represented. This kind of model, disregarding time and space and concrete implementation, has for obvious reasons not won approval in AI. One possible objective for making what looks like a distinction between an eternal omniscient observer's point of view (apparent relations) and the system's own point of view (operational relations), is to preclude precisely this type of interpretation of what is really represented. But, if whatever there's a program to extract is represented (without further qualifications) and if extraction operations can be combined, then what is represented will cover everything the system can, in principle, ever get out of the representation; "computable from" is a transitive relation just as "deducible from", and there is no differentiation between information that can be found instantly, and information that takes years to produce. The one is as much there as the other.

A second point about closure concerns, more specifically, pictorial or analog (Palmer's term) representation. Closure, it seems to me, is not compatible with one of the most important roles of analogies; namely to serve as heuristics. Very often we use analogies or models in the hope that there are more similarities present than we know of when the analogy is adopted; a property of the analog model is observed or discovered, translated back to the original domain to see if the corresponding property holds there. Sometimes it does, and the analogy has filled a heuristic purpose. In such a case we don't know in

11) For example, if it is known that A, and it is known that A implies B, then it is known that B.

- Stephen E. Palmer
advance what exactly is significant, that is, what exactly is represented, in the representation and it's precisely for this reason we're interested in it. What analogies, and I venture to say pictures in general, represent is often open-ended. More on this in chapter VI.

Intrinsic and analog

Palmer's analysis of the pictorial/verbal issue rests on a distinction between intrinsic and extrinsic representation. Representation is intrinsic whenever a representing relation has the same inherent constraints as its represented relation; it is extrinsic whenever the inherent structure of a representing relation is less constrained than the relation it represents. The representation of "taller than" by "larger than" in world F in figure III:3 is intrinsic because "larger than" is inherently asymmetric and transitive, just as "taller than". But the representation in world H of "taller than" by "is arrow-path-connected to" ("chains to") is not purely intrinsic; transitivity is inherent in "is arrow-path-connected to" but asymmetry is not; it's possible that there are paths between x and y in both directions, there's nothing inherent in the "is arrow-path-connected to" relation to prevent this. Thus asymmetry has to be externally imposed, meaning that "taller than" is extrinsically represented, if it is represented at all.

Now, the extrinsic/intrinsic distinction is essentially, according to Palmer, the verbal/pictorial distinction, or, as Palmer prefers to say, the distinction between propositional and analog representations.

"Propositional representations are simply those in which there exist relational elements that
model relations by virtue of themselves being related to object elements." (1978:294)

For example, "is under" is a relational element in "the ball is under the table"; "the ball" and "the table" are object elements. Propositional representations are extrinsic because object elements can, in principle, be arbitrarily connected by relational elements; in the example the physical relation 'is under' would then be represented by

"...whatever structure there is in a propositional representation exists solely by virtue of the extrinsic constraints placed on it by the truth-preserving informational correspondence with the represented world." (1978:296)

Analog (pictorial) representations are simply non-propositional representations, that is, representations that lack relational elements. It follows that properties must be represented by properties, and relations by relations, and that

"...whatever structure is present in an analog representation exists by virtue of the inherent constraints within the representing world itself, without reference to the represented world." (1978:297)

- that is, analog representations are intrinsic. Or does it really follow? I think not. Consider the example in figure III:4. (The represented world is still A in figure III:3.) It is an analog representation by Palmer's definition, because there are no relational elements, yet it's not intrinsic: the representing relation is neither asymmetric nor transitive.

Although it is rather obvious that intrinsicness has something to do with picturehood, I think Palmer is

- Stephen E. Palmer
on the wrong track trying to identify propositional with the presence of relational elements and analogical with their absence.

Virtually intrinsic

There's one snag with Palmer's definitions of extrinsic and intrinsic: what exactly is inherent? Palmer thinks that this problem is closely related to the philosophical debate about analytic and synthetic statements; intrinsic corresponding to analytic, and extrinsic to synthetic. This is a little discouraging since there is no received view on the analytic/synthetic issue. Actually, I think these definitions are in greater trouble than Palmer realizes. For, whereas it might be intuitively clear, given a representing world, if relationships are necessary or contingent, it may be less clear what the representing world is. Consider, for instance, world H in figure III:3. If we by definition exclude every configuration that contains some cycle (of arrows) from this world, we get a world in which asymmetry is inherent: if a is arrow-connected to b then necessarily b isn't arrow-connected to a. Moreover, this expurgated world will inevitably result if we enforce H to rightly represent A. And generally, at the level where we have enforced the structural restrictions of an "extrinsic" representation, we
have in effect an intrinsic representation.

I propose to modify the definition of intrinsic, transforming Palmer's absolute and non-operationally defined concept into a relative and operationally defined concept. I will say that a property or relation is (virtually) intrinsically represented (in a particular virtual representation) precisely if it is represented, and the processes using the representation need not be concerned with the maintenance of this property or relation.

For example, if R' represents R intrinsically, and R is transitive and asymmetric, then if a'R'b' and b'R'c', the representation is guaranteed to have also a and c represented as standing in the relation R to each other, in other words, that a'R'c' is part of the representation; there is no need for the processes using the representation to make sure that this is the case. And if changes are made to some instances of R', any additional change necessary to preserve transitivity and asymmetry occur without any explicit action of the processes using the representation. In short: no violation of the structural constraints can ever occur. The means by which this is achieved is of no concern to the processes using the representation. Possibly there is a complicated machinery behind the neat facade (demons, dependency-directed backtracking, etc.).

Note that this is not just a question of dividing properties and relations (or instances of properties and relations) into explicit and implicit. Structure generally means mutual dependencies, and preserving it necessitates mutual adjustments. Also, although not discussed by Palmer, dependencies across different properties and relations are certainly not uncommon; for instance, with an "intrinsic" representation of positions we would naturally understand a

- Stephen E. Palmer
representation that were guaranteed to maintain also appropriate distance relations. Intrinsicness, as far as a particular represented property or relation is not completely independent of other represented properties and relations, is a matter of degree.

It seems to me that this modified definition is more on a par with Palmer's demand for a rigorously operational view of representation, than his own definition is. The relativization of the concept of intrinsic parallels the transition from the absolute view of representations ("apparent relations") to the operational view, both being examples of a necessary shift of perspective, from a traditional philosophical point of view to a computational point of view.

I doubt, however, that Palmer would be very happy with the modified definition; he apparently takes representations in a computer, to be the paradigm of extrinsic representations. But what I have done is simply to relativize intrinsicness to a certain level of virtual representation, splitting the question in two: functionality - does it function as an intrinsic representation? - and implementation - how is intrinsicness achieved? A type of analysis routinely made in computer science.

As I mentioned at the beginning, Palmer doesn't believe that his analysis has any implications whatsoever for cognitive psychology other than to assist in the demarcation of the field. The reason is that he thinks: i) that the inherent nature of the internal representation concerns the physical medium that carries the information; and ii) that there's no way to find out about this without looking inside people's heads (thus a question for neurophysiology or physiological psychology).

I hope to have undermined his first point, which means that this discussion, and in particular the
intrinsicness concept, is relevant also to AI, severe restrictions on available physical mediums notwithstanding.

Palmer's second point also needs some comment. More generally, he believes that cognitive performance is indifferent to variations within the class of informationally equivalent representations, that is, representations that convey the same information about the represented world (Palmer's definition). Palmer believes that

"...the proper level of discourse for cognitive theory [is] ... the level of abstraction defined by informationally equivalent systems" (1978:277)

So all that matters is the content, not how it happens to be organized. 12

This is to be sure not an uncommon view of symbols and representations in general - but it is another telling example of how a traditional philosophical point of view clashes with computational considerations. Organization matters in questions of computation. Computer science has always taken as one of its main tasks to find efficient ways of organizing information, and it is well-known that different ways of arranging the same information may lead to drastically - qualitatively - different access times.

For Palmer's more specific claim, that psychology will never be able to disclose the basic nature of the internal representation, see the following section.

12) Apparently, Palmer does not wish to consider even time constraints.

- Stephen E. Palmer 137
John R. Anderson (1978, 1979), like Palmer, believes that "barring decisive physiological data, it will not be possible to establish whether an internal representation is pictorial or propositional." (1978:249). His general argument is based on a theorem that says that, under suitable assumptions, one system can mimic the behavior of another system, even if they use different types of internal representation.

Anderson's general framework for describing a cognitive system includes: representations $R$, encoding processes $E$, processes of internal transformation $T$, and decoding processes $D$. Formally, $R$ specifies a set of possible representations $\{I_1, I_2, I_3, \ldots\}$; $E$ maps external stimuli $S_i$ into internal representations $I_j$; $T$ is an operation on $R$, mapping internal representations on other internal representations; $D$ specifies how internal representations are manifested as responses. $E$, $T$, and $D$, should of course all be computable. A particular pair $\langle R, E \rangle$ Anderson calls a theory of representation.  

**The mimicry theorem**

Suppose now, that $M$ is a cognitive system with a certain encoding mechanism $\langle R, E \rangle$, and $\langle R^*, E^* \rangle$ is another encoding mechanism which divides the stimuli into the same categories as $\langle R, E \rangle$ (the only difference being that a different set of symbolic objects are used for the code), and we have a two-way converter $C$ (corresponding to a computable function $f^*$ and its

---

13) This framework is crude and constrained, and the talk about "encoding" and "decoding" is rather misleading. But for the sake of argument I will discuss in its terms.
inverse f) which converts between \(<R^*,E^*>\)-code and \(<R,E>\)-code - then we may take M, fit it with the \(<R^*,E^*>\) encoding apparatus plus the converter in front, and it will continue to work as before, only slowed down by the extra encoding and converting operations. Note that T operating on R can now equivalently be viewed as some "T*" (= f T f*) operating on R*, and similarly for D. (Call this embellished system M*.)

These rather trivial facts seems to be the basis of the mimicry theorem, which in full reads as follows:

Given a system \(M = <R,E,T,D>\) and a "theory of representation" \(<R^*,E^*>\) that preserves the internal distinctions of \(<R,E>\), there exists a system \(M^* = <R^*,E^*,T^*,D^*>\) which: i) preserves the input-output behavior of M, that is, maps stimuli into responses the same way; and which ii) mimics the internal transformations, that is, \(T^*\) and \(D^*\) are isomorphic to \(T\) and \(D\); finally iii) "since the mimicking model goes through the exact same steps, not only will it reproduce the same behavior, it will reproduce the same time relationships" (1978:266, emphasis added).

From the preceding i) and ii) are seen to be fairly obvious and trivial. To be sure, Anderson wants us to think of this as an existence proof, intimating that in most cases \(T^*\) can be computed in a simpler, more direct, way than as \(f^* T f\). But there's no proof of that. As far as I can see the theorem doesn't prove anything about representations, and it tells us nothing new about pictorial and verbal

14) One "theory of representation" \(<R^*,E^*>\) is said to preserve the internal distinctions of another "theory" \(<R,E>\) iff: i) there is a computable 1-1-mapping \(f\) from \(R\) to \(R^*\); which ii) has a computable inverse \(f^*\); and so that iii) \(E = f^* E^* \).

15) Proof: \(T^* = f T f^*, D^* = D f^*, \) where \(f\) is the distinction-preserving mapping.

- John R. Anderson 139
Apparently Anderson's so-called "theory of representation" is intended to correspond to the ordinary notion of representational system, a system of interpreted symbols. For example, it is these "theories of representation" that Anderson ascribes attributes such as "pictorial" and "verbal". Yet, the Ij appear to be just objects. As long as we don't specify which features of these objects are significant and what the elementary methods for accessing and operating on them are (as, I take it, T and D do specify), they are "representations" in an almost empty sense. In particular, it is hardly meaningful to say that one pair <R,E> is pictorial, and another <R*,E*> is verbal. Despite Anderson's talk of the importance of studying representation-process pairs, instead of "theories of representations" in isolation, he doesn't keep a clear distinction between (what Palmer calls) "apparent" and "operational" features.

With the virtual representation approach, it becomes evident that the constructions in Anderson's proof actually leave the mode of representation intact. If <R,E,T'> is pictorial, then <R*,E*,f*T'f> will be pictorial too - it will actually be the same virtual representation as before, just implemented in a roundabout way.16 (To be sure, a shortcut wouldn't make it a different virtual representation either.) So really, there's just one virtual representation involved in any interesting sense in the theorem.

16) I write T' here to indicate the elementary operations that define the virtual representation; whether T' would be simply a subset of T I'm not sure, it isn't very clear what Anderson intends to be included under T.
Levels

Due to the uncertainties about the intended concept of representation it is hard to tell what the mimicry theorem really is about. One would have expected the mimicry theorem to be a theorem about the possibilities and consequences of implementing one (virtual) representation in terms of another (virtual) representation, or the feasibility of implementing a particular behavior in terms of different types of (virtual) representations. Even if this (expressed in my terminology) isn't what the theorem really amounts to, it is nevertheless what it is purported to be about, and what the discussion of it centers on. Anderson's general point, transformed to fit the virtual representation concept, would presumably look something like this: a given behavior underdetermines the underlying virtual representation, and a given virtual representation underdetermines how it is implemented.

The third point in the theorem - the timing aspect - is crucial; will not timing constraints shine through to higher levels? Anderson thinks that by speeding up the mimicking or underlying processes we can always save the phenomena. Hayes-Roth objects that Anderson's explanation doesn't account for differences in complexity. For example

"...the use of simple threshold units or logic gates permits instantaneous template-based pattern matching on appropriate analog representations." (Hayes-Roth 1979:377)

whereas

"...template recognition within a propositional framework requires general graph matching capabilities, and this requirement places template recognition within the class of nondeterministic

- John R. Anderson
polynomial-complete (NP-complete) computational problems." (ibid.)

Pylyshyn (1979) makes essentially the same point. I quote an instructive example of his:

"Consider the case of two systems that do arithmetics using different number systems ... the conventional decimal system and a system that represents number according to their prime decomposition. ... The latter system makes multiplication very easy ... but ensures that addition is extremely complex. To define an addition operator to mimic the one in the decimal system, two translations between the prime decomposition and the place-valued decimal system would now have to be undertaken ... in general this translation has arbitrary complexity ... Furthermore, there is no way to mimic columnwise addition without first converting the entire number to decimal form, since the relevant columns are only defined relative to this notation" (Pylyshyn 1979:391)

Anderson tries to fend off these attacks by reference to assumptions about the type of machinery used, for example "If, rather than assuming our computing device is a serial computer or a Turing machine, we assume it is a Boolean network ... we would get linear time complexity functions [for the propositional template-matching task]" (1979:397), or even by speculations on the physical constraints on the geometry of the hardware. As a final resort, he refers to the finiteness of the human cognitive system:

"There are clear bounds on the complexity of the problem that can be presented to the human system. ... As long as there are complexity bounds
on the possible problems, the mimicking process can be sped up (with a finite bound on the amount of speedup) to be able to mimic the other processes on all problems that are actually possible." (Anderson 1979:405)

This is probably true, but isn't it beside the point? A universal Turing machine is generally thought to be a better model of a computer than a finite automaton is, although the unboundedness of the Turing machine is glaringly inconsistent with the finiteness of the computer.

But whatever we may think of this, it is quite clear that the general point of view of AI is very different from the psychological point of view. The approach of psychology, as manifested by Palmer and Anderson, is generally that of analysis. The approach of AI is as much synthetic as analytic. Rather than inferring internal representation from external behavior, we are trying to specify and implement representations suitable for creating a certain behavior. On this, Anderson's arguments have little bearing, whereas complexity arguments are compelling.

- John R. Anderson
What have philosophers said about pictorial representation? Two and a half millenia of philosophical production isn't easily surveyed, even if we may disregard a good part of it on the ground that it does not conform with certain reasonable demands on any theory of representation useful to AI; such as symbols being concrete (things, processes, states, or whatever) and representational systems being in principle possible to implement physically. I will give a sketch of some of the more salient lines of thought, without any pretense of doing justice to what philosophy may have to contribute to our understanding of pictures and pictorial representation, of potential interest to AI.

From an AI point of view, what can we expect from a philosophical theory of how pictures represent? We shouldn't expect a blueprint for a fullfledged working symbolsystem. The typical philosophical theory of representation is concerned not with how symbols are actually employed, but with general criteria for symbolic relationships. Let us consider what it may have to say on some of the fundamental questions about the reference of a symbol—here illustrated by the verbal symbol "green" and the set of green things:

1) why green things (and not, say, grue\(^1\))?

---

1) See Goodman 1983.
2) why is "green" associated to green things (and not, say, blue)?

3) how is "green" related to other symbols (in the same symbolsystem), and how is this structure related to the structure of the represented world?

4) how is the symbol "green" established given something green, and how is "green" tested for applicability to something given?

5) how is the fact that "green" applies to green things, taken account of, internally?

Some theories (Aristotle) have an answer to 1), others (Goodman) have a philosophical explanation why this ultimately isn't a philosophical problem (but a problem for cultural anthropology, biology, etc). I agree that the question probably doesn't belong to theory of representation proper. On the other hand, I believe it is essential that a representational system, to be useful in an AI-system, carves up the world in the right way, moreover that it will have to be a way largely determined by how humans do it.²

Question 2) is the piece de resistance of many theories of representation, giving criteria or even some sort of mechanism of reference, which possibly might provide useful hints at how a working symbol-system could be constructed. Other theories dismiss the question as falling outside philosophy. Largely, it is matter of symbolic relations being considered conventional or not. Also 3) has drawn much attention, perhaps more than 2) in recent theories. Here too, we expect to find important clues to the design of actual representational systems.

²) See the frame paper, part III: Metaphysics. Compare also question 3.
Yet the answers are abstract and general, and will not tell us in any detail how one practically goes about deciding if a given representation applies, forming a representation of a given scene, etc. Question 4) fall outside the scope of most philosophical theories. Neither does 5) usually get a practical answer. The typical answer refers to the dispositional properties attached to the symbols, and the meaning implicit in their relations to each other, but this is not much to go by when trying to build a working symbol system. Also, it is easy to be deluded by the parallels with man-made symbol systems used by human beings, where the user of the symbols usually has access both to the world of the symbols and to the world symbolized (or at least a similar type of world) on an equal footing. An AI-system will have access to the symbolized world only through the very symbols the relations of which to the symbolized world we are studying.

The classification of different theories for pictorial representation I have used below is of course rough but serves the present purpose. One should also bear in mind that some of these theories take a rather narrow view of pictures and pictorial representation, they are about pictures in a literal sense.

Congruence theories

Copy and illusion

The first four types of theories I want to consider have in common that they can be seen as progressively diluted versions of copy theories. Copy theories go back at least to Plato and Aristotle (mimesis); a
work of art is conceived as an (inferior) copy of the depicted object (which in Plato is an inferior copy of the object's Idea). There are restrictions as to what aspects of the original one actually attempts to copy - for example, a marble sculpture of a man doesn't copy his innards or his voice. Normally, in painting and sculpture copying is restricted to certain visual and (possibly) tactile properties of the object's surface. With a slight shift of perspective, we might say that we're not really trying to copy the thing, only (certain dimensions of) the thing's appearance, such as visual and tactile properties.

Of course, characterizing a portrait (say) as being simply a copy of the object's visual appearance is to simplify in the extreme. Consider the problem of viewing point. Or the problem of movement:

"...the portraitist who wants to compensate for the absence of movement must ... so exploit the ambiguities of the arrested face that the multiplicities of possible readings result in the semblance of life. The immobile face must appear as a nodal point of several possible expressive movements. As a professional photographer once told me ... she searches for the expression which implies all others" (Gombrich 1972:17)

With yet another slight shift of perspective, we have the view that it isn't appearances in the sense of properties but appearances in the sense of impressions that are to be copied. The picture should give an impression equal to the impression we get from the original object, thus creating an illusion of it. This particular variety of copy theories might consequently be called illusion theories; an illusion being a kind of virtual copy (although I will not
The most obvious interpretation of "impressions" here, is perhaps as simple sense impressions, sense data. Alternatively, one could choose some subtler or more complex interpretation (and possibly more ill-defined), and get other, more sophisticated varieties of illusion theories, which may accord better with various optical illusions, trompe l'oeil, and the fact that something may give us subjectively an equal impression under the most varying circumstances. In any case, to "copy" an impression consists in producing something which gives the same impression as the original. Working in different materials than the original, it may not be obvious what that something is (which makes it a problem for the artist), in particular from the point of view of some sophisticated illusion theories which do not allow of any point by point copying.

Copy theories typically embrace an assumption about the "innocent eye"; a copy theory recipe for making a faithful picture is to empty one's mind and make an imitation, or reproduction, of the original, just as it is, or just as one sees it. The finished picture is properly viewed with an equally empty and unprejudiced mind. Gombrich (1960) has made a strong case against this myth of the innocent eye. How we see, and even what we see, is related to wish and prejudice. Seeing is not passive registration, but active construction relative to habit, knowledge, and goals. 3 There are many different ways of interpreting

3) In view of recent research on human vision (see Marr 1982), the connection which Gombrich makes between prejudice and active construction, needs a comment. Certainly, the most deeply rooted prejudice is the "compiled", or built-in, into the very form of the representation, and the apparatus of perception. The "active construction" of the human vision up to the point where we start to classify objects is not problem solving, in the usual AI sense, but routine
paintings, associated with different styles of depicting. The parallel with the thesis about the "theory-ladenness of observation" in current philosophy of science, is no coincidence; Gombrich is an old friend and follower of Popper.

The postulate of the innocent eye does not appear to be a required ingredient of a copy theory of the illusion brand. One possible alternative is copying under a given interpretation: that is, the copy and the original appear equal under that interpretation. Another possibility is to have different interpretations for original and copy, which is really just a recognition of the obvious fact that when we're viewing a picture as a picture we're aware that it is a picture. Illusion should not be confused with delusion (pointed out, for example, by Black 1972); if a picture deludes me into believing that I have the depicted object in front of me, then the picture clearly isn't a picture to me. (Besides, it's a rare event that a picture, even with elaborate staging, really deludes a viewer.) One could even claim that to be able to see a picture as a picture in the proper sense, one must know and be attuned to the particular style of depicting so that the right interpretation is applied.

Gombrich's own theory certainly has affinities with the double interpretations proposal. He maintains that one of the artist's main problems throughout the history of art has been how to create illusions, how tasks of information processing. Goals are really very long-term, and prejudice at the low level is metaphysical assumptions built into the algorithms; for example, the (visual) world is made up of "things" composed of smooth surfaces, which move and change smoothly.

4) Of course, if the "interpretation" of the original, of the world, is only one and fixed, we may not wish to call it an interpretation.
to achieve desired effects, with the means he has at his disposal. The means are changing, and interpretations are in a flux; usually, there are even several contemporary ways of depicting and viewing art (schools). The means for depicting - materials and techniques - influence, and are influenced by the standards of viewing art, the interpretations.  

One consequence is that the way a particular painting is interpreted may have more to do with how similar paintings have been interpreted than with how the corresponding object is usually viewed. The single interpretation variant, on the other hand, would seem to imply that, as new means for depicting evolves, and thus new ways of interpreting pictures are acquired, we also interpret the world in new and different ways. (Of course, a similar effect is compatible with the double interpretation theory, as well.) In fact, it has been suggested by Gombrich, and more strongly by Goodman, that our ways of interpreting pictures, do have an impact on the ways we view reality; Gombrich in a rather literal sense, Goodman in a more general, cognitive, sense as our ways of understanding reality.

Interpretations would seem to vary grossly in their "naturalness", and the degree of conscious intellectual effort required. Is it just a matter of habit? "For the last seventy years enthusiastic aestheticians ... have assured us that once we become used to the new style in painting, Picasso's women will look to us just as natural and life-like as Vermeer's. Alas, the miracle has not happened." (Zemach

5) Gombrich tends to stress the struggle to meet the standards with available means - the "puzzle-solving" activity of the artist, if we extend his own Popperian interpretation of art history to a Kuhnian perspective - rather than the "revolutionary" process of inventing new means of depicting, new interpretations.

- Congruence theories
Let us consider what copy and illusion theories might mean to AI. The main objection, I think, is that we don't want copies (even if we could have them).

"Art is not a copy of the real world. One of the damn things is enough." (motto in Goodman 1968 - unfortunately he has been unable to locate the original source)

As with art it is with knowledge representation. We don't want the map to coincide with the terrain.

True, the prime purpose of a symbol (in knowledge representation) is to be at hand when the real object isn't available - because it isn't there, or isn't in full view, or doesn't exist, or must not, or cannot, be manipulated with, etc. A copy, to be sure, could sometimes serve some of these purposes, but not all - for instance, we can't have a copy of something which doesn't exist, and never has existed - and not in a sufficiently convenient way. The requirement for convenience, which really is just a natural extension of the requirement for availability, means that the symbol must be easier to inspect and handle than the real thing (even were it available) - quick, because reasoning should be faster than real actions, and small, because many symbols are needed.

When we copy a thing we copy its drawbacks too: if the object is opaque, neither can we study the inside of the copy, if the object is heavy and bulky, so is the copy, etc. On the other hand, for each of its drawbacks we drop, it becomes so much less a copy, and the corresponding information is lost; if we choose to make the copy of weightless material, weight is no longer represented, if we make it colorless, colors are no longer represented, etc.6

6) No doubt, the examples look silly applied to computers, which excel in symbols that have no weight
Equally important is perspective. A symbol should present a particular point of view of the object, or we have just deferred a difficult interpretation problem. A symbol is used with a purpose; then we don't want it to reflect aspects of the object that have nothing to do with that purpose, this would only be confusing and slow down reasoning by continually repeated interpretations of the symbol in the ways appropriate to the purpose. This is why a drawn map may be handier than a photograph. Indeed, some hold that symbols are necessarily interpretative. "A picture never merely represents x, but rather represents x as a man or represents x to be a mountain, or represents the fact that x is a melon." (Goodman 1968:9).

This is the general objection. Sophisticated illusion theories of the type Gombrich represents, are not quite as incredible as crude copy theories (for AI). But what could it mean that an internally stored picture gives an "illusion" of a real, external, object? It could perhaps mean that the net result of contemplating a pictorial representation would somehow be the same as of observing a corresponding real scene. One problem with this idea is that there could be many ways to create the same effect that we wouldn't like to call pictures. We might improve on this idea by also demanding that the mechanisms of "inner vision" and (outer) vision although facing distinct worlds, share a common apparatus deeper in. But how much of the mechanisms must be shared before it becomes a "pictorial representation"?

nor inertia, take up virtually no (physical) space, and yield to manipulations almost instantly, but of course the argument extends to more abstract properties of symbols.
Finally there is the general limitation of illusion theories to what is perceivable. Per definition, illusions are only of what can be perceived. Since computers usually don't have eyes - how can they have "pictorial representations"? A more general problem is how an illusion theory could ever be made to cope with "pictures" of things that cannot be perceived. Diagrams do not give illusions (in the sense of illusion theories).

**Similarity and resemblance**

Further removed from the original crude form of copy theories and of illusion theories, we have the corresponding weakened forms: similarity and resemblance theories. Similarity means that the picture and the object are similar, or that the picture is more similar to the object than to objects of any other class; resemblance means that the picture and the object give similar impressions, or that the impression of the picture is more similar to the impression of the object than to impressions of objects of any other class.

Inherent in copy and illusion theories, there is an idea of the perfect copy, or the perfect illusion, at least as an ideal. Different pictures of one object all point, as it were, in the same direction - the ideal picture. Similarity and resemblance theories are less committed to convergence to some supposedly ultimate ideal depiction, they are compatible with the perspectival demands on symbols. Also, whereas copy and illusion theories tend to indulge in a reproduction of low-level details, similarity and resemblance theories allow of a concentration on high-level agreements.

Most of the problems of illusion theories are shared by resemblance theories. And there are additional
problems: for example, doesn't a portrait resemble any other portrait more than it resembles the person portrayed? The main problem with similarity is that it is such a vague and elusive concept. Any two things are similar in some respect. Is it really possible to make sense of a general concept of similarity, that manages to pick out precisely the cases of genuine pictorial representation, and not anything else? Note however that, thanks to the varied and abstract relationships that may be taken as instances of "similarity", similarity theories, alone, have the potential to account for pictorial representation in the most general sense.

A standard formal objection to resemblance and similarity theories is that the logical properties of resembles and is-similar-to are incompatible with those of represent (e.g. Goodman 1976). For example, resemblance and similarity are symmetrical, whereas representation isn't: a tree doesn't (normally) represent a picture of it. The underlying asymmetry of copy and illusion, as we would normally understand these terms (we wouldn't say that a person gave an illusion of her own portrait), have thus been lost in the transfer to resemblance and illusion. But, as Black notes (1972:119n) there are closely related concepts in ordinary language that retain asymmetry, an example being "looks like".

A general problem for the whole group of congruence theories is that of pictures representing non-existent objects. What is it that a picture of a unicorn is a copy of, gives an illusion of, resembles, is similar to? Nothing. This is a particularly grave problem for knowledge representation and

7) Note that even we had a "perfect copy" (if there were such a thing), indistinguishable from the original, we wouldn't normally say that the original were a copy of the (true) copy.
AI, which makes such extensive use of hypothetical reasoning.

Information

An influential variety of resemblance theory is the information theory of James J. Gibson. A picture presents us with the same (or — in newer versions of the theory — partially the same) information as the real object. Gibson (1971) defines a picture to be:

"...a surface so treated that a delimited optic array to a point of observation is made available that contains the same kind of information that is found in the ambient optic arrays of an ordinary environment." (1971:31, my emphasis)

Note that this is a general definition of what a picture is. A particular picture represents a particular object, if the picture and the object give the same information.8 Gibson has since dropped the condition of same information (and removed the last remaining traces of his earlier sense-data illusionistic theory):

"There is no such thing as a literal representation of an earlier optic array. The scene cannot be re-established; the array cannot be reconstituted. .... Even a photograph, a color photograph at its technological best, cannot preserve all the information at a point of observation in a natural environment, for that information is unlimited." (1979:279)

Gibson energetically puts his information approach in contrast to illusion theories and any attempt to

8) Goodman didn't notice the difference in his 1971 comment.
associate pictorial representation with sensations; "an artist can capture the information about something without replicating its sensations" (1971:31). Nevertheless Gibson's theory belongs to the resemblance category rather than the similarity category, since the theory is based on the "phenomenal visual world" (although it would seem not to be resemblance in the ordinary sense). The information theory, Gibson points out, explains why

"when one sees a pictured object one ordinarily does not see its front surface only but the whole of it." (1971:31, my emphasis)

One sees a phenomenal object, not just some form in the phenomenal visual field.

The crux, of course, is the concept of information. What sense of information is intended? It would seem to have to be something quite close to the intuitive, content-related, notion. Unfortunately, a successful explication of the intuitive information concept has been lacking; as Black (1972) convincingly demonstrates, none of the then existent, quantitative, technical concepts of information have much to do with the intuitive, qualitative and content-related, concept. The more recent work by Dretske (1981) is a promising attempt to develop a theory of information which doesn't eschew the question of informational content.

Gibson's approach is apparently threatened by triviality: what symbol doesn't give us information? What

is the difference between information obtained from verbal symbols and information obtained from pictorial symbols? Why isn't a description a depiction? Gibson actually has in mind a very special and unusual concept of information that doesn't fit in well with the intuitive notion. Optical (read: pictorial)

"...information consists of invariants, in the mathematical sense, of the structure of an optical array." (1971:31)

It is a content related concept. Different objects have different optical invariants. What doesn't fit with our intuitive notion of information is that the form of the information seems to matter. If I gave a description of the optical invariants of some object, would I produce a picture of it then? Clearly not. The only explanation seems to be that depiction and description differ in the form of information they provide. Possibly, this is what "kind of information" alludes to, in Gibson's definition of picture quoted above.

Now, Gibson actually hints that the invariants cannot be described:

"Most of the formless invariants in the array from a picture could not be put into words anyway. They can be captured by an artist but not described." (1979:274)

What is not clear is if this inability is just a manifestation of human feebleness, or if it is an impossibility in principle. One would have thought that the concept of a mathematical invariant presupposes that it is in principle possible to describe as a mathematical formula. That the invariants really are in principle impossible to describe, would mean that one would have to give up all attempts to formulate
Despite Gibson's rejection of illusion theories, his theory still has in common with them, that all pictures of an object approximate a unique ideal "picture" of it, which would be the object itself; the information obtained from a picture (about its object) is always a part of the full set of information obtained from viewing the object itself.

Goodman (1971) argues (rhetorically) that the invariants may turn out to be - the object itself! What else is constant among the several views or pictures that can be made of an object? Now, Gibson (1971) actually concedes the existence of unfavorable points of view that do not provide the eye with the full set of invariants, and his (1979) discusses at length the fact that pictures always only preserve partial information about the object. That is, invariants, in contradistinction to the object itself, are not constant for all possible views. But with a slight modification of Goodman's argument, one could still perhaps identify the full set of invariants with the - phenomenal - object itself.

**History based theories**

The congruence family of theories relate the reference of a pictorial symbol closely to how the individual symbol manifestly is or appears. A radically different group of theories relate reference to the individual symbol's history; its origin, the circumstances of it creation, or adoption, etc.

There is a parallel to this genetical approach in Kripke's theory of how proper names in a verbal symbol system refer (Kripke 1980); for example, a person
gets his name by some kind of naming ceremony, which is the starting point of "chains of communication" linking each instance of use of the name by an unbroken causal chain back to the origin. I may, according to Kripke, successfully refer to that person by using his name, without knowing anything about him, so long as I'm at the end of such a chain. (I may have heard the name from my neighbor who read it in the newspaper in an article by a journalist who actually heard the person call himself by that name.)

**Cause**

One idea in this category is that pictures refer by virtue of a causal connection (of the right kind) between picture and object. The picture is a sort of trace of the object; the object is a salient factor in causing the picture. Photography—a highly mechanized technique of producing pictures which are commonly taken to be the most undisputable examples of pictorial representation—is a case in point.

Of course, such a characterization of pictorial representation must be hedged in a number of ways to be reasonable. Not every picture caused by an object is a picture of the object; differently put by Black (1972) as the problem of identifying the object of the picture. For example, the artist is certainly an important causal factor in making a portrait of some other person.

Consider a more difficult case: suppose that Steve Mahre wants to surprise his twin brother Phil, with a portrait of Phil Mahre receiving the World Cup price in downhill skiing. Steve sits for an artist who believes that Steve is Phil. Is the finished painting a picture of Steve, or of Phil, or of both? Intuition falters. A more complex case would be if both Phil and Steve, on different occasions, posed
for the portrait. Next step would be to imagine that the artist, never having seen, or heard of, the Mahre brothers, in a strike of divine inspiration paints something that could pass as a faithful portrait of either of them.

Black (1972) thinks there are limits to what we would allow of this, which would imply that there's something to the causal history idea at least. According to Black, if we found a rock formation that by some whim of Nature happened to look like Napoleon, we wouldn't be entitled to say that it was a representation of Napoleon.

An attacker of the causal theory might point out that such things simply do not happen, which explains our putative reluctance to accept it as a pictorial representation of Napoleon. In our world, as an empirical (statistical) fact, a detailed picture of a complex object simply doesn't come to existence (one way or the other) without a causal chain between object and picture - or possibly - from some originating event common to both the object and the picture. In the case of the Napoleon rock, one would immediately start searching for the causal chain everybody knows must be there (and similarly for the "divinely inspired" painter - is it perhaps that God has shown him how to do it?). If we lived in a world where a rock resembling Napoleon could form spontaneously we would be in a better position to estimate the validity of Black's contention.

The attacker might thus claim that Black has confused an empirical condition for the detailed depiction of complex objects in our world, with the general criteria for pictorial representation. A camera produces pictures, not by virtue of the process, but by virtue of the result (with the help of the process). If the photograph comes out as a grey blur, we don't
accept it as a pictorial representation, and, as indeed Black himself observes, "no genetic narrative of the photograph's provenance, no matter how detailed and accurate, can logically guarantee that photograph's fidelity" (Black 1972:103).

Another difficulty is how the theory is to account for the fact that a picture sometimes comes into existence before the depicted object does (as happens with many blue-prints), a reversed causal chain. And what about forking chains? In a CAD/CAM-system, some internally stored data may be the cause of both a picture on a video-screen, and the creation of an object in a milling machine. Is the picture on the video-screen not a picture of the object? Finally, we have the same general problem as with congruence theories, with regard to pictures of non-existent objects and objects that will actually never come to be - what are such pictures pictures of? Since representations of hypotheticals abound in AI planning and problem solving this is a particularly sore point for AI.

It seems unlikely that a causal theory, on its own, could give an acceptable account of the distinctive character of pictorial representation. Whatever it may yet have to contribute to a general theory of pictures, will be of limited interest to AI; it will not help in understanding how to test if a stored picture matches an observed object, or how to compare two pictures, or in understanding how the peculiar properties of pictures should be exploited, etc.

**Convention and intention**

A different type of attitude to the Napoleon rock, would be to simply decide to take it as a picture of Napoleon. Generally, the idea is to substitute the event of adopting a convention, for the event of
creating the symbolic object, as the decisive moment in the symbol's history which determines its reference. In one variety of convention theories, the attention is fixed on certain privileged private conventions. For example, a painting refers to whatever the painter intended it to refer to when he created it. Another variant is that the picture refers to whatever the viewer takes it to refer to.

Such a theory avoids some of the problems with causal theories, but leaves even more of pictorial representation unexplained. If we can't use any rock as a picture of Napoleon, then what is it about the Napoleon shaped rock that makes it susceptible of such a convention? (It is tempting to resort to some congruence theory.) Or, if we can take anything as a picture of anything, then what's the point with pictures? Are pictures just a sort of primitive words?

Summarizing the relevance of history based theories to AI: as far as the history of the symbol isn't written all over it, the history based approach isn't tractable for practical implementation. And should the history be written on the symbol, as efficiently accessible properties of it, then clearly there's no need for a history based theory of the type we have discussed here. This is not to dismiss history as generally irrelevant, rather a recognition that a working computer implemented symbol system based on the histories of individual symbols isn't practicable. I think that the relevant historical aspect should rather be sought in how the system of symbols has arisen and how it is related to previous systems of symbols.10

10) See the next section on Goodman's theory, and the following section on "Analog", subsection "Entrenchment".
Nelson Goodman is a particularly interesting philosopher from the point of view of knowledge representation. Not only because of his outstanding contribution of conceptual tools and thought-provoking hypotheses towards a theory of symbols, but also because his work is imbued with a pragmatic spirit most congenial with the aims and methods of AI. I will introduce those parts of his theory that I have found to be most useful for my present purpose. (Mainly in Goodman 1968, 1972, 1978, and 1984). In chapter VI I will develop my own ideas about pictorial representation and how it is related to verbal representation, essentially on this basis.

Goodman's major work on symbol theory, 'Languages of art', 1968 (second edition 1976), centers on symbols used in the arts. However, Goodman's basic ideas about symbols in the arts are quite as relevant to symbols used in science and technology. I'm also sympathetic to the thought of obliterating the traditional, sharp dichotomy between the arts and the sciences, which is one of Goodman's main themes; even the aesthetic experience itself "is cognitive experience distinguished by the dominance of certain symbolic characteristics and judged by standards of cognitive efficacy" (1968:262).

**Denotation**

Denotation is a basic concept in Goodman's theory of symbols. It is "the application of a word or picture or other label to one or many things" (1984:55). For example, a predicate, or more generally a description, denotes each entity it applies to severally, a name the individual it names, etc. Thus, for example, "horse" denotes each individual horse
separately; not the set of horses, nor the property of being a horse, nor the concept of a horse. This concept of denotation applies also to all sorts of non-verbal symbols like gestures, musical notation — and — to pictures. Indeed,

"denotation is the core of [pictorial] representation" (Goodman 1968:5).

Note that Goodman uses the term "representation" for what I call "pictorial representation". "Representation" and "represent" in the general sense implied in this paper — which is ordinary AI usage — agree reasonably well with Goodman's terms "symbol", "symbol system", "symbolization", and — note — "refer", which are Goodman's most general terms.

A picture of Churchill denotes Churchill. A picture of an eagle in a dictionary denotes each eagle; the picture is coextensive with "eagle". Any sort of symbol that denotes is a label. A predicate is simply a special kind of label; it is a label from a linguistic symbol system. Anything denoted by a label complies with it, and if we collect all things that comply with a label we get its compliance-class.\(^1\)

Many descriptions and predicates lack denotation, for example "unicorn". What about a picture of a unicorn? According to Goodman, both "unicorn" and a picture of a unicorn are labels that lack denotation.\(^1\) The problem is — if a picture of an eagle

---

1) That is, its extension in usual terminology, but remember that Goodman's concept of denotation is wider than what is usual.

12) Yet they are labels, which is a little confusing. What is the criterion of being a label if we drop the condition that it label something? A symbol always belongs to some system of symbols, within which it is recognized as being a label — denoting or not.

- Nelson Goodman
denotes eagles, how can we avoid saying that a picture of a unicorn denotes unicorns?

Goodman's solution is to distinguish carefully between what a picture denotes and what kind of picture it is. A picture of a unicorn can be classified as a unicorn-picture—that is, as being of the unicorn kind of pictures, similarly to how we may classify a piece of furniture as Chippendale—without any commitment to the existence of unicorns. Generally, saying that an object is "a picture of X" is ambiguous between saying that the picture denotes X and saying that the picture is of X-kind, or is an X-picture (in Goodman's handy notation). Goodman points out that there is a close parallel in language, for example: "Sancho Panza's master" and "the man who conquered Napoleon" are both man-descriptions, but only the second denotes a man.

So, a picture of a man can be a man-picture without denoting a man. More surprising, the converse is also true: a picture denoting a man, isn't necessarily a man-picture; take for example the caricature of Darwin as a monkey, or Louis Philippe as a pear. Generally,

"an object k is represented as a soandso by a picture p if and only if p is or contains a picture that as a whole both denotes k and is a soandso-picture." (1968:28)

(and similarly for description-as). A lion-picture denoting Churchill, represents Churchill as a lion. An ordinary picture of Churchill, represents Churchill as Churchill. And all pictorial representation is representation-as. Thus,

"the denotation of a picture no more determines its kind than the kind of picture determines the denotation" (1968:26)
"Almost any picture may represent almost anything" (1968:38)

Despite the striking element of convention in Goodman's conception of pictorial denotation, his theory of representation is structural rather than historical. The denotation of a symbol is determined by its place in a system of symbols. The place of a pictorial symbol in a pictorial symbol system, the relations it bears (as a symbol) to other symbols in the system, in short, which symbol it is in that system, is determined by its pictorial properties, roughly, properties which can be described in terms of "what colors the picture has at what places" (1968:42). Whereas the pictorial properties of a picture has no intrinsic connection with what it represents - contrary to naive congruence theories - they do determine the symbol's place within a given symbol system, and so indirectly determine the symbol's denotatum - contrary to history based theories.\(^{13}\)

What then is a symbol system? Some examples are: musical notation, written English, spoken French, Labanotation (for dance), oil paintings, black and white photography, a tailor's samples of cloth, poetry, a car's speed-indicator. Behind these examples we recognize a familiar, albeit vague notion. Goodman nowhere gives a very precise characterization of the general concept of a symbol system - a difficult task for a subject of such wide scope - but his extensive and penetrating explorations and explanations of various important subtypes, add up to form

\(^{13}\) This is certainly not the only interpretation of Goodman's theory of pictures. Robinson (1979), in a lucid article, clears up some misunderstandings, and brings out the pervasive theme of systematicity in Goodman's theory.
quite a powerful theory.

It will not be possible to recapitulate these investigations here, so a few comments on the general approach might be helpful.

i) A symbol system embraces both the symbols and their interpretation - it consists of a symbol scheme (uninterpreted symbols) correlated with a field of reference.

ii) A symbol is a symbol only by virtue of being part of a symbol system - symbols have no independent existence as symbols, no intrinsic relations to what they symbolize.

iii) A symbol as object can be a symbol in several different symbol systems.

iv) The exact content and boundary of a symbol system are often vague. One must be aware of a certain risk of dismissing difficult points by reference to their dependence on a symbol system that forever remains unspecified.

v) Denotation is not the only way a symbol may refer (or represent, in AI terminology); exemplification and expression are two other important symbolic relations to be explained below.

A note on terminology. Following Goodman, I use symbol as the most general term for signifying entity. Some writers prefer to use "symbol" in a much more restricted sense (in the sense that, for example, a cross is a symbol for the Church), distinguishing symbols from, for example, entities signifying concrete things, from entities signalling (initiating) an activity, from entities causally connected to what they signify (so-called natural signs), etc. This type of typology is probably of little interest to AI.14

14) When Goodman wishes to distinguish between type and token, he speaks of character and
There are other varieties of denotation besides denotation in natural language and pictorial denotation. Symbols in a natural language may be ambiguous (as for example, "bank"); two symbols need not be semantically disjoint (as for example, "father" and "son", or "triangle" and "polygon" are not); and they are not what Goodman calls semantically differentiated, which means that it may not be possible to determine which symbol(s) some given object is denoted by: whereas we may form arbitrarily precise descriptions in a natural language, we cannot make arbitrarily precise measurements, thus for example, it is conceivable that there is an object which we cannot decide if it complies with "one meter long, or less" or with "more than one meter long".

Notation, discussed at length in Goodman (1968), differs from ordinary natural language denotation not in how a single symbol is related to its denotatum, rather, being notational is a feature of the system of symbols as a whole. Notational systems, contrary to natural language, are characterized by nonambiguous symbols, non-overlapping denotata, and they are semantically differentiated. They share with natural languages the property of syntactical disjointness, that is, no token belongs to more than one type of symbol; and the property of syntactical differentiation, which means that it is always possible to decide which symbol(s) a particular token is an instance of. The standard system for musical notation is an example of a notational system. 15

Inscription respectively, but I have refrained from this usage; also there are in this context several fine points which I have decided to spare the reader. 

15) Notational systems are not plausible as general vehicles of knowledge representation; in an AI application it should be possible to represent a thing in different ways, from different aspects, and with varying degrees of abstraction, none of which is possible within a notation (in Goodman's sense). I suspect that even ambiguity may turn out to be

- Nelson Goodman
A different sort of classification of instances of denotation is the distinction between ordinary, literal, denotation, and metaphorical denotation:

"Metaphor arises by transferring a schema of labels for sorting a given realm to the sorting of another realm (or the same realm in a different way) under the guidance or influence or suggestion of the earlier sorting." (1984:61).

Once the metaphorical application of a few labels to an alien realm is fixed, the metaphorical use of related labels is implicitly determined by structural constraints transferred from the old realm. For example, if "high" is used metaphorically for certain types of sound, then it is immediately clear what types of sound "low" must metaphorically denote. And, on the other hand, the first application of "high" to a certain type of sound only becomes comprehensible by it being "natural" (or "right") to organize types of sound in a way structurally similar to how "high" and symbols semantically closely related to it organizes whatever realm they normally apply to.

Literal and metaphorical denotation differ not in genuineness - Nixon was a "hawk" as actually as he was a "president" - but rather in novelty and resistance to application. "Briefly, a metaphor is an affair between a predicate with a past and an object that yields while protesting." (1968:69).

essential, and that the impressive combinatorial power of natural language which notational systems so obviously lacks wouldn't be possible without it. Of course, at a low level, notational systems are appropriate and indeed indispensable - or rather a prerequisite - for a computer science based on digital computers.
Exemplification and expression

Evidently, many works of art do not denote (somewhat contrary, it seems, to the quotation above). A visitor at an art exhibition can hardly be a confident judge as to whether a picture of a landscape, say, has or lacks denotation. Yet, in most cases, this is of little consequence for his judgement of the painting as a work of art.

Granting Goodman's general tenet that a picture - as a work of art - should be understood in terms of different functions of symbolizing, there must be more to pictorial representation than mere denotation. What is it? Goodman points out two symbolic relations of fundamental importance: exemplification and expression.

One of Goodman's examples of exemplification is a tailor's booklet of swatches of cloth. Such a swatch exemplifies certain properties, or rather - to be a good nominalist - predicates (or any kind of label): a color (color predicate), a weave (weave predicate), a texture (texture predicate), a pattern (pattern predicate), etc. It doesn't exemplify all of its properties; not, for example, its shape or size, nor even all of those properties it shares with the given run of cloth, such as having been finished on a Tuesday.

Thus, contrary to the common view that exemplification is a relation between an object and a set (of - in some sense - similar objects), Goodman takes it to be a relation between an object and a label.16

An exemplifying symbol fulfills two conditions: i) it is a symbol denoted by the label it exemplifies; and

16) This is a good example, I think, of what makes nominalism attractive to AI.
ii) it also refers to that label:

"To have without symbolizing is merely to possess, while to symbolize without having is to refer in some other way than by exemplifying." (1968:53).

A 5 x 5 cm rectangular yellow swatch, exemplifies "yellow", because it is yellow and it refers to "yellow". It doesn't exemplify "5 x 5 cm" in the present context, because, although it complies with this description it doesn't refer to it. Similarly, it doesn't exemplify "rectangular" since it doesn't refer to it; although in this case "rectangular" happens to apply also to the run of cloth.

Expression is an even more complex symbolic relation. It is explained as metaphorical exemplification. That is, a symbol expresses a label by being metaphorically denoted by this label, and by referring to it. For example, a painting can express sadness by being a metaphorical example of sadness; that is, by being metaphorically denoted by "sad" and by referring to that label. Moreover, it can express sadness without denoting anything sad or indeed anyone or anything expressing sadness.

Exemplification and expression are not restricted to pictorial representations; they occur in verbal representations too. For instance, a passage may express speed or excitement by metaphorically being denoted by "fast" or "excited"; without describing fastness or excitement.

17) Remember that Goodman's sense of "refer" is very general.
Summary of concepts

To sum up, we have in Goodman's theory the following "attributes" of a picture or a verbal symbol: its kind, what it denotes, what it exemplifies, what it expresses, what it is denoted by, and its identity as symbol. These are generally distinct.

Generally, there need to be no congruence between what a pictorial or verbal symbol denotes, what it exemplifies, and what it expresses; "a tale of fast action may be slow, a biography of a benefactor bitter, a description of colorful music drab, and a play about boredom electric." (1968:91f).

Being a soandso-picture is neither necessary nor sufficient for denoting a soandso, as we have seen. Clearly then, since kind does not determine denotation, the kind of picture it is must not be identified with the picture's identity in a symbol system. Besides, a picture may be of several kinds. Similarly for descriptions.

As for exemplification - a picture need not exemplify its kind, but it is of course denoted by it. However, not just anything which a picture is denoted by has to do with its kind; "Representation of Churchill by a black ink drawing, or by a small painting, does not constitute representing Churchill as black, or as small." (1972:124). Similarly for descriptions.

Denotation, exemplification and expression do not exhaust the possible ways of symbolizing, but many cases of reference can be understood in terms of

18) At one place, Goodman writes that: "A picture of a green goblin ... is, and exemplifies being, a green-goblin-picture. Description-as and representation-as, though pertaining to labels, are likewise matters of exemplification rather than of denotation." (1968:66). This slip has since been corrected (1984:82).
chains with each link an instance of one of these three basic types of reference. For instance,

"a [picture of a] bald eagle denotes a bird that may exemplify a label such as 'bold and free' that in turn denotes and is exemplified by a given country." (1984:62)

Words and pictures - similarities

What is then the difference between linguistic and pictorial representation? Goodman is far more energetic in pointing out similarities than in finding differences, thus diverging from the mainstream of philosophical thought on pictorial representation. Pictures are as genuinely denoting as words, and as for any kind of symbol and any kind of reference, they refer not by virtue of some intrinsic relation to what they represent but by virtue of their place in a symbol system. Pictorial denotation is no less than verbal denotation a matter of convention. A picture, just as a description, may denote several things, a single thing, or nothing at all. We even find the same kind of ambiguity in speaking about null-denoting labels; "a picture of X", "a description of X".

"This all adds up to open heresy ... The often stressed distinction between iconic and other signs becomes transient and trivial" (1968:231f)

Goodman puts some effort into demolishing the group of congruence theories, base for the traditional strong discrimination between pictorial and verbal modes of representing. He joins Gombrich in the battle against the myth of the innocent eye: "The most ascetic vision and the most prodigal, like the sober portrait and the vitriolic caricature, differ not in
Some of Goodman's objections stem from his general philosophical convictions. A picture cannot be a copy of the object as it is, Goodman says, because there is no such thing as the way the world is. Neither can it copy all the ways the object is - because they are infinite and indefinite, nor can they be conjoined since "conjunction itself is peculiar to certain systems" (1968:6n). Neither can a picture copy any particular way the object is, any particular aspect of it, because "an aspect ... is the object as we look upon or conceive it, a version or construal of the object. In representing an object, we do not copy such a construal or interpretation - we achieve it." (1968:9)

As we have seen, there are also ontological problems with congruence theories. Is-a-copy-of, gives-an-illusion-of, is-similar-to, resembles, are all binary relations. They seem to imply that if a picture exists then its object must exist, too. Goodman's theory is designed to solve that problem; a picture of a man can be a man-picture without denoting a man.

Goodman's theory of pictures is radically conventionalistic, as opposed to Gombrich who yet remains a kind of illusion theorist. For instance, Goodman calls in question the non-conventionality of perspective in art, vigorously defended even by Gombrich. Still, I wonder if the issue between a congruence theory with interpretation pluralism and Goodman's point of view may not be more "ideological" than factual. If we understand a way of interpreting as something systematic, orderly and relatively stable, its qualities and its basic role have a remarkable similarity with a symbol system. Both serve as a
framework necessary to determine a symbol's func-
tions. It is true that Gombrich tends to stress pro-
cedural and subjective aspects in his concept of an
interpretation, whereas Goodman's investigations
center on structural properties, but this is hardly a
serious obstacle to a reconciliation. We might even
choose to think of this concept of interpretation as
a presystematic counterpart, a forerunner, to
Goodman's more articulated concept of a symbol sys-
tem.

Despite the parallels between pictures and words, of
course Goodman doesn't deny that there are differ-
ences. But they are, with one exception, differences
of degree. Pictorial symbol systems do not have the
organizational complexity of a natural language, but
then again, many verbal representations don't, so in
this respect pictorial representations are comparable
with simple forms of verbal representation. Pictures
possibly rely more on expression (metaphorical exem-
plification) than descriptions do. Pictures are in a
sense more replete, meaning they have more dimensions
of significant variation, however, this is in con-
trast also with diagrams that are pictorial in a wide
sense (but typically lack repleteness, and so is con-
sidered as something less than full-fledged pictures
by Goodman).

The difference

The exception, the decisive qualitative difference
between words and pictures according to Goodman is
that pictorial systems are analog - that is, syntac-
tically and semantically dense - whereas verbal sys-
tems are not. More exactly: only syntactic density
is required, but pictorial systems are normally
semantically dense (1968:227n).
A syntactically dense symbol system is a system such that between every two symbols there is a third symbol. A semantically dense system is a system such that between every two compliance-classes there is a third compliance-class. In the 1976 edition, these definitions were amended with the provision that no insertion of other symbols (or compliance-classes, respectively) "in their normal place" (1976:136) may destroy density. The motive was that without these provisos certain types of "gaps" are not ruled out; for example, a symbol scheme consisting of symbols that "correspond to all rational numbers that are either less than 1 or not less than 2" (ibid.), is dense by the old definition.19

In these definitions, it is important to realize that it is the ordering of symbols and compliance-classes provided for by the symbol system that counts, not which symbols and compliance-classes happen to be actual. For example, a symbol system with but one actual symbol (token) may yet be syntactically dense.

I find the notion of inserting "other" symbols "in their normal place" dubious. Contrived examples apart, it is mysterious how alien symbols (from a completely different system) can have a "normal place" in a well-entrenched symbol system. The only plausible occasions would seem to be, as in the example above, when the alien symbols belong to an extension of the system in case. But aren't there many extensions to a given system? If all possible extensions would count, then we could always destroy density.

Goodman states that density implies total absence of differentiation, no symbol-token can be determined to

19) Strictly speaking, the amended definitions are of density throughout, but Goodman usually leaves out the "throughout" for convenience.

- Nelson Goodman
belong to one rather than to many other symbol-types. Intuitively, this appears to be correct, but a second look reveals a slight mismatch between the density definitions which are relative to betweenness orderings, and the differentiability definitions which are of an absolute flavor. To see this, let us consult the exact definitions of differentiation:

A system is **syntactically differentiated** if for every two symbol-types and every symbol-token that doesn't actually belong to both types it is theoretically possible to determine either that it doesn't belong to the first symbol or that it doesn't belong to the other. Similarly, a system is **semantically differentiated** if for every two symbols with non-identical compliance-classes and every object that doesn't actually fall into both classes it is theoretically possible to determine either that it doesn't belong to the first class or that it doesn't belong to the other.  

These definitions are relative only with respect to the interpretation of "theoretically possible" — which we are invited to make "in any reasonable way" *(1968:136)*; however I take it we are not allowed to postulate at random what is theoretically possible, moreover that anything which is practically possible is also theoretically possible. A consequence is that density is sometimes compatible with differentiation. For example, in the example of a symbol scheme with a gap cited above, the symbols that correspond to rational numbers could be fully reduced Arabic fractional numerals; although clearly differentiated, with the numerical betweenness ordering indicated, the scheme would yet be dense if we just fill in the gap.

---

20) The exact terms actually read "finite differentiation".
There are more problems with density. In most of Goodman's examples and discussions it is evident that the semantic ordering, the betweenness relation referred to in the definition of semantic density, is to be understood as something apart from and independent of syntactic relations.

For example (1968:163,163n), consider the subsystem of English consisting of the terms "halfway between a and b", "halfway between a and halfway between a and b", "halfway between b and halfway between a and b", "halfway between a and halfway between a and halfway between a and b", and so on ad infinitum. "This system is syntactically differentiated but semantically dense" (ibid.). In the accompanying footnote — which gives it an air of self-evidence — we are informed that the syntactic ordering of the terms is ascending according to their length, and terms of equal length are ordered alphabetically; the semantic ordering is according to left-right position of the compliant points.

But in some passages it appears that the semantic ordering is induced by the syntactic ordering: three compliance-classes stand in a between relation precisely if the corresponding symbols do.

A case in point is the following example (1968:240n) of a symbol system: the symbols are straight marks of lengths in inches equal to rational numbers; each symbol denotes an integer, namely the ordinal number of the corresponding rational number in some particular enumeration of the rationals. Clearly, the system is syntactically dense. But it is also, according to Goodman, semantically dense, which "results from the way the whole numbers are related in the 21) Elgin (1983:102) concludes that "a discursive language is syntactically ... differentiated. It is semantically ... dense".

- Nelson Goodman 179
system to the characters [symbols] referring to them" (ibid.).

Such an interpretation of semantic density results in a hollow and redundant notion: any syntactically dense system which has enough objects (and doesn't corefer too much) will automatically be semantically dense, and no syntactically non-dense system will be semantically dense.

Unwanted consequences

One consequence of Goodman's criterion for picturehood is that any digital pictorial representation — in fact, any syntactically non-dense pictorial representation — will disqualify as pictorial in Goodman's sense. And there are lots of them: ordinary color prints, TV-pictures, computer art, etc.

As a matter of fact, converting a picture into digital form is usually not accompanied by any significant change in our apprehension of the picture. In many cases an ordinary observation will not make out any difference.

Perhaps, our inability to discern its digital character should be taken to indicate that it is syntactically dense to us under those particular circumstances. But isn't it rather annoying that moving a color magazine a few centimeters closer to one's eyes will suddenly shift (what we normally call) the pictures in it to an entirely different kind of symbol system? And notice that even a digitalization that is conspicuous to an observer can still be a recognizable portrait to the same observer from the same point of view.

22) Notice, however, that Goodman's criterion permits, for example a geometrical digitalization as long as tint or shade remain continuously variable.
If we take picture in the most general sense, there are all sorts of examples of non-analog (in Goodman's sense) pictures: chemical structure formulas, electric wiring diagrams, itineraries used by medieval pilgrims, etc. And on the other hand, many symbol systems that are pictorial according to Goodman's criterion, fail completely to conform with our intuitions about pictures. For example, reverse the direction of all instances of denotation in a natural language (or some suitable subsystem) - it will result in a syntactically dense system - but is it any more reasonable to take a dog as a picture of "dog" than vice versa?23

Another trick is to supplement a semantically dense but syntactically differentiated system, with enough additional vacant symbols (that is, symbols with null-denotation) to make it syntactically dense, and thus pictorial. Once again, consider some part of a natural language that is semantically dense but syntactically differentiated. Any term may be viewed as a number in base N notation, where N is the number of different types of letters and delimiters (including space); a space is identified with 0, the letter A with 1, letter B with 2, etc. If we now just supplement this system so that there is a symbol corresponding to every rational number (in N-base point notation), it will become a syntactically dense system. (Notice that the original syntactic ordering is reverentially preserved.)

Are the benefits of Goodman's systematizing really worth its price of "unnaturalness", that is, bad

23) To be sure, if we retain the alphabetical ordering of terms as semantic ordering in the new system, it will be a semantically non-dense system, and as such an "abnormal" pictorial representation (cf. 1968:227n). But why not use the induced syntactic ordering as the new semantic ordering, similarly to Goodman's integer example cited above?

- Nelson Goodman
entrenchment in presystematic terminology? In chapter VI I will outline an alternative theory of pictures which is less at odds with existing classifications, and which hopefully is at least as systematic.

Let us examine closer what "analog representation" could mean, and what - if anything - it may have to do with pictures. To facilitate the discussion it will be useful to have a more formal definition of a denotative symbol system.

A denotative symbol system\(^{24}\) \(<S, O, R>\) has three parts: a symbol space \(S\), an object space \(O\), and a semantics \(R\). Intuitively, the symbol space contains all acceptable symbols, the object space consists of all objects that the symbol system is to represent, and the semantics consists of the correlations between symbols and objects. Formally, the symbol space is a set of symbols \(s\), and a set of first-order relations on \(s\), \(s[n,i]\), \(s[n,i]\) being relation number \(i\) with arity \(n\) \((0 \leq n, 0 < i)\). Likewise, the object space consists of objects \(o\), and relations \(o[n,i]\) on \(o\). The semantics is a relation with domain a subset of \(s\) and range a subset of \(o\). Notice that the semantic mapping is not a function, since that would rule out the possibility that a symbol associates directly to

---

\(^{24}\) This definition is provisional; its purpose is to illustrate the discussion of analog. In fact, I'm not sure that there is a meaningful general formula for denotative symbol systems, and I will shortly argue that it is probably a mistake, for AI purposes at least, to build on the notion of a rigid semantics.
several objects without some prior packaging or collecting into a set or some other complex object.

Continuum analog

I think there are two main interpretations of analog. The first and most popular interpretation takes analog to be a purely syntactic property of symbol systems: a symbol system is analog if there are no "gaps" in the symbol space, if it is a continuum. Let us call it the continuum sense of analog. One way of making this more precise is - in the manner of Goodman - to take it to mean that for every two symbols in the symbol space there is a third "between" them. "Between" may be taken as a primitive that satisfies certain conditions (if x is between y and z then y can't be between x and z, etc) or it may be defined in terms of some more primitive relation. A slightly different way to formalize the idea about "gaps" is to demand that there are other, distinct symbols, arbitrarily close to any given symbol, assuming that we have a metric for the symbol space.

Whatever our decision, it is clear that analog in the continuum sense is a relative concept. It is relative to a particular metric, or between-relation, or whatever we have chosen as our geometrical basis. For simplicity I will discuss continuum analog defined in terms of a metric; it has the convenience of being a well-known, mathematically precise, concept. Only given a specific metric can we decide if

25) It may seem ridiculous to use such a sophisticated and finely tuned mathematical apparatus for a subject of such indistinctness and uncertainty, but I do it just for the convenience of this particular discussion of analog. The distinction between the two senses of analog, doesn't depend crucially on the particulars of defining continuum analog. Generally, a definition in the vein of Goodman's definitions of differentiation would be more recommendable.
a symbol system is continuum analog. Digital, as it is normally understood, is the opposite of continuum analog. A scheme is digital according to Goodman's definition if it is syntactically differentiated throughout. The really interesting consequence, from the point of view of knowledge representation, is that a digital scheme allows of implementations that are not subject to decay of tokens: a slightly deviant "token" can always be restored to its proper state. In particular this means that "repeatability of readings" (Goodman 1968:161) and "flawless copying" (Haugeland 1981:213) are feasible. This, I believe, is why we use digital computers (and should continue to do so).

Analogy analog

The other interpretation of analog is related to analogy; I will call it analogy analog. This interpretation focuses on the relation between symbols and objects, the semantics of the symbol system. In the most general sense it is the idea of a structure-preserving mapping between symbols and objects; in a more specialized sense it denotes a continuous mapping between symbols and objects.27

A continuous mapping can be viewed as a mapping that preserves "closeness", which makes it a special case of a structure-preserving mapping. Clearly, this continuous mapping variant of analogy analog also relies on a specific variant, moreover, we need a

26) Or system – however Goodman demands of a digital system that it also be semantically differentiated, a condition which is of no interest for the present point.
27) Generally, the "mapping" is not really a function, so strictly speaking we cannot apply concepts such as continuity, but I think that more technical machinery would only obscure the real issue.

184 IV. PHILOSOPHY
metric for the object space as well. Notice that, since a continuous mapping need not have continuous range and domain, a symbol system may be analog in this sense without being analog in the continuum sense.

In the more general case of a structure-preserving mapping, obviously we now have to be specific about relations in general. Let us say then, that a particular relation o[j,n] is preserved if and only if there is a corresponding relation s[i,n] such that whenever s1, s2, ..., sn in the symbol space are related by s[i,n], and s1Ro1, s2Ro2, ..., snRon, then o1, o2, ..., on in the object space are related by o[j,n]. Thus, we can speak of the structure being partially or completely preserved, a mapping preserving more of the structure than another, etc. Of course, a structure preserving mapping is completely compatible with discrete symbol space and object space.

Analog and pictures

What, then, has analog to do with picturehood? Continuum analog, I think, very little. (Although this is, in effect, Goodman's basic criterion for picturehood.) If we take a broad view of pictorial representation, many pictorial systems are not continuum analog, among them - obviously - all representational systems that could exist in a digital computer.

That pictorial systems generally are analogy analog, on the other hand, seems to be widely accepted. Not all analogy analog representation would seem to qualify as pictures, however. Consider figure III:4 once more; if we take the symbol space to consist of all rectangles, then the "taller than" relation of the object space is clearly preserved by the "has a longer side than the shortest side of" relation, and
yet it seems difficult to accept this representation as pictorial. It's too arbitrary. What is lacking, I think, is intrinsicness. Compare the example in figure IV:1, where the symbol space is now restricted to squares.  

Entrenchment

But there is a problem with this whole discussion, namely that the question if something is analog (in either sense) seems to get too arbitrary an answer. For instance, take a paradigm case of a non-analog (in both senses) representation such as the ordinary way of representing rational numbers with Arabic rational numerals; it could be turned into an analog representation simply by introducing the right metric. Without restrictions on metrics permitted we could make any symbol system with a continuous object space continuum analog simply by adopting the metric induced from the object space. And then there would be no substance left in the concept of continuum analog. But how do we distinguish the "proper" metric from all the mathematically possible metrics?  

28) However, as will appear in chapter VI, I think an analysis in terms of analog, intrinsic, etc, only goes part of the way to a full understanding of what pictures fundamentally are.  

29) Cp the discussion above in the subsection on Goodman: "The difference".
Similarly with analogy analog - any symbol system could become analogy analog by just filling it up with relations, and the concept of analog would be just about vacuous. But, again, how can we distinguish between "proper" relations and "ad hoc" relations?

Notice how Hayes escapes this kind of problem in his definition of "direct" representation:

"each medium-defined relation used in constructing configurations . . ."

(Hayes 1974, my emphasis)

This is a constructive approach to symbol systems, not an analytic (which incidently reminds us - where do all the relations, s[i,n] and o[i,n], come from?). For several reasons, it is non-trivial to convert Hayes's constructive approach into a more generally useful, analytic, characterization:

i) The part-whole relation assumed by the constructive approach is problematic. Is a general characterization of subsymbol-ship possible? Also, being able to pick out subsymbols, doesn't necessarily imply that they can be neatly arranged in finite hierarchies and viewed as constructions. Subsymbols may intersect and multiply beyond control, etc. How is a photograph constructible from primitive symbols?

ii) It is quite conceivable that different constructions will yield the same symbol system.

iii) The constructive relations may not exhaust the set of relevant relations.

Now, the relations s[i,n] in the symbol space must certainly be computable from the symbols, meaning, among other things, that there must be effective procedures for individuating and identifying the
symbols; we can't allow that two computationally equivalent symbols map into different objects. This, rather obvious requirement (for a theory in AI), excludes many of the mathematical possibilities. But not enough, as seen by the example above: a metric that turns decimal point notation into an analog symbol system is perfectly computable, and it's not even a complex computation.

I will suggest a pragmatic solution. Symbols are not pure abstractions - the stuff they are made of, the medium, retains its properties even in the symbol system. The symbols come from a particular realm where they are typically characterized by a particular collection of properties and relations. More precisely, there is a prior symbol system. To take a very concrete example: using pebbles as symbols - terms for location, size, weight, and color, are part of an established system of labels used to characterize them.\(^{30}\) Some of these properties become part of the symbol space, building up its structure.

Consequently, "natural" man-made symbol systems are non-arbitrary (although conventional) as to what features and relations of the symbols that count, and a particular metric will be judged by its degree of entrenchment\(^{31}\) in that collection of "medium-properties", and by how well it competes with other metrics in terms of entrenchment. Similarly with structural relations of the symbol space in general. And similarly with the object space.

Thus, we are reluctant to take Arabic rational numerals as analog, for the simple reason that certain visual, "geometric", features of the notation

\(^{30}\) And there are other such well-established systems, e.g. from the point of view of mineralogy, of archaeology, etc.

\(^{31}\) See Goodman 1983.
are better entrenched in the realm of "marks on a paper", than are the relations that would have made it an analog system. (It's possible that a mathematician would take a different view.)

One might perhaps think that an "artificial" symbol system designed for use by a robot, isn't subject to such entrenchment considerations. Actually, I think knowledge representation faces entrenchment restrictions - in the guise of efficiency considerations - on two fronts. The symbol system has to be efficient relative to the underlying software or hardware - which can be viewed as a primitive, prior symbol system. Some data-structures, some algorithms, although in principle possible to implement, would become just too awkward and inefficient. Secondly, the representational system of a robot that is with us in this world, will ultimately have to take our own symbol systems into account (as I argue in the frame paper, part III: Metaphysics).

Some remarks on symbols, spaces and worlds

It is important not to confuse properties of symbols with properties of the symbol space. For example, that a particular symbol "mirrors" an object doesn't mean that the system is analog. Analog is either a relation among the symbols of the symbol space, a property of the symbol space (continuum analog), or a property of the general mapping, a relation between the two entire spaces (analogy analog) - not a relation between individual symbols and objects.

The main source of this kind of confusion is the notion of complex symbols.\textsuperscript{32} A complex symbol is, to

\textsuperscript{32} Notice that complex symbols were not given any special attention in the definition of a denotative symbol system above; a "complex" symbol is just a symbol that has various "super-symbol" relations to other symbols ("subsymbols"). I'm
be sure, structurally related to those other symbols in the same symbol space, that are subsymbols of it; and in many cases, preserving subsymbol relations goes a long way towards preserving relations in general. Thus the fact that a complex symbol "mirrors" its object will often indicate that the symbol system is— to some extent— analogy analog. But sometimes there may be other relations of interest. For example, between different photographs there is no general relation relating photographs of one and the same object: a front view need not have any syntactically identifiable relation to a back view of the same object. (Compare this with scale models.) That is, the identity relation is not preserved.

The kind of symbol space I have been talking about here is a very abstract and static thing, where all the possible symbols exist side by side in an arrangement determined by "closeness", "topology", and other structural properties (s[i,n]) of the medium, that are considered relevant. The concept of a symbol world— as in Palmer (1978)— lies closer to AI practice, and seems to offer certain advantages: (i) symbol worlds are more concrete; and (ii) the notion of manipulating symbols, (that is, transforming symbols into different symbols), ties in more naturally with the symbol world concept. Note, however, that the concept of symbol space is the more general: every symbol world generates a symbol space (its state space), but not every symbol space can easily be thought of as generated by a symbol world— for instance, what would be the symbol world underlying the space of ordinary photographs? The relation between the notions of symbol space and symbol world would be worth a closer study.

sceptic about the existence of a general and useful syntactic concept of "composition" of symbols.
V. ARTIFICIAL INTELLIGENCE - EXPERIMENTS

Although many AI-systems in one way or the other have pictorial elements, and although there are areas of computer science - such as computer vision and picture processing - entirely devoted to pictures, or pictorial input or output, there are surprisingly few experimental studies explicitly designed to explore and exploit the possibilities offered by pictorial representation. Here, I will discuss what I think are the two most important contributions in this area, namely the geometry machine of Gelernter et al, and Brian Funt's WHISPER.¹

The geometry machine

An early attempt to exploit diagrams in mechanical problem-solving was the geometry machine of Gelernter et al (Gelernter & Rochester 1958, Gelernter 1963, Gelernter & Hansen & Loveland 1963). The geometry machine could construct proofs of theorems in plane geometry from a set of axioms on parallel lines, congruence, and equality - axioms of the type found in ordinary textbooks on elementary geometry. The main object was to study the use of heuristic methods in problem solving.

For the present purpose, the interesting part of the geometry machine is the so-called "diagram computer" and how it is exploited in the proof-search process.

The diagram computer holds a diagram illustrating the current premises (hence, hopefully, also the current theorem to be established), stored as a coordinate representation, but accessed and manipulated as a picture:

"although the procedures of analytic geometry are used to generate the description, the only information transmitted to the heuristic computer ... is of the form: 'Segment AB appears to be equal to segment CD in the diagram,' or 'Triangle ABC does not contain a right angle in the diagram.' The behavior of the system would not be changed if the diagram computer were replaced by a device that could draw figures on paper and scan them." (Gelernter 1963:139, my emphasis)

That is, the diagram computer provides a virtual picture of a particular triangle, which happens to be implemented verbally, but (barring practical problems) could have been implemented as a real picture. For example, checking that P1, P2 and P3 are collinear is done by evaluating and comparing \((y_3 - y_1)/(x_3 - x_1)\) and \((y_2 - y_1)/(x_2 - x_1)\), where \((x_1, y_1), (x_2, y_2), (x_3, y_3)\) are the number pairs representing P1, P2 and P3, respectively. If the difference is reasonably close to zero, the answer is "yes", otherwise "no".2

How can a diagram be used in the search for a proof? The main point (but not the only point) is that anything which is false in the diagram is impossible to prove from the premises. This follows from the

2) It's interesting to note that Pylyshyn (1973) rejects the thought that the diagrams in the geometry machine are pictures on the ground that their physical nature is irrelevant, only their status as models. Again this confusion of pictorial with physical, inherited from the copy theory of depiction.
To give an idea of how this works in practice, I will use and modify an example from Elcock (1977), which although not an exact description of how the geometry machine solves the problem, demonstrates clearly the role of the diagram.

Given that:
1) ABC is a triangle
2) M is the mid-point of segment BC
3) BD is the perpendicular from B to AM
4) CE is the perpendicular from C to AM
5) D, M and E are distinct

Prove that: segment BD = segment CE

The following diagram is constructed:

Among the axioms, there is one that says that:

Al: If two triangles are congruent, then their corresponding sides are equal.

This could be used to prove that BD = CE, given that we can prove that BD and CE are corresponding sides.

3) See Gelernter 1963, for the authentic protocol of this problem.
of two congruent triangles. So, we should set up a subgoal to that effect. There are several possibilities:

a) triangle BDM is congruent with triangle CEA
b) triangle BDM is congruent with triangle CEM
c) triangle BDA is congruent with triangle CEA
d) triangle BDA is congruent with triangle CEM

Checking with the diagram we find that a), c) and d) can be pruned since they are obviously false in the diagram, leaving subgoal b) as the only remaining possibility. But before that, we should make certain that BDM and CEM really are triangles. (The points could be collinear, or some segment might be amiss.) It could of course be proved from the premises (3 and 5) but we may also take it as "evident" from the diagram. Note that this use of the diagram has a different character (and is not as unimpeachable).

The careful reader will have noticed that not all the syntactically different possibilities were enumerated. Triangle BDM, and triangle DMB, for example, are two syntactically different representations for the same triangle (in our semiformal notation), syntactically they satisfy the equality relation (=). Certainly, the diagram would help to handle also those additional possibilities, but in the actual geometry machine, syntactic symmetries were dealt with by a separate, completely syntactically based mechanism. This mechanism was also used for discovering and exploiting semantically relevant syntactic symmetries. For example, if parallel goals were syntactically symmetric (even if semantically distinct), establishing one of them would be taken as establishing the rest of them by symmetry. For instance, if we have a proof that M is the mid-point of AC (see figure V:2)
- then by symmetry there's also a proof that \( M \) is the midpoint of \( BD \). (Doing this by semantical means, using a diagram, would amount to finding transformations of a certain class that map the diagram onto itself.)

Continuing with the proof, the current subgoal is now:

triangle \( BDM \) is congruent with triangle \( CEM \).

The following axiom appears useful in proving this goal:

A2: If triangle \( XYZ \) and triangle \( RST \) are such that segment \( XZ = \) segment \( RT \) and angle \( XYZ = \) angle \( RST \) and angle \( XZY = \) angle \( RTS \), then the triangles are congruent.

But how should it be instantiated? There are several possibilities of matching the two triangles with each other:

a) \( BDM \) with \( CEM \) (in that order, that is, \( BD \) corresponds to \( CE \), \( DM \) to \( EM \), and \( MB \) to \( MC \))

b) \( BDM \) with \( CME \)
c) \( BDM \) with \( MCE \)
d) \( BDM \) with \( MEC \)
e) \( BDM \) with \( ECM \)
f) \( BDM \) with \( EMC \)

and for each of these possibilities, there are 3 different ways of instantiating A2. For example, a) can be instantiated as:
Again, a look at the diagram tells us that a1) is an appropriate choice - most of the other subgoals can be pruned by using the diagram, and the remaining are recognized as leading to circularities. It seems wasteful to generate all those subgoals and then start pruning. A quicker way of finding a subgoal that is likely to succeed (which normally will be the only attainable subgoal), would be to pick one of the segments (say BD) in the triangle BDM and find a (normally the) segment among CM, ME, and EC, in the diagram that it is (almost) equal to. Then matching an adjacent angle (angle BDM, say) with the two angles adjacent to the matching segment would pick out a plausible instantiation, if there is one. (Normally, it would be the only possible.) By now, we have three new subgoals:

1) segment BD = segment CE
2) angle BDM = angle CEM
3) angle BMD = angle CME

The first subgoal is a premise. The second can be proved with

A3: Right angles are equal.

That angle BDM and angle CEM are right angles is given in the premises. The third subgoal can be proved with the axiom

A4: Two vertically opposite angles are equal.

given that D, E and M are collinear, and that B, M and C are collinear, which is evident from the diagram. And that completes the proof.
The real geometry machine didn't actually construct the diagram, but was given it as an additional input. Plans to make it automatic were never realized, although one obviously felt confident that this was just a minor technical exercise. The diagram capabilities were however extended in a different way, by introducing a limited form of constructions: should the current search space become exhausted, the geometry machine could introduce new points by connecting two previously unconnected points by a
segment and constructing its intersections with existing segments. 4

Uses of the diagram

The above example demonstrates three different ways of using the diagram:

i) to eliminate possibilities (negative heuristics);

ii) to suggest plausibilities (positive heuristics);

and

iii) to provide evidence.

By far the most important role of the diagram in the geometry machine, was to eliminate goals, pruning the search tree. Obviously a very powerful negative heuristics: it was estimated that over 99% of the subgoals were eliminated with the help of the diagram (plus a circularity discovering mechanism). I have found no clear indication that the real geometry machine employed the diagram also as a positive heuristic for attacking alternative subgoals in an order according to their plausibility. Positive heuristics seems to have been based solely on syntactical methods, such as measuring some sort of "distance" between goal strings and the set of premise strings. In the example above I suggested how one could use the diagram to direct the attention to the most plausible instantiation of A2, which is one

4) To be exact, there was, even before the introduction of this construction capability, a possibility to introduce new line segments between unconnected points - but without constructing their intersections. In effect, the difference between an existing segment and a segment "constructed" in this manner, only served as a weak heuristic aid, helping to concentrate on segments "mentioned" already in the premises, before beginning to explore other segments.
example of how a diagram could serve as a positive heuristics.

Let us reflect a little on the different characters of negative and positive heuristics. First some general properties. A negative heuristics decides if a given possibility should be rejected or not. A positive heuristics picks out, or generates, a plausible (or the most plausible) alternative from a range of possibilities. A negative heuristics is (usually taken to be) irrevocable: a possibility ruled out stays out. Mistakes can't be afforded, leading to a cautious, conservative attitude. A positive heuristics, by contrast, is tentative: alternatives not chosen are not ruled out forever. Since no permanent damage is done by an occasional error, there's room for bold and brilliant, but perhaps badly supported and possibly mistaken, suggestions. On the whole, guiding by rejecting is very different from guiding by directing; success in "Twenty questions" depends very much on good questions.

In the case at hand here: when the picture is used as negative heuristics, it is the verbal (syntactic) apparatus that generates the different alternatives - the picture vetos suggestions it doesn't like. When the picture is used as positive heuristics, it is the picture that is the source of inspiration.

Elcock (1977:27) remarks that

"we must simply leave as an open question whether exploration of truth in the model should always be in response to goals in a developing syntactic proof or whether some prior exploration of truth in the model might suggest appropriate sets of syntactic possibilities and facilitate their ordering for detailed consideration."

- The geometry machine 199
I believe diagrams have quite as important a role providing positive heuristics as of giving negative heuristics. A picture can be used to expose unfaithful descriptions, as well as to inspire to the formulation of adequate descriptions. And it seems likely that suggestions based on verbal (syntactic) considerations are generally different from, and complementary to, suggestions based on pictorial (diagrammatical) considerations.

The third use of the diagram — as evidence — was frequent in the geometry machine for the kind of details usually not given a formal treatment in elementary textbooks. For example, it can be used to verify that a particular triple of points actually defines a triangle, that one point lies between two others, that two points are on opposite sides of a certain line, and for other betweenness relations.

A particularly tedious detail to handle by formal means is the identification of different names for the same angle. For example, in the diagram in figure V:4, there are (excluding symmetries) four different ways of representing each base angle (angles $\angle ACE$, $\angle ACD$, $\angle BCE$, $\angle BCD$, and $\angle CEA$, $\angle CEF$, $\angle DEA$, $\angle DEF$).

Consequently there are 16 different ways of representing their possible equality. With the diagram it is easy to check if two syntactically different expressions refer to the same angle.
Clearly much effort can be spared by using the diagram as evidence. Indeed, Gelernter & Rochester believes that

"traditional geometry excludes proofs of betweenness, and most mathematics appears to lack rigor because many matters are settled by heuristic methods rather than formal proofs. It seems clear that the machine must be able to work this way if it is to become proficient." (1958:341, my emphasis)

Reliability and efficiency

How reliable are these methods? There are two sources of error:

1) the diagram is not general
2) the diagram is not exact

Care was taken that the diagram input to the geometry machine should "reflect the greatest possible generality" (Gelernter 1963:141). A description can be general or not general; the nearest an object (such as the diagram) can "approximate" generality would seem to be by being "typical". So probably, what Gelernter means is that the diagram should be "typical": it should have as few "special" properties as possible (in the total set of diagrams satisfying the premises). That is, intersection points should not coincide, angles not be right angles, segments not be of equal length, etc, if it could at all be avoided.

Nevertheless, theorems (the interesting ones) are universal, diagrams particular. Not everything true of the diagram, not every truth expressible in the language of the theory, follows from the premises. We must not confuse "typical" with "general". A theorem about triangles is valid not only for a
"typical" triangle, but for all triangles.

Incidently, what is typical? Intuitively, the typical triangle is neither right-angled nor isosceles. The simple idea that to be typical is to be in majority, meets with immediate difficulties; there are, for example, as many non-right-angled triangles as right-angled, as many non-isosceles as isosceles triangles. Perhaps it is more reasonable to say that a property is special (non-typical) if it occurs (in the right way) in some axiom; right-angled triangles are special just because we have some special axioms about right angles. But that doesn't explain why isosceles triangles are special. Then, substitute "theorem" for "axiom" in the above. But surely there are as many theorems about non-isosceles triangles as about isosceles triangles. Substituting "interesting theorem" for "theorem", would only get us in deeper water. Maybe there still is a way (though I doubt it) to straighten out our intuitive notion of "typical" (in the special sense we are talking of here); unfortunately, in many cases we don't even have an intuition. What is more typical - a convex or a concave quadrilateral?

None of this affects the reliability of using the diagram for pruning the search tree. If something is false in any diagram that satisfies the premises, we can safely exclude that possibility. Efficiency, however, suffers - not every goal that should be pruned (because false in some diagram) will get pruned.

Similarly, using the diagram as positive heuristics doesn't depend on the generality of the diagram for its reliability (since positive heuristics is tentative, it isn't committed to anything) only for its efficiency.
Using the diagram as evidence is different. Because then we're assuming that a property of the particular diagram at hand is shared by all diagrams satisfying the premises. Things can go wrong. An interesting feature of the geometry machine was that it made a record of every assumption based on diagrammatical evidence - in effect extending the set of premises. The risk of mistakes might also be alleviated by

"having the machine draw alternate diagrams to test the generality of its assumptions"
(Gelernter & Rochester 1958:341)

This important stratagem, never realized in the geometry machine, would as well enhance the efficiency of pruning, and the precision of positive heuristics based on the diagrams.

The other general problem, is that of exactness. Above, we have reasoned on the assumption that the diagram is exact, that everything that follows from the premises is true of the diagram. Generally, this will not be the case. The diagram is represented using finite precision arithmetic, which means, for example, that tests for equality must have a certain leeway to make equal things come out equal in the diagram, hence that things may appear equal which really aren't. On the other hand, things will occasionally be deemed unequal which really are equal, especially if the machine had unlimited capacity of constructing new points and lines, meaning that small uncertainties would inevitably multiply to big uncertainties. Because of this inexactness, the geometry machine did, now and then, prune a true subgoal, but luckily always managed to find an alternative (possibly longer) proof. In a less redundantly axiomatized theory, one may not always be that lucky.

Positive heuristics based on the diagram is not seriously affected by inexactness. But the method of
using the diagram as evidence, is clearly vulnerable to bad precision, although intuition tells us that a diagram is much less likely to be mistaken for a more "special" than for a more "typical" diagram, in which case no serious harm is done.

On the whole, I agree with Gelernter & Rochester, when they say

"To serve as a heuristic device in problem solving, the model need not lie in rigorous one-to-one correspondence with the abstract system. It is only necessary that they correspond in a sufficient number of ways to be useful." (1958:338)

In particular, this would seem to apply when we approach problems in less highly formalized systems with methods similar to those used in the geometry machine.

WHISPER

WHISPER (Funt 1976, 1977, 1980) is a problem-solving system that analyzes and makes qualitative predictions of what will happen to unstable structures as they collapse, in a two-dimensional "blocks world". The most prominent feature of this system is its use of pictorial representation techniques to simplify problem-solving. There is a diagram depicting various stages of the collapsing structure, and a retina that "looks" at this diagram.

With a diagram, representing the initial configuration of (arbitrarily shaped) blocks, as input, a series of diagrams, snapshots of the collapsing structure, are produced as output. Each new diagram is constructed with the help of WHISPER's general knowledge about
the stability and motion of falling, rotating and sliding objects, and based entirely on the information contained in the previous diagram. Thus, the basic problem-solving process simply iterates over the diagrams.

One cycle of problem-solving (simplified by excluding sliding problems) consists of the following phases:

1) All instabilities are determined. In this process, the retina is used to locate objects, determine their centers of gravity, and find their support relationships.

2) The dominant instability is chosen.

3) By looking at the diagram the pivot point for rotation of the chosen object is determined; it will be the support point closest to the object's center of gravity.

4) By a process of "visualization" the image of the object is rotated in the retina until it is seen to collide with some other object or begin to fall freely.

5) The diagram is redrawn accordingly, to produce a new output and starting-point for the next cycle of problem-solving.

The retina consists of "bubbles", receptors arranged in rings (concentric circles) and wedges (rays from the center). Each bubble covers a small portion of

- WHISPER
the diagram and is able to detect the "color" of that region. (Different objects or shapes are individuated by their unique "colors".) The retina can fixate any point in the diagram. Notice that the resolution declines towards the periphery; a design dictated by "economy of receptors and processors" (1980:205). Each receptor has its own processor, connected to its nearest four neighbors and to a supervisory processor. These processors work in parallel so that fixation and other operations of the retina are fast.

The top-level

The problem-solver's interface to the retina, the questions that can be put to the "perceptual apparatus", the manipulations that can be performed on the image in the retina – perceptual primitives, Funt calls them – define a certain virtual representation. A brief survey of some important primitives will give us an idea of what this representation is like.

The center-of-area primitive determines the center of area of a given shape in the visual field of the
retina. (On the assumption that the objects are homogeneous, it can be used to find the center of gravity of an object.) Knowing its own size and position relative to the retinal center, each bubble can compute its own contribution to the total area vector, and all the retinal supervisor has to do is to sum up the incoming contributions and divide the sum by the total area. If the computed center is far off the center of the retina, in the low-resolution area, the accuracy can be improved by fixating the approximate center of area and repeating the calculations.

The other primitives also take advantage of the parallel processing facilities. For example, the retina-visualization primitive rotates a given shape about the center of the retina, step by step. The diagram is not involved in this process. In each step, any bubble marked by the color of the shape to be rotated, passes on that color to its ring neighbor (in the direction of rotation) and then erases its own color marking (and attends to any incoming message that may have arrived). A collision can be detected by a bubble receiving a mark message when already marked by some other color.

The contact-finder primitive finds an object's points of contact with other objects. (Important in finding the supporters of an object.) Find-nearest finds the bubble closest to the retinal center that satisfies a given condition. (Useful, for example, to find pivot points.) The ring structure greatly facilitates this task.

The symmetry primitive tests for reflective symmetry about a given axis by shifting markings wedge-wise; clockwise on one side of the axis, counterclockwise on the other. If the shape is symmetrical, each pair of marking messages that meet along one half of the axis will be equal. The retina-scaling primitive
changes the scale of a given shape (without affecting the diagram), by shifting ring-wise one or several steps outwards (enlarging) or inwards (reducing).

The similarity primitive combines translation, rotation and scaling operations in the retina, to find if two objects are similar in shape. The curve-features primitive is used to trace and analyse contour curves. Once the bubbles on the curve have been sifted out, they can test for different curve features in parallel. For example, a bubble can detect a sharp bend by comparing the number of neighbors on each side of the curve, that do not themselves belong to the curve.

To sum up: the top-level representation of WHISPER gives a distinctly pictorial impression with its access-methods and operations that are naturally associated with pictures in a rather narrow sense. But, of course, it is only as the problem-solver interprets this as representing various configuration of blocks that it really becomes a pictorial representation.

The middle- and bottom-levels

The middle-level representation is the (virtual) diagram, defined by the fixation procedure that maps a portion of the diagram onto the retina, and by the redrawing operations used to update the diagram according to what the problem-solver has decided on the basis of observations and "visualizations" made through the retina. These operations include translations and rotations of given shapes. Even if the repertoire of methods for accessing and manipulating the diagram is frugal, it has an unmistakably pictorial flavor. Once again, of course, it becomes a pictorial representation of block configurations only by being thus interpreted.
In principle, this could be where it stops - the virtual diagram being a real diagram - but in fact it doesn't. WHISPER's diagram is implemented as a dot matrix (plus, of course, the required procedures to handle the operations on it). Funt also discusses an implementation similar to a type common in computer graphics applications: the picture is stored as a list of elementary picture elements - usually line segments (pairs of coordinates). A draw-back of this approach is that the time to fixate a new point of the diagram will not be constant but a linear function of the number of picture elements.

Uses of retina and diagram

Funt describes the theoretical principles behind WHISPER's use of pictorial representations as follows. The standard approach to representation in AI is to formalize the problem domain. A theory $T$ represents a problem $M$ by having it as a model. $T$ exists in the problem-solving system in the form of statements and a deductive mechanism. The non-standard approach exemplified in WHISPER is to represent $M$ by a second, auxiliar model $M'$ of $T$. $M'$ is used

"to provide information about $M$ without deriving it from $T$. What is required is that some of the predicates, functions and individuals of $M'$ correspond to some of the predicates, functions and individuals of $M$ in such a way that it is possible to translate the results obtained when these predicates and functions are applied to individuals in $M'$ into the results that would be obtained if the corresponding predicates and functions were to be applied to the corresponding individuals in $M$. The similarity between $M$ and $M'$ means that experiments and observations made in $M'$ yield results similar to those that
Funt assigns the diagram and the redrawing procedures, the role of M'. What is modeled at this level is primarily elementary geometrical properties of the blocks world. For instance, that blocks are rigid, non-overlapping things that retain their shape, center of gravity, etc, under rotation and translation. As we step up to the top level more of the kinematical properties get to be appropriately modeled; for example, only at this level are collisions modeled. Hence, it would seem even more appropriate to take the diagram as viewed by the retina - what I've called the "top-level representation" above - as defining the model.

The fundamental theoretical basis is the same in WHISPER and the geometry machine; namely the notions of model and theory, and their relation. In both cases there's a general theme of cooperation between theory and model, between description and depiction, but there are important differences. No doubt, most of these variations can be ascribed to the fact that the geometry machine was designed to handle generalities, whereas WHISPER was designed to handle particular, individual situations, and the fact that the end products of the geometry machine are verbal, whereas the end products of WHISPER are pictorial. Nevertheless the differences have a more general interest.

Although there are parts of the modelling in the geometry machine that lack a verbal underpinning within the system - as in the case of betweenness relations - these are exceptions. Almost all of T exists within the geometry machine side by side with M', and almost all results obtained with the help of M' are independently checked in T. In WHISPER, the picture more often supplants a formal, descriptive, representation. Large parts of T fall outside the
scope of the system's formal capacities. If two objects "look" to be in contact, WHISPER will take that as a fact and not try to verify it by formal-verbal means. In fact, it couldn't, since the shape of an object - as well as most other information about individual objects - is only represented (and representable) in pictorial form. Similarly, there is no theoretical account within WHISPER of what happens to blocks as they are rotated, for example that they retain their shape. Etc.

Thus, in WHISPER, the "cooperation" between verbal and pictorial modes of representation consists mainly in a division of responsibility. Some parts of the total representational task are taken care of by verbal means, some parts by pictorial means. By contrast, in the geometry machine, verbal and pictorial modes of representations cooperate in resolving common issues.

In a sense, WHISPER's pictures are somewhat less pictorial, less symbolic, than the geometry machine's, and this is precisely because some of what goes on in the picture (relevant to the problem) cannot be internally described. To see this more clearly, consider WHISPER-2 which is an imaginary variant of WHISPER that is all diagram and no description, no theory: it consists simply of a little scale model of the set-up plus a camera that takes snapshots of the model as it collapses. No doubt, we might find the model an excellent picture of the real (or imagined) problem. But it is hardly meaningful to claim that it is a picture to the system of which this "representation" is a part, WHISPER-2; there doesn't even seem to be any "system" beyond the "representation" itself.

Yet, there are some interesting examples of interplay between the parts of T represented in WHISPER, and M'
about the same issues. Experimental feedback, Funt calls it. For example, step 4 in the basic problem-solving cycle goes in detail like this: in the retina, the object is visualized as rotating until a collision occurs, but due to the low resolution of the retina outside the central region, this result is only approximate. To improve it, a measure—a description—of the angle of rotation of the visualization (rounded off downwards so that the rotation will not overshoot), is used to redraw the diagram. There will now probably be a slight gap (in the modified diagram) at the estimated point of contact. The retina is fixated on that point to get a good measure of the size of the gap, and the verbal information is subsequently used to make a final redrawing of the diagram, closing the gap.

The way WHISPER handles sliding problems, gives another illustration of the "experimental feedback" technique. It is not possible to visualize an object sliding along an arbitrary curve. Instead, WHISPER examines the curve itself. We will skip the details of this examination and proceed from the moment when the point where the slide will end has been established. The new position of the object is found by the following method. First, the object is translated by calling the appropriate redrawing procedure for the diagram. The object will now probably not be correctly oriented. It seems a reasonable approximation to assume that the object should have the same points in contact with the surface it is sliding on, at the end of the slide, as it had at the outset. So, the second step is to visualize the rotation of the object, in its new position, watching for these contact relationships to be re-established. The angle is measured, the diagram redrawn accordingly, and the retina moved to examine any remaining gap. If there is one, a final corrective redrawing of the diagram is made.
By the employment of a pictorial representation, WHISPER manages to solve, elegantly and efficiently, many problems that are very difficult to handle efficiently and reliably by verbal, formal, methods. One example is the problem of detecting collisions and other abrupt changes in the motion of objects. Formal equations of motion do not provide for any easy way of determining the points where boundary conditions change. And it would be difficult to devise a verbally based heuristics that wouldn't overlook cases such as the impending collision in the following example (from Funt 1980):

Another example of a task where WHISPER's pictorial capacities outperforms verbal methods is the type of curve following needed to solve sliding problems. The crux is that a smooth curve may be formed by the accidental alignment of two or several distinct objects. In a system that relied on separate descriptions of each object, this would cause considerable trouble:

"To find it [smooth curve formed by accidental alignment] would first of all require the built in expectation that it might happen. Then its existence would have to be continually checked. This check would involve establishing all the contact relationships of the object on which the sliding object is initially resting with all the other objects in the universe, already a difficult problem, followed by the amalgamation of the descriptions of the two separate curve descriptions into a new curve description."
WHISPER just sees it.

One of the points made in Funt (1980) is that the pictorial techniques used in WHISPER enables one to make do with fairly simple theories - qualitative reasoning.

"What we can see from all this is how experimental feedback combined with a first order theory of sliding motions results in a very natural form of qualitative reasoning." (1980:222)

That is, the full, detailed, theory may be beyond what the system can explicitly describe, but given that the auxiliary model M' is a model of that theory just the same, a cheaper and simpler theory can be successfully employed within the system as long as it is being corrected by M'.

(1976:82)
VI. PICTURES AS MODELS – OUTLINE OF A THEORY

Up to this point I have reported on and discussed a wide variety of views on pictorial representation; now, I want to bring all this together by outlining a new theory. Some desiderata for a theory of pictorial representation suitable for application in knowledge representation have emerged in the foregoing. The main point, of course, is that it must not be a theory of representation so unworldly as to resist interpretation in terms of computer implementation. Preferably it should be a theory which supports such interpretations, and which is rich in suggestions of how pictures can be cognitively exploited. Representations of hypothetical objects and situations should not be subject to any special treatment in the theory. The theory should have a view of pictures wide enough to include the sorts of pictorial representation we can expect to have in computers; if possible – it is at least worth an attempt – wide enough to form one half of a dichotomy with verbal representation, understood in its broadest sense, as the other half. Also, it would be desirable to have a theory that sets verbal and pictorial representation in relation to each other.

The key idea in my proposal is that pictures are examples or models; models that comply with a label, a description, a theory, or any kind of verbal, non-pictorial, representation. Thus, contrary to Goodman's thesis that denotation is the core of pictorial representation, I propose to build a theory of pictorial representation on a modified version of Goodman's exemplification concept. Examples are models in a sense that can be thought of as an
The extension of the model concept of logic. The specialized concept of a standard-model or intended model is also borrowed from logic and plays a role that parallels the distinction between "literal" and "metaphorical" reference. That X is a picture of Y, can then roughly be analyzed as X being a model for some description that has Y as its standard model.

The sense of model intended, differs from the ordinary logical concept in that modelling is not restricted to sentences, but extends to a wider range of verbal categories, including names and terms. Thus, for example, a blue elephant is not only a model for "the elephant is blue" - but also for "blue" and "elephant". This is mostly a matter of terminology, I think, but also suggests a less prominent role for the concept of truth.

A perhaps more consequential difference is the extension of the model concept to non-formalized discourse. Although it means some loss of precision and clarity, it is a natural step to take, and indeed necessary to get a vantage-point from which verbal as well as pictorial types of representations of a wide variety can be covered.

I will make a distinction between genuine pictures, and quasi-pictures. A representational system in operation may use symbols that to the outside observer appear to be pictures of what they represent, and yet the symbols are actually used in no way different from how non-pictorial symbols would be; such "pictures" I will call quasi-pictures. To function as a genuine picture, requires that the picture is being used pictorially, that is, as an example, as a model. This means, for example, that a picture can act as a source for generating verbal representations (labels, descriptions, theories) that apply to it. Such verbal offspring from a picture
can be tested for applicability to a situation in the
world, or to a different picture; it can be processed
by some verbal inference engine to produce new verbal
representations, etc. There is a multitude of possible
ways of explicitly using pictorial symbols, which
have in common the picture's indirect application.
Contrary to a common assumption that pictures are a
more direct type of representation than words, my
point is that genuine pictorial reference is always
mediated by non-pictorial symbols.

Generally, pictures stimulate the search for, or
creation of, descriptions that apply to them. An
important point here is that a single picture can
normally be verbally represented in multiple, even
indefinitely many and varied, ways; highlighting dif­
ferent aspects, being abstract or concrete, using
different languages of description (which may not be
reducible to each other), etc. Often, it is very
much left open which description to apply: pictures
are open-ended in the sense that it isn't usually
given what the description of a picture is, rather
the picture implies a whole spectrum of descriptions,
with rather diffuse boundaries.

By contrast, verbal representations, stimulate creat­
ing, or searching for, models for them. Applying a
description implies trying to find a model for it.
The creation and modification of internal models –
that is, pictures – for verbal representations have
several interesting possible uses. For example,
models can be used to calibrate the meaning of
descriptions; to change perspective, translating
between different ways of describing something; and to
support analogy reasoning.

Verbal and pictorial representations, each have their
particular advantages and disadvantages. Their dif­
ferent natures are seen particularly clearly, when we
let them interplay with each other. More than a point by point examination and comparison this brings out a fundamental duality between words and pictures, and an important interplay between describing and depicting.

One symptom of the fundamentally different characters of words and pictures, the fundamentally different direction of verbal (converging, contracting) and pictorial representation (diverging, expanding), can be observed in the process of adding to a representation. Adding to a description in a way that augments its information content, generally means that there will be fewer models left, to which it applies. Whereas, adding to a picture so that its information content is augmented, will generally allow for additional applicable descriptions. A description very rich in information applies to very few models, conversely a depiction rich in information satisfies many descriptions. Generally, a model has more in it than the description that directed its generation or discovery prescribes. The model yields a surplus of applicable descriptions.

Of course, this needs some elaboration. What follows is a preliminary first attempt; it is to be understood as a report of work in progress.

Application

Truth and denotation

I will speak not so much in terms of "denote" and "denotation"1 as in terms of "apply" and

1) I am now thinking of these terms in their "ordinary" sense in philosophy, not in Goodman's broad sense.
"application"; "truth" I will also avoid, and for similar reasons. Truth and denotation may be fine for doing philosophy - for AI those concepts are generally too blunt, they make distinctions which AI has little use for, and they exemplify a bodiless, eternal, objective, omniscient, and omnipotent, perspective, that scorns the mundane representational systems within the scope of AI, subject to all sorts of restraints: materiality, locality, computability, efficiency, etc. This is not the place for a very careful and elaborate argument along these lines - you may take it simply as the down-to-earth practical systems designer's point of view. Here I will just briefly indicate some motives.

First, the question of making the right distinctions. Truth is in a sense a very specialized concept, that applies only to propositions. But, straightforward logical representations excluded, there are in applied AI generally no very sharply pronounced divisions between propositions, questions, attributes, descriptions, etc. What I mean is that, for instance, representations corresponding to "cat is black", "a black cat" and "black and feline" are handled in rather similar manners (within the same system). Generally, a representation is not "true" or "false", rather it is "right" or "wrong", "good" or "bad", it "fits" or "doesn't fit". The ordinary concept of truth simply does not emerge very naturally from AI's concern with modelling. Denotation, on the other hand, is closer to what is needed and what is used in practice, but in the usual Fregean interpretation stops short of providing a useful analysis of sentences. And of course, the whole idea of truth gives a bias to verbal representations, since there are no clear and generally accepted criteria for what _____________2) In Frege's analysis, sentences denote truth-values.___________
it would mean for a model or a picture to be "true".

Truth is an all-or-nothing affair, it is absolute and unqualified. Similarly, with denotation - a certain thing is denoted or not, and there is a definite extension associated with each term, whether we know it or not. In practice, however, it is impossible to get on without the idea of a statement being "more or less true", or something being denoted "more or less" by a term, in a sense which - note - has nothing to do with uncertainty or lack of knowledge (or "fuzziness"). If "circle" could be applied only to true circles, we could never use it on anything in the world; yet we do it, and a competent speaker might very well apply it to something blatantly non-circular. In one situation it may be right to call something a "circle"; in a different situation it may be wrong to call the same thing a "circle". Speaking of "truth" and "denotation" in this way would seem to be a complete distortion.

Moreover, stressing the notion that something is true, tends to draw attention away from how it is true. Let me explain with a few examples. To ask how "Johnny is tall" is true, is to ask who this "Johnny" is, in what way "tall" applies to him, etc. (In a room full of Johnnies of varying age, some of which are pygmies, the answers are not without interest.) Similarly, that "there is something on the table" is true, will not tell us if there are one or several things, nor what those things are. Equivalently, the truth-value "false" often effectively hides information of how something failed to be true; suppose it is false that "this vase is red" - if it isn't red, then what color is it?

Yet, these questions about the mechanism, so to speak, behind something being the case (or not), are often settled in the process of computing the truth-
value. It is interesting to note that even pure theorem-provers will provide at least partial answers in some circumstances, for example, proving existence by finding a particular instance.

Underlying this, of course, is the tacit assumption that a term always denotes in the same way, and a sentence is always true in the same way, so that the "how"-questions have trivial answers. Such sentences and terms may exist (I don't know), but they are rare in natural language. As for logic, this assumption means that there would be at any moment just a single interpretation, which can change only by extension, leading to an enormous profusion of symbols. Also, it seems that much of the power of formalism is lost if interpretations are fixed. From a practical point of view, the value of arithmetic theory lies in that it applies not only to numbers, but to apples and cows and all sorts of things.

Some very difficult problems to handle with this kind of fixed semantics are: i) indexicals, cameleont symbols like "he", "now", "this", which change reference with the situation; ii) discriminative use of descriptions, which let us classify a single piece of cloth in one situation as "blue", when there's a choice between "blue" and "red"; and in another situation classify it as "green" and "not blue" in a choice between "blue" and "green"; and iii) metaphorical use of descriptions. 3

Actually these ideas are consonant with Goodman's concept of denotation, and comes quite close to his "rightness of fit":

3) And I strongly suspect that every use of language is more or less discriminative, more or less metaphorical, and that all words are more or less indexical - and moreover that far from being something to deplore, it is a prerequisite for an efficient and economical representation.

- Application 221
"Rather than attempting to subsume descriptive and representational [pictorial] rightness under truth, we shall do better, I think, to subsume truth along with these under the general notion of rightness of fit." (Goodman 1978:132)

Last, let me explain that even if I don't think that truth and denotation are the right concepts for a theory of representational systems for robots, I do not deny that they may be handy in the actual construction of a robot with a world-view similar to our own.

Apply

It seems to me that most theories of representation (reference) in philosophy are guided by a string metaphor: symbols and denotata are tied together by little strings. Although these strings are more or less undestructible and eternal, they are remarkably intractable. One cannot start pulling in one end of the string to find out what is at the other end. They are useless for that. Of course, it's reassuring to know that the little strings exist, but they are of little practical help.

I propose as a complementary metaphor more suitable for theories of representation for AI, the following keys-and-locks metaphor: symbol and denotatum are related to each other in the way a key is related to a lock that it fits. To find out if a certain key fits a certain lock, we try to apply it to the lock; this is the fundamental operation. There is no invisible link between key and lock. Given a lock it is always possible to fabricate a key that will fit. Given a key it is often possible to fabricate a lock that will fit. There are certain standard locks which a key was made to fit, but it will usually fit many others as well. Two keys that fit the same lock
represent different perspectives or different degrees of abstractions of something. Sometimes a key is calibrated by adjusting it to certain lock or locks, internal or external. Studying a lock that fits a certain key may be helpful in refining the key to fit better a certain class of locks, even when the lock studied doesn't itself belong to that class. Etc.

I think that **apply** and **applicability**, understood with this metaphor as a background, are notions well suited to the task of modelling in AI. These terms accord with many ordinary uses of "truth" and "denotation". If a name or term denotes something, then it applies to it; if a sentence is true, then it applies to the situation described by the sentence, roughly along the lines of Barwise and Perry's semantics for sentences (Barwise & Perry 1983). I will sometimes use Goodman's term **comply** as the converse of **apply**.

Applicability is a matter of degree. A rough, qualitative, classification of instances of application will be useful. I will say that a verbal representation applies **positively**, if indications for its applicability outweighs any indication that it doesn't apply. Similarly, a verbal representation applies **neutrally**, or **forcedly** when positive indications balances, respectively is outweighed by, negative indications. Seeing a car I recognize as being a Volvo, I can apply "Volvo" positively; but if the car is far away and I can't tell whether it is or isn't a Volvo, applying "Volvo" to it will be in the neutral sense; finally if I recognize the car as not being a Volvo and yet labels it "Volvo" - then I've

4) You will perhaps perceive a bias towards verbal representations in this choice of term. If so, it is allright, because it will normally be used of verbal representations. A possible pictorial parallel would have been **match**.
made a **forced** application. Henceforth, I will use "apply" and "applicable" without qualifications, as a shorthand for "positively apply" and "positively applicable".

From the practical point of view, obviously in a working representation system, there must be **application procedures** which can perform the necessary tests to decide if a label applies or not (if the key fits the lock or not). Applying a description means then, in practice, to use an application procedure directly attached to the description — if there is one — and/or testing indirectly by applying related labels and putting the results together appropriately, using their application procedures, and/or indirectly, ... and so on. There is no definite end to this, but of course, in practice, how wide the ripples of application are allowed to spread is a question of economy versus reliability.

Exemplification

Exemplification is in my view the most revolutionary and the most fruitful concept in Goodman's theory of symbols, in getting to understand pictures. It is, I think, a concept remarkably well suited for the purpose of explaining the essence of pictorial representation, but as we have seen, Goodman assigns exemplification a much less prominent role; **denotation** is designated the core of pictorial representation.

If pictures really denote, we need details on how they do it. The information we have from Goodman is mostly negative. A picture may exemplify labels which do not apply to the denotatum, so exemplification doesn't explain denotation. Neither does the kind of picture it is determine what it denotes, if
anything. Pictorial properties determine denotation, but this is not much to go by - the same is true of words written on paper, although of course many pictorial properties are then irrelevant: color, size, etc.

Yet, why is it that Churchill-pictures on the whole represent Churchill? After all, they could represent anybody, or anything, or nothing, since kind does not determine denotation. Why do man-pictures usually denote men? Eagle-pictures eagles? Etc. There's a remarkable regularity in all this.

Certainly "Churchill-picture" is a convenient label for a class of pictures that, exceptions notwithstanding, by and large denote Churchill. But this cannot be the whole story. Suppose I had met Churchill in person but never seen a Churchill-picture before - surely I would be quite proficient in recognizing Churchill-pictures among pictures in a symbol system I knew well. I might even be able to make them. I am sure that Goodman would agree. That a certain type of picture generally denotes a certain type of object, is not some sort of coincidence, it is conventional but systematic. This is in complete accord with the main message in Goodman's theory of symbols: symbols refer by virtue of being part of a system.

What then is the system? - what is it that connects a Churchill-picture and Churchill? Goodman has chosen to draw the line for his analysis at this point. My efforts to continue has led to an answer that makes the denotative role of pictures only secondary, but remains consistent with Goodman's general approach to representation.

About Churchill-pictures we may (rightly) say: "it's Churchill". Metaphorically, "Churchill", or some suitable description of him, applies to the picture.
A picture is a Churchill-picture because it exemplifies certain Churchill-descriptions that also apply to Churchill. In short: a Churchill-picture is an example of Churchill, it exemplifies certain characteristics of Churchill.

Thus Goodman's notion of a denotative pictorial symbol system dissolves into a composite system: an exemplifying system of pictures with labels from a verbal denotative symbol system. Pictorial "denotation" derives from exemplification.

This is, of course, very preliminary; there are a number of problems that must be cleared up. Most of them have to do with exemplification.

Exemplification reinterpreted

Let us return for a moment to the tailor's swatch again. The swatch is a sample of color. Suppose it is yellow. On Goodman's account, the swatch is so to say telling us: "I'm labeled 'yellow'". As customers we are indifferent, to say the least, to this information. We already know what yellow looks like, what we want to know is the color of a particular run of cloth. Clearly, the point with the swatch (as far as concerns color) is to inform about the color of this particular bolt. How does it manage to do that on Goodman's account?

As far as I can see, it doesn't. Even if we also knew that the swatch was taken from the bolt, even if the swatch were telling us "I come from the bolt" or

5) As I do not keep a very strict distinction between literal and metaphorical denotation, I will not in this context distinguish carefully between exemplification and expression. Besides, some of the descriptions involved may actually apply literally.

6) Of course, the swatch and the cloth need not have a common origin at all.
"I denote the bolt", it doesn't follow that the bolt is yellow. It would be different if we knew the swatch to be a fair sample of the run of cloth with respect to color, but this would be to beg the question.

Anyway, leaving the question to pragmatics is unsatisfactory: it would take the sting out of the concept of exemplification; it would degrade exemplification to little more than converse denotation.\footnote{Which, incidently, may be why several critics of Goodman have had trouble to see the difference between exemplification and converse denotation.}

Now, there are other labels - coextensive with "yellow" - which could alternatively be taken to be exemplified. Among them "the color of the bolt", which seems a much more reasonable choice than "yellow", considering that it is indeed the color of the bolt we are trying to get at. And here hides, I think, the missing link between swatch and bolt: the bolt satisfies the label "the color of the bolt" per definition, it is so to say the intended object, or standard model, for that description.\footnote{Of course, descriptions that have \textit{y} as a standard model do not generally have the form "the \textit{x} of \textit{y}".}

That a symbol sometimes exemplifies only one of two coextensive labels, is discussed (in a different context) at some length in Goodman (1968:54-56). His general conclusion is that the class of coextensive labels that also are exemplified can be broad or narrow depending on circumstances. However, he does not say along which dimensions we can loosen or tighten this class.\footnote{Although there are hints; for example, synonyms in the same language are said to be closer than synonyms in different languages.}

My interpretation of the example with the swatch, is that having the bolt as a standard model is a necessary condition on the labels that are...
exemplified by the swatch.

Model and standard model

Let us examine then this triple that is involved in an exemplification: exemplifying symbol, exemplified description, and intended model. Holding it all together is the relation of modeling: the exemplifying symbol as well as the intended object are models of the exemplified description. Now, the notion of a model is reputed to be muddled. With good reasons, no doubt, yet I think that the three main approaches to the model concept - the logician's, the ordinary abstractive view, and the exemplar or prototype view - have much more in common and are much less confused than is commonly thought. Consider the diagram in figure VI:1.

![Diagram](image)

Different readings of this simple diagram illustrate different emphases in what is in fact a single schema.

One reading is: "X is a model of Y" - meaning: X is a model of a description D that has Y as its standard model. This is the abstractive view. Examples are scale models, pins and balls models of molecules, and pictures in the narrow sense. Some mathematical models belong to this category; in cases where there
is clear distinction between a theory of some phenomena, and a mathematical model of the same phenomena. Other "mathematical models" are really theories; this I consider a misuse of the term "model".  

Another reading of the diagram is: "Y is a model for X". This view of the model as an exemplar is the one we meet in art, and in the use of a model as a prototype for industrial production.

Underlying any reading of the diagram is what I take to be the fundamental model concept - "X is a model of D" -, a model as something complying with a description, a notion which has found its clearest formulation in logic.  

The specialized concept of a standard model, or intended model, plays a crucial role in any endeavor to formally characterize something given. One of the most shocking results of modern logic is the general impossibility to pin down a particular domain of objects with a formal system. Geometrical theories are satisfied not only by lines and points, but by all sorts of things: light-rays or sticks - which may be precisely what we want, to be sure, but certainly not what the geometrician usually thinks he is primarily describing - and even by chairs and tables and beer mugs in certain arrangements. An example more frustrating to the mathematician is that any theory of the real numbers have some subset of the natural numbers as a model, a model which doesn't have the right (=the intended) cardinality.

---

10) But it still fits into the diagram, even though the labels have been transposed.
11) Related ideas can be found, e.g. in Suppes 1960, Wartofsky 1979, and Apostel 1960.
12) One of Hilbert's examples in Grundlagen der Geometrie, 1899.
13) Generally, the Löwenheim-Skolem theorems assert the existence of models of various
To any description of something given there are always extra models, sometimes wildly different from the intended model,\textsuperscript{14} the standard model. Certainly, one may go the other way around: start with a theory and then try to find out what it is a theory of. This is the road taken by large parts of modern mathematics. Even then one would probably be lost without some standard examples to go by. Anyway, such an aprioristic approach is obviously not open to knowledge representation. Nor do natural languages work in an aprioristic manner. However we formally define "chair" there are likely to be things, real or imagined, which fit the description but are not of the kind we want to think of as genuinely "chairs".\textsuperscript{15}

Reference to exemplified labels

But this is still much too loose. I agree with Goodman that complying with a label, being a model of a description, is not sufficient for exemplifying it. Anything is a model for an unlimited number of quite different, contrary, incompatible, and even incommensurable descriptions. Still, I wonder if Goodman doesn't sometimes overemphasize the distinctness and definiteness of the second element of exemplification: reference to the exemplified label.

\textsuperscript{14} Notice that the term "intended" model is a little misleading, since something isn't a standard model of something just because someone intends it to be. Generally, the standard model of a complex description is systematically related to the standard models of the parts of the description (the standard model of "wooden table" is obviously related to the standard models of "wood" and "table").

\textsuperscript{15} Cp Putnam's discussion of water, pencils and cats. The notion of standard models, normal examples, and their indispensability for verbal representations - and hence, indirectly, for pictorial representations - accords with many of his ideas about "stereotypes" (Putnam 1975).
Exemplification, just as application, is not on an individual symbol per symbol basis, it is systematic. The "reference" of a specific exemplifying symbol to the labels it exemplifies, can be understood in terms of a general reference - common to all symbols in the exemplificational (that is, pictorial) system - to a certain vocabulary, a certain verbal symbol system. The labels exemplified are then those in the vocabulary that apply to this particular symbol, and that converge on a common standard model. An exemplificational (pictorial) symbol system, is thus organized by a verbal system, in a sense it owes its existence to it.

For example, the tailor's swatch refers to a system of labels for color, texture, pattern, etc. From this collection of labels, those that apply to the swatch and have the bolt as standard model are exemplified. Consider a less extreme form of pictorial system, such as portrait paintings: in this case the vocabulary will include labels for hues and shapes, features and facial expressions etc. The exact extent of this vocabulary is difficult to ascertain, although there will usually be an undisputable core.

For many exemplificational symbols even the whereabouts of the verbal symbol system that organizes the exemplifying symbols, are unclear. When Goodman says that Socrates exemplifies "rational" (1968:54), what he means is that we can take him, or he is sometimes taken, to exemplify rationality. There is such a system. But of course he is not intrinsically part of such a system, and someone might very well take Socrates to exemplify "henpecked husband" (assuming now that he was henpecked), which puts a completely different perspective on Socrates. It is characteristic of many exemplificational systems16 that its symbols (as objects) are not very

---

16) Certainly not all - not e.g. wiring diagrams - but I think that such systems typically contain
tightly associated to any one particular symbol system. By contrast, most verbal representations in use have symbols peculiar to that system - a quick look at a symbol usually suffices to establish which symbol system it belongs to.

One manifestation of the indefiniteness of pictorial symbols is their general undecisiveness about degree of abstraction. Goodman takes a picture of an eagle in a dictionary - presumably accompanying a definition of "eagle" - as an example of a picture with multiple reference; the eagle picture denotes each eagle and nothing else (and is thus coextensive with "eagle"). But the same picture, the same pictorial object, might illustrate an article about the favorite pet of Charles V,17 and so represent a particular eagle; or it might illustrate an article on golden eagles; or it may illustrate an article on birds of prey; etc. If we don't understand the language of the dictionary, it is hardly possible to attach definite degrees of abstractness to the illustrations.18

Undecisiveness in emphases is also characteristic of pictorial representation:

"The picture of a huge yellow wrecked antique car ... does not commit itself to any of the following statements:

verbal elements, and that it is these elements that help to identify the system; in a wiring diagram the symbol for a diode may be such an element.

17) A fictive example.

18) Incidentally, I think that the main point with a picture such as this in a dictionary is - not to "recapitulate" in pictures what is already said in words - but to calibrate a certain word or description, that is, helping the reader to adjust some verbal label to a better fit with the example which the picture is. (See further below.) A second purpose is to exemplify properties not mentioned in the text.
The huge yellow wrecked car is antique
The huge yellow antique car is wrecked
The huge wrecked antique car is yellow
The yellow wrecked antique car is huge" (Goodman 1978:131)

Precisely. But it lends itself to any of these descriptions. It can be taken to exemplify any of them.

And in general, pictures allow multiple interpretations. The eagle picture once again, might exemplify the totem of the Hidatsa, or the largest bird of prey in Northern Europe, or the favorite game of the Algonkin.19 This undecisiveness of pictures is their weakness and their strength. Why it is a weakness needs no comment. But in what ways can the undecisiveness be an advantage?

Pictures are decisions withheld.20 Let us compare a picture with some description it (putatively) exemplifies. The description is i) selective relative to the depiction, it selects a definite degree of abstraction, it picks out and puts emphases on different details, etc (an individual eagle, or all eagles, the big, strong bird of prey with great power of vision, or the strong bird of prey of large size with great power of flight); ii) the description puts a more pronounced perspective on what is represented (the eagle as an object of religious worship, as an ornithological object, or as food); and iii) the description involves committals relative to the depiction.

19) Do not take these examples for facts! Incidentally, this illustrates my general point that examples are so important that if we can't find real ones, we invent examples.

20) Some decisions that is, certainly not "all", whatever that would mean. Every representation involves decisions.

- Exemplification 233
The selective aspect has been aptly characterized by Dretske:

"To describe a process in which a piece of information is converted from analog to digital form is to describe a process that necessarily involves the loss of information .... Until information has been lost, or discarded, an information-processing system has failed to treat different things as essentially the same. It has failed to classify or categorize, failed to generalize, failed to "recognize" the input as being an instance (token) of a more general type." (Dretske 1981:141)

Here, we may read "pictorial" for "analog", and "verbal" for "digital" and remain, I think, consistent with Dretske's intentions.

"I will say that a signal (structure, event, state) carries the information that s is F in digital form if and only if the signal carries no additional information about s, no information that is not already nested in s's being F. If the signal does carry additional information about s, information that is not nested in s's being F, then I shall say that the signal carries this information in analog form." (Dretske 1981:137)

Paraphrased: a representation r conveys information i in digital form if and only if r conveys no other information than i and whatever follows from i; r conveys i in analog form if r conveys information in addition to i and all its implications.21 A minor complication is that information, according to

21) Cp the previous discussion of additions to a representation - chapter II: Patrick Hayes - which made essentially the same point.
Dretske, can only be of what is the case, it is per
definition "truth". To cover representation in gen-
eral such a definition would have to be based not on
the concept of information, but on some notion of
"purported information".

Thus, an important aspect of the pictorial-verbal
distinction, is that a description exemplified, in a
sense is a distillation of the exemplifying depic-
tion. This notion of selection presumes a certain
interpretation of the picture that - in principle at
least - settles its information content. This
interpretation, this particular point of view, or
perspective, on the picture, is determined by the
verbal symbol system chosen to describe the picture.

Concomitant with the verbalization of a pictorial
representation are the new commitments they imply
relative to the picture - turning a picture into
words means taking risks, formulating hypotheses.
And this is because of the open-endedness of (many,
or most) pictures: it is not known exactly what in
the picture is significant and what isn't, and the
descriptions that can be taken to be exemplified -
even within a given perspective - isn't given once
and for all, discoveries are possible.

From this we can see at least one way pictures can be
useful, namely as a sort of general point of support,
something to fall back on should a verbal represent-
tation turn out to be faulty, or not serving its pur-
pose. A categorization may turn out to emphasize the
wrong things, or be too coarse-grained or too fine-
grained, or it may turn out to be simply mistaken -
then we may return to the picture and try again, more
carefully. Yet a case where this strategy is useful
is when something has been correctly classified,
correctly described, but it turns out that the
categorial framework, the vocabulary used is

- Exemplification 235
A picture of pictures

Hopefully, the picture is becoming clearer. Here, I think, is the germ of a theory that breaks with tradition in that it neither excludes computers from the use of pictures, nor takes to ordinary (untenable) congruence theories, yet preserves many of our strongest intuitions about pictures. It is interesting to note that my account of pictorial representation, developed from Goodman's theory of exemplification, accords quite well with Johnson-Laird's theory of "mental models" (Johnson-Laird 1983) which builds on a completely different, psychological, tradition (Craik 1943). Certainly, loose ends remain and I will briefly indicate how some of them may be tied up, and also make other comments on my proposal.

Representation-as

How does this new theory fit in with representation-as? The perspective a pictorial symbol gives on what it represents, can be directly associated with the vocabulary, the verbal symbol system, involved in the exemplification. A picture representing Churchill as an officer exemplifies a description of Churchill (partly) in military terms. A picture representing Churchill as a lion exemplifies a description of Churchill in terms of leonine features and properties. However, there seems to be a little problem here; Robinson makes the following interesting observation:

"... to understand the point of such pictures [Churchill as a lion] we need to know what properties Churchill and a lion ... have in common,
but the relevant properties are precisely those which are not represented in the picture. Churchill does not have four paws but he is fierce and courageous ..." (Robinson 1978:52)

I would say that such a picture doesn't represent Churchill as either fierce or courageous, but simply as "leonine". If it is right to picture Churchill as a lion, it is because he is leonine. He isn't leonine in the same way as the picture is, nor in all the ways that a real lion is, but leonine he is nevertheless. He is leonine in some of the ways a human being can be leonine, just as the picture is leonine in some of the ways an oil on canvas painting can be leonine (obviously, it can't roar).

Exactly what these ways are is an open question. It is reasonable to think that a human being can't have paws and claws, whereas he can be courageous and fierce. Is a man represented as a lion also lithe? (Yet an illustration of the open-endedness of pictures.)

Then why isn't it simply a picture of a lion? Because it doesn't simply exemplify "leonine". As Robinson points out, there are two gross categories of pictures of Churchill as a lion: straightforward lion-pictures, and pictures where a most extraordinary lion "is smoking a cigar, haranguing the beleaguered troops or lounging on the Front Bench of the

22) Things are complicated by the possibility that lions do not have the purported properties, perhaps they really are docile and cowardly. It would hardly be right then to depict Churchill as a lion, or describe him as a lion. Unless, that is, lions are taken as symbols straightforwardly denoting courage, etc, which means that we have here a longer chain of symbolic reference. A case in point is the ass, which - as everybody knows - really isn't "asinine", and yet functions as a symbol of stupidity.
Commons." (Robinson 1978:45). The latter kind of picture is easily identified as belonging to some political cartoon variety of symbol system, and Churchillian features are straightforwardly represented - cigar, profile, etc. Together with the leonine features, the picture forms an exemplification of - not simply "leonine", but, something more like "leonine cigar-smoking with a suchandsuch looking face" does not have a lion as a standard model even if the standard model for "leonine" is a lion; this standard interpretation is ruled out by the rest of the description.

Similarly for the first kind of picture, although here we need more context to understand which symbol system it belongs to - it might be a symbol system representing heads of states as different animals, where this particular picture exemplifies a "leonine Prime Minister".

Null-denoting and impossible pictures

A unicorn-picture exemplifies a unicorn-description, a description the standard model of which happens not to be materialized in the actual world. There is nothing mysterious about a standard-model which doesn't exist, although there is then nothing to play the role of a standard for the description. From the description we can yet have quite detailed knowledge what it would be like if it did exist.

That leonine features are interpreted less detailed and concretely than the human features is part of the system of political cartoons.

Presumably, it could in a different context pass as a representation of a lion.

And here we deviate from the usage of the logicians and mathematicians, for whom existence is - usually - equal to possibility.
An impossible picture would exemplify an inconsistent description, which is impossible since an inconsistent description has no model; one cannot exemplify something which has no examples. Hence, there are no impossible pictures! What shall we say then of Reutersvärd's and Escher's drawings, the Penrose triangle, etc? Such "pictures" look enough like symbols in certain well-known pictorial symbol systems that we're perplexed to find that the labels we try to apply doesn't quite fit - typically they fit locally but taken globally they are not coherent. We end up with a locally wrong but consistent description, or a locally right but inconsistent description, of course, neither of which the putative picture is an example of.

To really be a picture the "impossible picture" must be taken to exemplify some consistent description (hence to depict something possible), which is sometimes feasible, but then in some unusual or contrived interpretation. A case in point is the Penrose triangle of which even a photograph has been made - of course the object photographed wasn't a triangle at all but an open zigzag-shaped structure which looked like a closed triangle from a particular point of view.

Now, "impossible" might be taken to refer to some empirical impossibility, not logical. Still, if the picture is subject to the same general empirical constraints as the depicted object, essentially the same argument applies - see further below in the section on "Analogy reasoning".

This would explain our intuition that pictures are less than words prone to represent impossibilities: there are simply no impossible pictures. A different problem then comes to the fore: recognizing whether a given object really is a symbol in a given pictorial

- A picture of pictures
symbol system.

Genuine pictures and quasi-

It is often said that pictures are a more direct form of representation, a view which has its roots in congruence theories of depiction and the idea of an intrinsic relation between picture and what is depicted. In Goodman's theory, pictorial and verbal representations are on an equal footing in that respect. A consequence of my proposal is that genuine pictures are even less direct than words.

The difference between genuine pictures and quasi-pictures is the difference between a system which explicitly maintains and manipulates descriptions exemplified, and a system which doesn't treat the "pictorial symbols" different from how any denotative symbol would be handled. That is, a quasi-picture is purely denotational, although it could have been exemplificational if supplemented with the right descriptions. In other words, what matters is where the exemplified verbal labels are located: if they exist within the system then picturing is genuine, if they exist only without the system as supplements provided by some observer, then the symbol system is quasi-pictorial.

An illustration will make this more clear. Imagine a system using sequences of dots to represent quantities of apples, say. The symbol for one apple is a single dot, for two apples, two dots in a row, etc. Assume also that the interface to the representation

26) See for example The handbook of artificial intelligence, Barr & Feigenbaum 1981.
27) A different sense in which pictures have been held to be more direct is that they are supposed to lack the so-called "double articulation" of language. This is dubious - a thoroughly digitalized picture is also doubly articulated.
has every appearance of a pictorial representation.

To the outside observer it will appear as a pictorial representation: the number of dots can clearly be taken to exemplify "the number of apples". But the actual user of the symbols may happen to use them as simply denotative: this particular arrangement of dots is a symbol which denotes that there are two apples, that particular arrangement a symbol which denotes that there are three apples, etc. To the user then, the symbols simply do not depict apples; the system would work equally well if two dots represented three apples and three dots two apples, for example.

I think, although it may sound paradoxical, that the more a pictorial symbol system interacts with verbal symbols, the more genuinely pictorial it is.

Analog and intrinsic, formal properties

On the proposed account of pictorial representation, depiction will naturally be analogy analog - properties and relations of the exemplifying model corresponding to properties and relations of the standard model that model the same labels in the exemplified description.

Pictorial representation will also naturally be (virtually) intrinsic - at the level where the model makes contact with the description, labels correspond to properties and relations of the model, and general constraints prescribed by the description must per definition be satisfied by the model.

The logical properties of the relation between picture and depicted object which issues from my explanation of pictorial representation are the ones we expect of any relation of representation: non-transitive, non-reflexive, and non-symmetric.
Uses of pictures - and words

We saw above that pictures can be generally useful to back up verbal representations. There are several variations on this theme, as well as different themes. Apparently, there is a multitude of possible uses for pictures, based on an interplay with verbal representation. I will briefly describe some of them, making no pretense to exhaustiveness. Fundamental to many of these ways of using pictures is the capability to generate descriptions of a picture, and the capability to generate examples and counterexamples.29

An important use of pictures is to check the consistency of some verbal representation. Success in trying to make a model of a description, shows that it is consistent. Repeated failures indicates imminent danger of inconsistency. A related use is to check two (or several) descriptions for coreference. Success in trying to make a model that fits both descriptions and which cannot be separated into two models (in the same medium) each of which complies with only one of the descriptions, indicates that they are possibly referring to the same thing.31

28) The theme of a dialectic between formal characterizations and examples is of course an old one. More recently, Edwina L. Rissland has brought out this theme anew, partly building on the philosophy of science tradition of Kuhn and Lakatos. Rissland (1984) gives a summary and references.


30) But could, of course, be due to inadequate modeling capabilities.

31) In one rather reasonable way of understanding coherence, two descriptions are coherent if they have a common non-separable model. In particular the descriptions could label situations rather than objects.
Related to this is also the use of models to consolidate or amalgamate\textsuperscript{32} descriptions. A common model is created, linking the descriptions together. Although the condition of non-separability is here not a necessary requirement, it is still something desirable—the degree of separability can be taken as an inverse measure of how tightly one has succeeded to consolidate the descriptions. This model may be used to create a new, simplified and comprehensive description of it, which is an amalgamation of the original descriptions. An important point here is simplification, how several description may fuse to form a solid whole without joints; a good illustration would be WHISPER's way of finding a smooth curve formed by the accidental alignment of several objects, if we assume that the objects were originally given as descriptions.

A different role for pictures is as counterexamples. A counterexample of a description is something which isn't a model of it; it is a non-example which should have been an example, useful to expose descriptions that do not apply to the class of objects they claim to.

There are several variations in how counterexamples arise. A counterexample may be the (perhaps unexpected) outcome of an attempt to create an example, or it may result from deliberate efforts to create a counterexample. There may also be standing representations serving as touchstones which verbal descriptions may pass (example) or fail (counterexample). This is the predominant role of the diagram in the geometry machine; the diagram is an example of the original description (premises and axioms), and since every description on the road to a proof of the given theorem (known to apply to everything the premises

\textsuperscript{32} I have borrowed this term from Funt (1976).
and the axioms of plane geometry applies to) must be exemplified by this diagram, when the diagram is a counterexample of a description it means that the description is not on the path to a proof.\(^{33}\)

Pictures can be used to accomplish a change in perspective. Given a description, a model is created to which a different vocabulary is applied, creating a new description with a different perspective. The importance of having the right representation, the right point of view, in attempting to solve a problem, has been stressed repeatedly, even to the point where having the right representation is taken to be almost tantamount to having the solution. Since problems that arise in a natural context seldom present themselves in the proper format for solving them (or else they wouldn't be problems) it is essential to be able to change between different perspectives. It is doubtful if perspective changes are even possible without the help of a model.

The two concluding examples of varieties of pictorial use - calibration and analogy reasoning - I have chosen to elaborate a little more on.

**Calibration**

Sometimes a description loosely fitting a picture is stipulated to fit the picture tightly, possibly forcing more or less temporary adjustments and refinements of the semantics, the application procedures. I will call this process of fine tuning a verbal label to a particular example, calibration. Elgin (1983:74) has an example that makes a nice illustration of calibration:

\[33\] Lakatos (1977) gives an insightful and instructive demonstration of how counterexamples may be used not only to discard unsuitable formulations but also to suggest how they should be amended.
"... consider the note produced by an oboeist to which other members of an orchestra tune their instruments. The note exemplifies a certain pitch, but which one? To simply identify it as an A is far too crass, for then all the musicians need do to play in tune is hit the right notes. ... the note he [the oboeist] produces sets the standard for that level of attunement."

To Elgin, a careful expounder of Goodman's theory, this example poses a problem. For one thing, because

"... we may have no sufficiently precise verbal classification of the note he produces. ... If the note's exemplification is null ... it cannot perform its normative function." (ibid.)

Second, and more serious I think, is the problem of what pitch the note A really denotes. If it is a fixed precise pitch, 440.00 Hz say, then spurious performances are almost the rule. And if it isn't a fixed pitch, then why tune the instruments?

Both problems dissolve if we make use of the notion of calibration, abandoning the idea of an absolutely fixed, rigid semantics. The note A initially denotes pitches roughly between 430 and 450 Hz, say, but is calibrated to a precise pitch whenever a score is to be performed. The relations between pitches are much more precisely and rigidly designated, which means that the calibration of a single note suffices to calibrate all the others.

In this particular example, the calibration relies on a sound, something not obviously symbolic outside the particular context of calibration.34 It could have been done with the help of something more obviously

34) Although something could be said for the idea that music is symbolic.
pictorial, for example using an oscilloscope. There is no principal difference in the tuning process in these two cases, even though the types of application procedures involved are different.

I think the technique of calibration opens up a general strategy that could be most useful: having examples at hand, permits application procedures to be initially rough and simple, by calibration they can be made as refined and precise as is called for on any particular occasion. (Note that the initial precision of a label may sometimes be sufficient to pick out the example to use in a calibration, even in the absence of other indications.) The economy in this is to be found both in the restrictive use of refined application procedures and in having a single symbol (system) cover a wide range of cases without giving up precision.35

**Analogical reasoning**

Analogical reasoning has been one of the more popular subjects for experimentation and speculation in AI. I do not intend here to add to the know-how of analogy use. Rather, I want to discuss what analogical reasoning is and how it can be justified, partly in the context of scientific research methodologies.36 Generally, I find the parallels between artificial intelligence and philosophy of science most illuminating, and it seems to me that they are not exploited nearly as much as they could be.

Basically, analogical reasoning is to proceed from a description exemplified by a picture, to a different

---

35) Incidentally, this is, I believe, essentially how indexicals work.

36) Which means, among other things, that the term "theory" will be used, however, the discussion is valid for any type of verbal representation.
description that applies to the same picture. Details may vary: the picture may have to be created from the original description, or it may already exist, the picture may be static or be dynamic (that is, involve the depiction of actions and events), etc.

Why does reasoning by analogy work? Compared to the huge efforts spent in trying to justify induction, the parallel task of justifying analogy has received surprisingly little attention. Even among the relatively few philosophers with a benevolent view on analogy, it's difficult to find serious attempts to work out an acceptable rationale for using analogies. The following quote from Black (1962:238) is representative.

"...the successful model must be isomorphic with its domain of application. So there is a rational basis for using the model. In stretching the language by which the model is described in such a way as to fit the new domain, we pin our hopes upon the existence of a common structure in both fields. If the hope is fulfilled, there will have been an objective ground for the analogical transfer. For we call a mode of investigation rational when it has a rationale, that is to say, when we can find reasons which justify what we do and that allow for articulate appraisal and criticism. The putative isomorphism between model and field of application provides such a rationale and yields such standards of critical judgement."

It all agrees quite well with our standard intuitions about what analogy is, and something of what it takes to be a good model. But does it help in understanding why it is, by and large, worthwhile to engage in analogical reasoning? Surely, that a proposed model in the end turns out to be successful, that the model
in retrospect is shown to be "isomorphic" to the domain, cannot be a rationale for adopting it.

Now, there are cases where success is guaranteed, where the "common structure" is established beforehand. In such cases, just as in certain proofs in geometry, or in the use of a slide-rule, we are simply exploiting known congruences, and analogical reasoning reduces to a sort of calculating aid in making deductions; there's absolutely no risk involved. This is one - important - category of analogical reasoning, inherent in many uses of pictorial representations, and discussed at length by Sloman (1978).

But it is not the kind of analogical reasoning I'm primarily thinking of here, nor is it, I believe, what Black has in mind in the passage quoted above (see also the following quote). The use of analogy that has interested scientists and philosophers of science is where it - just as induction - goes beyond deduction, where analogy is an aid in making genuine discoveries and forming hypotheses. Black writes (1962:233):

"A promising model is one with implications rich enough to suggest novel hypotheses and speculations in the primary field of investigation."

This is perhaps the most celebrated use of models in science; for instance Bohr's billiard ball model of the atom, Huygen's water wave model of light, the model of electricity as a liquid, etc. Here, the point with the model is that we don't know in advance to which extent it agrees with the domain. In other

37) I do not wish to imply that something deduced can not be a genuine discovery, only that deduction is not a powerful method of making interesting discoveries, and that it has definite limits - one cannot deduce something which doesn't deductively follow.
words, such pictures are open-ended from the user's point of view.

Note that even where a general and complete (and "right") theory of the domain is available, analogical reasoning of the second kind can be extremely useful.

As an example, consider diagrams in the geometry machine. For instance take the diagram used in the proof that if ABCD is a quadrilateral with BC parallel to AD and BC equal to AD, then AB is equal to CD (Gelernter 1963:149). The open-endedness of such a picture is unmistakable: the diagram represents every quadrilateral that conforms to the premises, but it is - from the practical point of view - an open question which labels are exemplified in addition to the premises. That the diagram is suchandsuch (for example, that ABD is congruent with CDB) suggests that it might exemplify "suchandsuch" (being one of the necessary conditions for exemplification) but is obviously insufficient to ascertain if this property really is represented or not. And of course this will be relevant if the geometry machine is to use the diagram as positive heuristics.

There is a different reading of Black. It could be that the "common structure" he is talking of is established beforehand and hypotheses are formed on the basis of the model, that are not simply part of what is already known to be common, but really go beyond that. In other words, one expects that a partial "isomorphy" can be extended to a fuller "isomorphy". Still, there's no explanation why, if two things are

38) It is convenient to speak of different "kinds" of analogical reasoning, but potentially misleading since there is to be sure no difference in the process, only in the epistemic status of the results.
similar in some respects, we should generally expect that they are similar also in other respects.

To me, a *rationale* is a rule recommending (dissuading) a particular action, giving rational reasons why that choice of action is better (worse) than random, based only on what is known before the choice. There are some alternatives for what exactly "rational" should mean in this context. To say that it's all right to use analogies because it usually works, is to be sure a sort of rationale, but certainly not the sort of rationale that has satisfied philosophers. The effect is just to reduce the problem of justifying analogy to the problem of justifying induction, that perennial stumbling-block.

I will try to be strict about it and suggest a rationale for analogical reasoning, that — who knows? — might be acceptable even for die-hard positivists. But first, let us examine the common assumption that analogy is just a special kind of induction. I think it's fairly easy to see that there is an important difference. One black crow gives by analogy a rationale (assuming now that there is one) for forming the hypothesis that any other particular crow will be black too. So far it looks very much like induction. The difference shows when we bring in the second (black) crow: it doesn't strengthen the analogy rationale for the hypothesis a bit. A thousand black crows wouldn't make the analogy stronger.

Now, this is not how induction is supposed to work. Analogy, provided we can find a rationale for it, supports any hypothesis that a particular crow is black, as well as the hypothesis that all crows are black, but the number of known positive instances are

39) An assumption that may have blocked philosophers from further examination of analogy.
in both cases immaterial for the strength of the support. One is enough. More instances doesn't add anything. In fact, negative instances are also immaterial to the analogical support for the hypothesis that a particular crow is black: a white crow would just give support for a different (competing) hypothesis, and the original hypothesis will not be destroyed or undermined.\(^\text{40}\)

In short: whereas induction claims to indicate the probable, analogy only claims to indicate the possible or plausible.

There's still hope then, for a justification of analogy that avoids the pitfalls of induction. My proposal is based on the assumption that there are certain general constraints on our descriptions of the world (or on the world itself, if you prefer). We don't need to assume any particular constraint, just that there are constraints. At the minimum I will assume there are the logical constraints to consistent theories; it will suffice to prove the general point, while demonstrating the general and simple line of reasoning.\(^\text{41}\)

Suppose we have a theory and wish to add some hypotheses to it. A hypothesis that is inconsistent with the original theory is clearly a bad choice, it is doomed beforehand. Certainly it is rational to avoid the investigation of inconsistent extensions.

How can this be avoided? To do it by purely syntactic means can be quite awkward.\(^\text{42}\) A most useful solution is to employ a model: as soon as we feel assured

\(^{40}\) A white crow would of course be devastating to the hypothesis that all crows are black.

\(^{41}\) See Nickles 1977, for some related ideas.

\(^{42}\) Cp Hayes's motive for proposing the consistency based method for handling the frame problem, above.
that a model exists we may also feel assured that any theory it happens to comply with is consistent (satisfies the logical constraints). Thus, if we never attempt extensions that we don't have some model for, we're safe. Of course, the most important model for a theory is the intended model, the domain for the theory. But if we really are in a situation where extending the theory has become something of a problem, it means that we do not have full access to the intended model, it isn't available, we cannot access it in the way we need to, etc.

This is where analogy comes in. An auxiliary model - a picture - is chosen (or is ready at hand for such occasions), that also complies with the theory, and that has a higher degree of availability and convenience than the standard model. These auxiliary models are precisely those things usually called "models" by philosophers of science. The theory is then extended to cover some of the additional features of the auxiliary model; the extension - the new hypothesis for the intended domain of application - is thereby guaranteed to be consistent. And that's that.

What is the power of this rationale? Does it really bring about any significant reduction of the class of candidates for a theory extension? This certainly requires further study, but I think the potential efficiency should not be underrated. If the original theory is sufficiently dense - that is, has a certain degree of complexity and degree of interrelation between concepts - and we restrict extensions to cohere with the original theory⁴³ it would seem that the rationale can be quite efficient. Compare the efficiency of using the diagram for pruning in the

⁴³ All of which are reasonable demands on any mature theory. Cp also Hayes's general reflections in the introduction to the Naive physics manifesto, Hayes 1978.
geometry machine, and consider that any description excluded by pruning would also be excluded by analogical reasoning. The more dense the theory, the more inadequate hypotheses will be excluded by the use of analogy. It is a general characteristic of verbal representations, that their range of applicability shrinks as they are made more elaborate while remaining coherent.

If we assume other, non-logical, constraints, the rationale might gain considerably in power. (Some conceivable examples of such constraints are: signals travel with finite velocity; there is no spontaneous change; two things cannot occupy the same position at the same time; one thing cannot occupy two different positions at the same time.) The argument will be completely parallel to the one just given, only that now the model is guaranteed to satisfy not just the logical constraints but also more restrictive, non-logical constraints; if we can find a model, then we know that the description is not empirically impossible (in the particular sense of the assumed constraints). The effect is simply to restrict the class of possible extensions further. Notice also that we may have particular reasons to assume that the auxiliary model and the intended model fall in the same area, and so have reason to expect more narrow constraints.

On the assumption that relatively new-born theories (generally - initial descriptions of something) are not very dense but become more dense as they develop, then the above explanation of analogical reasoning leads to the prediction that analogical reasoning is not so efficient in the initial attempts to verbally characterize something, as after a while when the description is beginning to consolidate. Hence, in an initial phase using pictures as counterexamples (negative heuristics) might prove more efficient than
using them for analogical reasoning (positive heuristics).

To sum up, I suggest that there is a single very simple principle underlying analogical reasoning, namely - Do not attempt verbal representations that there is no known model of! - and the rationale behind this recommendation is that by avoiding descriptions that have no known model, we avoid impossible descriptions, and thus a potentially large class of descriptions that cannot possibly apply to whatever we are trying to describe.

Prospect

As I said at the outset, this is to be understood as a report on work in progress. I hope to have conveyed some interesting ideas that, given more work, can be developed and elaborated into a more comprehensive and full-blown theory. It is interesting to note that some recent research on machine learning fits in with my basic ideas about pictorial representation and the role of exemplification. This work will be of great value, I think, for the future development of my theory.

Looking back at the questions I formulated in the Introduction, it now seems to me that instead of taking pictorial and verbal representation as primarily competing alternatives, it is more fruitful to view pictures and words as complementary forms of representation mutually supporting and depending on each other.

44) See Michalski, Carbonell & Mitchell 1983.


Eastman, C.M. (1973) Automated space planning Artificial Intelligence, 4:41-64.


Feigenbaum, E.A. & Feldman, J., eds. (1963)
Computers and thought

Fodor, J.A. (1975)
The language of thought
Crowell, New York.

Representations
Harvester Press, Brighton.

The modularity of mind

Funt, B.V. (1976)
WHISPER: A computer implementation using analogues in reasoning
Report No. 76-09, Computer Science Department, University of British Columbia.

Funt, B.V. (1977)
WHISPER: a problem-solving system utilizing diagrams and a parallel processing retina

Funt, B.V. (1980)
Problem-solving with diagrammatic representations
Artificial Intelligence, 13:201-230.

*Funt, B.B. (1983)
A parallel-process model of mental rotation

Gelernter, H. (1959)
A note on syntactic symmetry and the manipulation of formal systems by machine
Information and control, 2:80-89.

Gelernter, H. (1963)
Realization of a geometry theorem-proving machine in Feigenbaum & Feldman (1963). (First published 1959.)


Gelernter, H. & Rochester, N. (1958)
Intelligent behavior in problem-solving machines
Gibson, J.J. (1971)  
The information available in pictures  
Leonardo, 4:27-35.

Gibson, J.J. (1973)  
On the concept of formless invariants in visual perception  
Leonardo, 6:43-45.

Gibson, J.J. (1979)  
The ecological approach to visual perception  
Houghton Mifflin, Boston.

An examination of the geometry theorem machine  
Artificial Intelligence, 1:171-187.

Gombrich, E.H. (1960)  
Art and illusion  
Phaidon Press, London.

Gombrich, E.H. (1972)  
The mask and the face: the perception of physiognomic likeness in life and in art  
in Mandelbaum (1972).

Goodman, N. (1968)  
Languages of art  
First edition, Bobbs-Merrill, Indianapolis.

Goodman, N. (1971)  
On J.J. Gibson's new perspective  
Leonardo, 4:359-360.

Goodman, N. (1972)  
Problems and projects  
Bobbs-Merrill, Indianapolis.

Goodman, N. (1976)  
Languages of art  

*Goodman, N. (1977)  
The structure of appearance  
Third edition, Reidel, Dordrecht. (First edition 1951.)

Goodman, N. (1978)  
Ways of worldmaking  
Harvester Press, Hassocks.

Fact, fiction, and forecast  
Harvard University Press, Cambridge, Massachusetts. (First edition 1954.)
Of mind and other matters

Haugeland, J. (1981)
Analog and analog
Philosophical Topics, 12:213-225.

Hayes, P.J. (1974)
Some problems and non-problems in representation theory
AISB Summer conference, University of Sussex, pp 63-79.

Hayes, P.J. (1977)
In defence of logic
IJCAI-77:559-565.

Hayes, P.J. (1978)
The naive physics manifesto
Institut pour les etudes semantiques et cognitives, Universite de Geneve, Fondazione Dalle Molle, Working papers No. 35.

Hayes-Roth, F. (1979)
Distinguishing theories of representation: a critique of Anderson’s "Arguments concerning mental imagery"

*Hermerén, G. (1978)
Depiction: some formal and conceptual problems

*Hermerén, G. (1983)
Aspects of aesthetics
CWK Gleerup, Lund.

*Hesse, M. (1966)
Models and analogies in science
University of Notre Dame Press, Notre Dame, Indiana.

The logical structure of pictorial representation
Theoria, XL:76-109.

Johnson-Laird, P.N. (1983)
Mental models

A three point perspective on pictorial representation: Wartofsky, Goodman and Gibson on seeing pictures
Erkenntnis, 15:55-64.
References

A psychology of picture perception

*Kjærup, S. (1978)
Pictorial speech acts
Erkenntnis, 12:55-71.

Wittgenstein and the philosophy of pictorial languages
in K. S. Johannessen & T. Nordenstam, eds.,
Wittgenstein - aesthetics and transcendental
philosophy, Hölder-Pichler-Tempsky, Vienna.

Images, symbols, and skills

Kosslyn, S. M. (1978)
Measuring the visual angle of the mind's eye

*Kosslyn, S. M. (1981)
The medium and the message in mental imagery, a
theory
in Block (1981).

On the demystification of mental imagery
in Block (1981).

Kosslyn, S. M. & Pomerantz, J. R. (1977)
Imagery, propositions, and the form of internal representations
Cognitive Psychology, 9:52-76.

Kripke, S. (1980)
Naming and necessity
Blackwell, Oxford. An earlier version was published
1972 in Harman & Davidson, eds., Semantics of Natural
Language, Dordrecht, Reidel.

Lakatos, I. (1977)
Proofs and refutations

Metaphors we live by
University of Chicago Press, Chicago.

*Lewis, D. (1971)
Analog and digital
Nous, 5:321-327.
Picturing

Mandelbaum, M., ed., (1972)
Art, perception, and reality
Johns Hopkins University Press, Baltimore.

Vision
Freeman, San Francisco.

Non-monotonic logic II: non-monotonic model theories

McDermott, D. & Doyle, J. (1980)
Non-monotonic logic I
Artificial Intelligence, 13:41-72.

Machine learning, an artificial intelligence approach
Tioga, Palo Alto.

*Morawski, S. (1978)
Three observations on Languages of art
Erkenntnis, 12:119-128.

Nickles, T. (1977)
Heuristics and justification in scientific research: comments on Shapere
in F. Suppe, ed., The structure of scientific theories, University of Illinois Press, Urbana.

Paivio, A. (1971)
Imagery and verbal processes

*Paivio, A. (1977)
Images, propositions, and knowledge

Palmer, S.E. (1975)
The nature of perceptual representation: an examination of the analog/propositional controversy
TINLAP-1:165-172.

Palmer, S.E. (1978)
Fundamental aspects of cognitive representations
Putnam, H. (1975) 
*Mind, language and reality*, Philosophical papers, volume 2

Pylyshyn, Z.W. (1973) 
What the mind's eye tells the mind's brain: a critique of mental imagery 

Pylyshyn, Z.W. (1975) 
Representation of knowledge: non-linguistic forms. Do we need images and analogues? 
*TINLAP*-1:174-177

*Pylyshyn, Z.W. (1978)*
Imagery and Artificial Intelligence 

Pylyshyn, Z.W. (1979) 
Validating computational models: a critique of Anderson's indeterminacy of representation claim 

Computation and cognition: issues in the foundations of cognitive science 

Pylyshyn, Z.W. (1981a) 
Complexity and the study of artificial and human intelligence 

Pylyshyn, Z.W. (1981b) 
The imagery debate: analogue media versus tacit knowledge 
*Psychological Review*, 88:16-45.

*Reiter, R. (1972)*
The use of models in automatic theorem proving 
*Technical report* 72-09, Department of computer science, University of British Columbia, Vancouver.

*Reiter, R. (1976)*
A semantically guided deductive system for automatic theorem proving 

Example generation 
The ubiquitous dialectic
ECAI-84:663-668.

Constrained example generation: a testbed for
studying issues in learning
IJCAI-81:162-164.

Robinson, J. (1978)
Two theories of representation
Erkenntnis, 12:37-53.

Robinson, J. (1979)
Some remarks on Goodman's language theory of pictures
British Journal of Aesthetics, 19:63-75.

Portraying analogy

*Rudner, R.S. (1978)
Show or tell: incoherence among symbol systems
Erkenntnis, 12:129-151.

Sandewall, E. (1985)
A functional approach to non-montonic logic
Report LITH-IDA-R-85-07, Department of computer and
information science, Linköping university, Linköping.

*Savile, A. (1971)
Nelson Goodman's 'Languages of art': a study
British journal of aesthetics, 11:3-27.

*Schon, D.A. (1963)
Displacement of concepts
Tavistock Publications, London.

Imagery - there's more to it than meets the eye
in Block (1981).

*Scruton, R. (1974)
Art and imagination
Methuen & Co Ltd, London.

*Shepard, R.N. (1970)
Second-order isomorphism of internal representations:
shapes of states
Cognitive Psychology, 1:1-17.

Shepard, R.N. (1975)
Form, formation and transformation of internal
representations
in R.L. Solso, ed., Information processing and
cognition, the Loyola symposium, New York.


Zemach, E.M. (1975) 
Description and depiction 
Mind, LXXXIV: 567-578.