10.02.91/NOB/final

JOHNSON-LAIRD's THEORY OF MENTAL MODELS

Niels Ole Bernsen

What follows is a discussion of Johnson-Laird's theory of mental models, mainly as presented in his 1983 book on the subject (*Mental Models. Towards a Cognitive Science of Language, Inference, and Consciousness*). The notion of mental models seems to be gaining prominence in cognitive science. Johnson-Laird's work is central to the understanding of this notion, and since his views on mental models do not appear to have changed significantly since 1983 (some 1983 views would appear to have been abandoned, however, cf. Johnson-Laird 1989), it seems worthwhile to take a look at the 1983 theory. I propose to look at mental models in deductive reasoning (sect. 1), mental models in discourse (sect. 2), mental models in general (sect. 3), and finally discuss a number of points (sect. 4).

1. Mental Models in Deductive Reasoning.

A deductive inference is valid if and only if there is no interpretation of the premises that is consistent with a denial of the conclusion. One possible assumption is that deductive reasoning and inference is based on the use of a formal mental logic such as the propositional calculus. It seems, though, that no one has proposed that the much more complex first-order predicate calculus is mentally implemented.

Problems with the assumption are:
1. People make mistakes in performing deductive inference.
2. Formal logic today comes in many different varieties. It is not clear which variety (of, e.g., modal logic) is identical to the presumed mental logic. Mental logic may be axiomatic or may use inference schemata as in natural deduction. Again it is not clear from experimental evidence which formal logic, if any, is in the mind.
3. If there is a mental logic, it is presumably innate. We need an account of how this might be possible.
4. In formal logic, deductions are valid by virtue of their form, not their content, since formal rules of inference work in a purely syntactic way. But problem content does affect inferential performance as shown, e.g., in the Wason/Johnson-Laird card selection task (Wason and Johnson-Laird 1972) and in the many subsequent experiments performed within this paradigm. How does a mental logic theory account for this ?
5. Subjects do not draw just any one out of the infinite number of valid conclusions deducible from a given set of premises. Why ? Perhaps because they use triviality-filtering heuristics such as that no conclusion should contain less semantic information than the premises on which it is based or should fail to express that information more parsimonously, or that conclusions should not repeat what is already obvious such as simple categorical premises already stated. There may, of course, be many more principles involved here. And in many cases in ordinary life, we add information when drawing (consequently deductively invalid) conclusions.
6. Even *modus ponens* can be suppressed under certain circumstances, as in:

If it rains, she gets wet.
If she goes out, she gets wet.

It rains.

In this particular context, many subjects suppress the conclusion that she gets wet (Byrne 1989, Johnson-Laird 1989).

The assumption of a formal mental logic is also problematic with respect to non-deductive inference. Many ordinary inferences are nonmonotonic. Nonmonotonic (formal) logic is not likely to be able to model much of this since nonmonotoniticity often stems from contents rather than from logical form. Many inferences in daily life are not derivable within a formal calculus because they depend on the particular situation to which the premises refer, or are just plausible on the basis of general knowledge, or derive from premises which can never be rendered sufficiently complete to ensure validity, or are inductive. It is a well-known fact that students of formal logic have to learn the skill of transforming natural language sentences into statements of formal logic, and, at least so far, this skill is non-algorithmic.

It is no satisfactory alternative to the hypothesis of some formal mental logic to assume instead that all inference is based on content-specific rules of inference such as production rules or connectionist pattern-matching. The reason is that also our ability to reason in a content-independent way has to be psychologically explained. The same is true of our ability to reason non-deductively. It remains true, however, that Newell's state-space model of human problem-solving (implemented as SOAR) is somehow closely related to mental model theory. Newell (1987) proposes that problem-solving can be seen as a series of transformations of a model of the initial state of affairs in a problem.

Johnson-Laird proposes an alternative to the hypothesis that, normally, deductive inference is based on a mental logic with formal rules of inference. Mental model theory does not assume rules of inference of any sort, either formal or content-specific, but instead assumes that reasoning depends on the manipulation of mental models.

In order to be able to perform deductive inference in the propositional calculus, it is sufficient to master the meaning (semantics, truth conditions) of the logical connectives in the form of truth-tables and to apply this knowledge in systematically eliminating models of the premises which are inconsistent with that semantics. In this process, no formal rules of inference are applied. However, experimental evidence also goes against assuming that people use mental truth-tables, complete construction of all possible models of the premises, and systematic elimination.

In non-formal deductive inference problems, evidence is that people do not use truth-tables, substitution of truth values for premises, or inference schemata (as in "natural deduction") at all. They reason by constructing a representation of the events described by the premises based on the meanings of the premises (including the meanings of the logical connectives appearing in them), on context, and on general knowledge. If their conclusion is challenged, they look more closely at the meanings of the premises and on their present interpretation of them and try to construct alternative interpretations of the premises to see if the conclusion can still be maintained. The representations that people use are more likely to resemble a perception or conception of the events than a string of symbols directly corresponding to the linguistic form of the premises.

The logical properties of connectives derive from their meanings. When, e.g., people interpret a conditional, they do not add its antecedent to their stock of beliefs, and then evaluate its consequent: they simply do not have ready access to their stock of beliefs. It is more plausible to assume that they use the beliefs "provoked" by their interpretation of the conditional to construct a mental model of a scenario in which the antecedent is realized, and then interpret the consequent with respect to that model or to a scenario based on it. A mental model based on the antecedent of a conditional is a fragment of many "possible worlds": it is consistent with many alternative complete specifications of how the world might be, because many propositions will be neither true nor false in the fragment. "If" is a verbal cue to consider (i.e., construct mental models of), in context, possible, hypothetical, or imaginary situations.

Theories of the Syllogism:

Any theory of reasoning should be evaluated according to (at least) the following criteria:
1. The theory should account for the evaluation of conclusions, the relative difficulty of different inferences, and the systematic errors and biases that occur in drawing spontaneous conclusions.
2. The theory should explain the differences in inferential ability from one individual to another.
3. The theory should be naturally extensible to related varieties of inference rather than apply solely to a narrow class of deductions.
4. The theory should explain how children acquire the ability to make valid inferences.
5. The theory must allow that people are capable of making valid inferences, that is, they are potentially rational.
6. The theory should shed some light on why formal logic was invented and how it was developed.
7. The theory should ideally have practical applications to the teaching of reasoning skills.

One theory of syllogistic inference is that non-logicians form topological mental models which are isomorphic to Euler circles and that they base their conclusions on combined models of the premises (Erickson 1974). This often generates a combinatorial complexity which is too large for subjects to keep in working memory, and the theory therefore assumes that subjects often only construct one, random, topological representation of the premises. The theory does not explain, Johnson-Laird argues, why subjects often falsely maintain that no valid conclusion can be drawn from the premises. A simple extension of the theory would seem able to do this, however.

Sternberg (Guyote and Sternberg 1978) proposed that subjects represent the Euler premises accurately and completely in symbolic form. One among several problems with this theory is that it assumes errorless operation by subjects in some very complex symbol manipulations. Furthermore, Sternberg admits that people may often be using spatial models instead of symbolic ones.

Venn diagrams are likewise topological representations of premises and they are superior to Euler circles in that they include a systematic method of making sure that one has considered all the different ways of combining the representations of premises. All the possible combinations of the premises are represented as distinct areas in a *single* diagram. Venn diagrams (like Euler circles), claims Johnson-Laird, are not "natural" mental models, i.e., they are remote from the perceived structure of situations. Neither Euler circles nor Venn diagrams can handle premises with more than one quantifier (such as "Everyone loves someone").

Johnson-Laird's theory of the syllogism is proposed as a special case of a more general theory of reasoning. A standard example of Johnson-Laird's scheme for representing mental models is the following, which represents the premises "All the artists are beekeepers" and "All the beekeepers are chemists":

```
artist = beekeeper    = chemist
artist = beekeeper    = chemist
artist = beekeeper    = chemist
         (beekeeper) = (chemist)
         (beekeeper) = (chemist)
                       (chemist)
```

The number of tokens is arbitrary. Parentheses indicate possible existence. This representational convention ensures that each premise requires only a single mental model and thus makes it possible to avoid the combinatorial explosion caused by the use of Euler circles. The psychological assumption here is that humans use some kind of representational device (but presumably not parentheses) to represent possible existence in order to map one premise into just one mental model. It is easy to see that one valid conclusion is that all the artists are chemists. In some cases, there is only one possible integrated model (given the arbitrariness of the number of tokens in a model). In many cases, says Johnson-Laird, two or (at most) three different models of this type are needed to represent different possible interpretations of the premises. Given a slight extension of the use of parentheses above, however, it seems possible to contain the representation of any one set of syllogistic premises within just one model. But Johnson-Laird claims that such an extension does not obviate the need to consider the different possible models of a set of premises and hence are only notational variants of his own. For instance, given the use of parentheses (representational devices) accepted as being psychologically realistic by Johnson-Laird, the premises "All of the B are A" and "None of the B are C" require the construction of three different models. However, just one model would suffice, if we represent the premises as follows:

```
c   (= a)
c   (= a)
------
    b =  a
    b =  a
         (a)
```

In this example, the extension of the interpretation of the parentheses used is simply that the parentheses above the broken line (another representational convention used by Johnson-Laird) should be interpreted so that both, one, or neither of the identities stated may apply. If parentheses are conventional stand-ins for representational devices signalling possible existence, then this is not really an extension of their significance at all. The situation, then, is no different from that of representing the premise "All the artists are beekeepers" as one mental model rather than as two distinct models. One model is sufficient, namely:

```
artist = beekeeper
artist = beekeeper
artist = beekeeper
         (beekeeper)
         (beekeeper)
```

This point limits the validity of some of Johnson-Laird's conclusions (see below).

Johnson-Laird (1983) proposes the following **theory** of syllogistic inference:

Step 1. Construct a mental model of the first premise.
Step 2. Add the information in the second premise to the mental model of the first premise, taking into account the different ways in which this can be done.
Step 3. State a (non-trivial) conclusion to express the relation, if any, between the 'end' terms that hold in all the models of the premises.

A more comprehensive, three-step formulation containing the notion of "scanning" of a mental model (Johnson-Laird 1989) is:

Step 1. Construct a mental model of the state of affairs described in the premises, taking into account any relevant general and specific knowledge.
Step 2. Formulate a novel, putative conclusion based on a scanning and subsequent description of the model constructed.
Step 3. Search for counterexamples, i.e., alternative models of the premises, to the putative conclusion.

Johnson-Laird (1983) appends three **hypotheses** to the theory:

(1) The more alternative models of the premises that have to be constructed in order to ensure valid inference, the heavier the load on working memory and the more errors should be made.
(2) Some syllogistic figures make it harder to integrate premises *via* the middle term than others and hence should lead to more errors.
(3) It may be assumed that the natural order in which to state a conclusion is the order in which the terms were used to construct a mental model of the premises. For example, premises in the A-B-B-C figure - the one in which integration is easiest - favours conclusions of the A-C type rather than of the C-A type.

The following comments may be added at this point:
The present author does not use models like the above in syllogistic reasoning. This point may be trivial in that Johnson-Laird might be the first to admit that his way of representing mental models involves a number of conventions (parentheses, broken lines, tokens symbolically expressed, etc.) that only symbolize, but do not resemble, the representational devices actually used in mental models. Indeed, given the whole tenor of mental model theory, Johnson-Laird's way of presenting syllogistic premises is almost provokingly symbolical. Instead, the present author seems to be using at least something like: (a) containers with a finite (or empty) set of tokens in them; (b) the notion that some tokens are inside and some are outside the containers in the model; (c) the notion of possible tokens, i.e. tokens that may or may not exist inside or outside containers. However, since considering "may's" and "may-not's" does not lead to valid conclusions it is possible that expert reasoners have learnt to completely disregard them . Clearly, we

still need (1) an adequate abstract listing of the elements needed in mental models of syllogisms, and (2) ideas as to what these elements actually "look like" when implemented as mental representations. That mental models contain tokens is part of an answer to (1); that mental models involve containers is part of an answer to (2). Johnson-Laird's symbolic scheme for presenting syllogisms is merely an illustration of the use of such elements and seems to be rather less "natural" than, e.g., Venn diagrams.

Moreover, as we saw, that scheme is inconsistent in the following respect: Johnson-Laird cannot consistently claim both that one model is sufficient for the representation of any one premise and that two or even three distinct models are needed to represent certain combinations of premises. It follows that the notion of a certain number of different mental models of a given set of premises which *have* to be considered in order to make sure to have drawn valid conclusions has not been demonstrated. It is possible that both expert reasoners and others manipulate just one model in any case and do so in ways that are not clarified by Johnson-Laird's theory. Hypothesis (1) above (p. 6) is therefore dubious. Similarly, Johnson-Laird has no convincing account of the order in which the (claimed) different mental models of a given set of premises are constructed by subjects. He claims, for example, that in modelling the premises "All of the B are A" and "Some of the C are B", subjects construct a model of the second premise and then make a renewed interpretation of the first premise in order to make the middle term integration easy. This is arbitrary. People would seem just as likely to simply construct a container model of the first premise and then integrate the second. Once the model of the first premise is there, it is no problem to augment it with the contents of the second premise. Johnson-Laird's account of the order of mental model construction takes too much for granted from his peculiar way of representing mental models of premises. As long as we do not know what mental models of premises actually look like, or even whether they look alike from one subject to another, it makes little sense to propose detailed representational theories of figural effects. Furthermore, there may be all sorts of semantic effects from the domains considered which are not accounted for in terms of formal order effects anyway.

This is not to deny the intuitively obvious point that different syllogisms with different combinations of "all", "some", and "none" have different internal complexity and hence are more or less difficult to construct and manipulate in terms of mental models of the premises. Nor is it to deny that subjects often respond on the basis of a wrong (i.e., non-representative) model of the premises. Subjects may often jump into forming a non-representative model of two premises because of the difficulty of combining them into one model. And the effort to free themselves from this interpretation and consider other possible interpretations may exceed their working memory and model-manipulating capacities. What is not obvious is that syllogism complexity is a product of the number of different mental models which have to be constructed. The syllogistic task is a very peculiar one. It is not evident that difficult syllogisms should be classified as ambiguous discourse, which clearly does call for several different mental models for its interpretation (see below). Syllogistic discourse is not ambiguous. On the contrary, to understand syllogistic premises one has to understand that quantifiers are used in a technical sense which is different from the way these quantifiers are understood in ordinary discourse. This is particularly true of the quantifier "some". It is quite possible that some of the difficulties evident in naive syllogistic reasoners arise from the fact that they interpret occurrences of "some" in ways which are not permitted by the technical vocabulary of syllogistic reasoning. For instance, if "some" in the premise "Some members of staff are members of the Concervative Party" were taken to imply, as it well might be in many cases of ordinary discourse, that there are also members of staff who are not members of the Conservative Party, then the naive syllogistic reasoner may well start off on the wrong foot, may have to deal with ambiguity, and may have to construct several alternative mental models of this premise.

Johnson-Laird's recent (1989) position is still not satisfactory. He admits that there may be important individual differences in the ways that people form mental models and operate on them: some may construct an initially misleading model and then revise it, whereas others may discover the existence of different possible models (sic) from the outset. The precise number of distinct models that a subject constructs on any occasion is uncertain. Unless subjects are using visual images, they have no conscious access to the mental models they are using. Little is known about the nature of the processes that generate counterexamples. Without a training in logic, ordinary people do not have a simple standard procedure for dealing with syllogisms.

Thus, rather than supporting a detailed theory of the abstract elements, the mental mechanics, and the mental implementation of syllogistic inference, Johnson-Laird's discussion supports some general features of mental models which will be presented below.

Mental Models and Inference in General:

In discourse comprehension, we make many implicit (rapid, effortless, outside conscious awareness) inferences which are usually deductively invalid though plausible because of utterance meaning, context, and general knowledge. The hypothesis is that we do so in the process of constructing a default single mental model of the discourse. We only search for an alternative model if this becomes necessary because of new information. Implicit inference differs from explicit inference in that only in explicit inference there is a *deliberate* search for alternative (or revised) models of the discourse that may falsify putative conclusions. Explicit inferences based on mental models do not need to make use of a formal mental logic (deduction rules, inference schemata, etc.). "Natural" mental models can represent the content of any sentence for which the truth conditions are known. Any form of deductive reasoning in a finite domain can be based on semantic procedures for constructing, interpreting, and manipulating mental models. Mental models are easily constructed for statements containing relations and quantifiers and it seems clear that the logical properties (like, e.g., transitivity) of such expressions emerge directly from their semantics rather than being an explicit part of that semantics.

In conclusion: mental models provide a basis for representing premises, and their manipulation makes it possible to reason without logic. The search for alternative interpretations of the premises, however, requires an independent representation of the premises as a kind of short-term memory source for the construction and manipulation of mental models and that representation is, Johnson-Laird claims, in some sense to be clarified, propositional.

2. Mental Models in Discourse.

Arguments against Model-Theoretic Semantics (MTS):

- MTS cannot handle the semantics of sentences about propositional attitudes since it does not accept the notion that the human mind is acting as an intermediary between language and models (or the world);
- meaning postulates are superfluous: once we have a semantic interpretation function which introduces the entities referred to by terms into a mental model, then the contents of the semantic interpretation are sufficient to constrain the interpretation of lexical items;
- the relations between expressions and the model-structures of formal semantics are very different from their relations to the world;
- MTS does not provide an analysis of the meaning of words;
- the mind is finite but there are infinitely many possible worlds in which a given assertion would be true. The mind cannot possibly represent and manipulate that infinity.

Arguments against Three Theories of the Meanings of Words:

Theory 1. Semantically complex words are not represented in a mental dictionary in such a way that, during comprehension, their meanings are decomposed into more primitive semantic components such as 'semantic markers' (Katz and Fodor 1963) which may define a word in terms of necessary and sufficient conditions, in terms of prototypes, or in other ways. It is not clear from experimental evidence that the decomposition of the meanings of words required by this theory actually takes place during language comprehension.

Theory 2. A second theory of the mental representation of the semantics of words is in terms of semantic nets: the meaning of a word is a set of acquired verbal associations which involve a number of different sorts of associative link, including class-inclusion, part-whole, property-of, and variable relations specified by a third defining word (Collins and Quillian 1972). The theory fails to predict observed differences within categories. The empirical import of the theory is unclear, since there are no principled constraints on the processes that can be employed in setting up or interrogating semantic networks.

Theory 3. Meaning postulates are used in formal semantics to specify necessary relations between predicates but have also been proposed as a theory of the mental representation of the meaning of words. The claims are that there are no semantic primitives to be used in the decomposition of the meanings of words and hence that there are no mental dictionary entries representing the meanings of words (Fodor, Garrett, Walker, and Parkes 1980). There is just strings of unanalysed tokens in *mentalese* + meaning postulates on some analytic relations between the meanings of words. No adequate definitions of the meanings of words exist.
Against this position, Johnson-Laird argues that:

- the meanings of most English words derive solely from definitions;
- the meanings of prototypes and stereotypes (schemata) can be articulated;
- the meanings of (relatively) more semantically primitive words are difficult to express using other words. In general, the meanings of such words have to be acquired by other means (ostensively, through context) and such words tend to have a greater diversity of meanings. But the less semantically primitive a word is, the easier it is to define using other words;
- it seems clear from experimental evidence that semantic information is mentally organised in the form of dictionary entries.

Theories (1)-(3), though of some use to formal and machine implementational theories of language processing, do not account for how language is related to the world and hence only create an illusion of semantic significance: they are not theories of cognitive semantics. They jointly assume that the meanings of words (intensions) are autonomous and do not have to be understood on the basis of an understanding of their real-world extensions or their reference.
Arguments against this assumption are:
- reference (context) and world knowledge and inferences based on these (and not only selectional restrictions - cf. theory (1) above) are used in word disambiguation;
- context constrains the interpretation of words;
- logical inference based on descriptions cannot solely be based on meaning postulates or other autonomous intensional machinery but also requires knowledge of context;
- metaphorical and other non-standard uses of language violate the autonomy of intensions which works best with "literal meaning".

The theories (1)-(3) do not attempt to *represent* the truth conditions of sentences and discourse. They only represent their logical structure by means of some propositional representation plus the semantics of lexical items through semantic decomposition, semantic networks, or meaning postulates. This apparatus is assumed to take care of the inferences made during discourse. The procedural construction of mental models and hence to some extent of their truth conditions, on the basis of the mental lexicon and world knowledge makes further assumptions concerning meaning postulates in order to represent, e.g., transitivity or other formal relationships needed for capturing entailments, superfluous (see below). From the mental model representation of the truth conditions of discourse, further semantic properties emerge naturally without the need for a mental logic. Such properties may be more flexible than standard mental logic ones: the scanning of a mental model of people sitting around a round table will, e.g., easily establish the limited transitivity of the relation "next to".

The meanings of many words are mentally represented as prototypes or stereotypes, i.e., schemata of characteristic instances defined by default values. The (often fuzzy) boundaries of a word's intension and extension are set by the taxonomy in which the schema occurs. Many semantic fields have a more complex structure than class inclusion, e.g., those of spatial prepositions.

Discourse Comprehension:

In discourse comprehension, mental models are constructed by processes (of psychological semantics) that map "propositional representations" (strings of symbols) into models.

In the **first stage**, superficial understanding of an utterance produces "a symbolic propositional representation close to the surface form of the sentence". This representation determines the truth conditions of the utterance, a representation of which are central to stage two. This first stage involves mental parsing. There is psychological evidence against transformational grammar and in favour of the assumption that meaning is recovered directly from surface structure: subjects remember sentences either *verbatim* or they remember only their meaning (a mental model). Furthermore, transformational rules are not necessary to the analysis of English sentences. Context-free phrase-structure grammar rules seem sufficient.

Given the ambiguity of grammars of natural languages, mental parsing has to be non-deterministic (*pace* the deterministic, look-ahead Marcus-parser; see Marcus 1980) argues Johnson-Laird, using either backtracking or the construction of a well-formed substring table of parallel analyses of a sentence, possibly combined with lookahead. Both bottom-up and top-down (predictive) parsing are probably involved as in, e.g., a left-corner parser. Since humans have difficulties in understanding self-embedded sentences, non-determinism is presumably not handled by humans using backtracking (to any greater depth) and a stack. Lookahead would also seem to be excluded since sentence comprehension at all levels starts almost immediately at sentence onset. Another possibility is that the mental parser constructs and maintains a table of possible analyses of a sentence. This leads to four constraints on the design of the mental parser:

- the parser delivers an almost immediate propositional representation of a sentence constituent by constituent, or word by word. It does not set up a representation of syntactic structure: lexical entries contain information not only about the most likely syntactic frame, but also about potential referents for the different arguments. Selectional restrictions, default values, and factual information for the use of an inferential component may all be used to make predictions, and sentences that conform to them will be easier to interpret than those that do not. The propositional representation of one constituent could help determine the proper analysis of other constituents. The identification of the referents of expressions could influence the process of parsing. The implausibility of the interpretation of a sentence as a whole may lead to the rejection of its propositional representation;
- the parser uses semantic information from several sources to help it parse (prior context, meaning and reference of previous constituents, general knowledge);
- the parser uses both top-down and bottom-up procedures, perhaps integrated within a left-corner parsing system;
- the parser copes with local ambiguity arising from dislocated constituents either by maintaining a table of possible analyses, or by reparsing the ambiguous constituent. It does not make any systematic use of either backtracking or lookahead.

The interpretative process is syntactically driven. The parser uses the grammar to build up a propositional representation based on the lexical semantics. Whenever it has a choice, it can be guided by further semantic information in the lexicon, by the mental model of discourse, or by general knowledge (for examples of proposed heuristics used by the mental parser, see Johnson-Laird 1983 p. 332). The parser constructs the propositional representation as a semantically interpreted tree.

In the **second stage**, the propositional representation is used as a partial basis for constructing a mental model whose structure is analogous to the state of affairs described by the discourse. The relevant context of an utterance can be represented in the mental model, and the "significance" of the utterance is established by relating its propositional representation to this model and to general knowledge. This process, Johnson-Laird says, may occur clause by clause, or constituent by constituent, rather than at the level of complete sentences. A model goes beyond the literal meaning of discourse, because it embodies inferences, instantiations, and references, so that the meaning of the sentence is not recoverable from the model. The mental model is constructed on the basis of the truth conditions of the propositions expressed by the sentences in the discourse. The truth conditions of the proposition expressed by a sentence depend on the meaning of the sentence, its context of utterance (as represented in the current mental model), and the implicit inferences that it triggers from background knowledge.

Assumptions about stage two, i.e., the procedural semantics are:

- the processes by which fictitious discourse is understood are not essentially different from those that occur with true assertions;
- in understanding a discourse, one constructs a single mental model of it. A mental model is a single representative sample from the (possibly indefinitely large) set of models satisfying an assertion;
- the interpretation of the discourse depends on both the model and the processes that construct, extend, revise, and evaluate it;
- the processes that construct, extend, evaluate, and revise mental models, unlike the interpretation functions of model- theoretic semantics, cannot be treated in an abstract way;
- a discourse is true if it has at least one mental model that satisfies its truth conditions that can be "embedded" in the model corresponding to the world.

Procedures used in stage two include (Johnson-Laird's formulation in boldface):

1. A procedure that begins the construction of a new mental model based on the propositional representation and its truth conditions (introducing - sometimes arbitrary or default or prototypical - semantic structures by using the system's mental lexicon and knowledge representation of the domain as elicited by the discourse) **whenever an assertion makes no reference, either explicitly or implicitly, to any entity in the current model of discourse** (as established from scanning of this model)**.**
2. A procedure which, if at least one entity referred to in the assertion is represented in the current model (possibly determined through scanning of the model)**, adds the other entities, properties, or relations to the model in an appropriate way** (using the system's mental lexicon and knowledge representation of the domain as elicited by the discourse. This procedure may note (depending on the difficulty of this task and on how discriminatively the discourse is attended to) if there are (unused) alternative possibilities without necessarily specifying these completely. The system assumes by default that the speaker intends to communicate one consistent mental model and that its task therefore is to reconstruct this model on the basis of the speaker's communications).
3. A procedure that integrates two or more hitherto separate models if an assertion interrelates entities in them (by scanning models and using the system's mental lexicon and knowledge representation of the domain as elicited by the discourse).
4. A procedure which, if all the entities referred to in the assertion are represented in the current model, verifies (possibly through scanning of the model) **whether the asserted properties or relations hold in the model (and adds what has not been included so far - cf. 2).**
5. A revision procedure checking whether an assertion discovered (possibly through scanning of the model) **to be false of the current model can be rendered true by recursively modifying the model in a way consistent with the previous assertions. If not, then the assertion is inconsistent with the previous discourse.** (If it can, then the model is non-monotonically revised accordingly).

6. A (surely optional, mainly for solving the specific task of valid deductive inference) **revision procedure checking whether an assertion true of the current model can be rendered false by changing the model in a way consistent with the previous assertions. If not, then the assertion is a logical consequence of the previous assertions.**
(We may add the following, cf. the section on discourse below:
7. A set of procedures which evaluate the truth or plausibility of the model (at least given non-fictitious discourse) with respect to world knowledge from perception or memory. In some cases, this evaluation may lead to renewed search for alternative interpretations of the discourse).

The mechanisms used in stage two are illustrated through a purportedly psychologically realistic computer program which constructs configurations in spatial arrays using the above procedures (Johnson-Laird 1983 p. 252 ff.). For example: the mental lexicon contains procedures (simulated in the computer model in a way which is admittedly simplified and too rigid) for constructing the relation expressed by the preposition "in front of" between entities in the mental model. This may be, in fact, an example of the procedural use of lexically represented image schemata (Lakoff 1987) in mental model construction and discourse comprehension although image schemata are not included in Johnson-Laird's account of the lexicon. The example illustrates how the semantics of terms like "left" or "right" that are indefinable through other terms, are procedurally defined through the construction of a mental model of the truth conditions of discourse using, Johnson-Laird assumes, procedural primitives.

Given that children do have the ability to construct mental models, what they have to learn in order to have learnt the meaning of a word is its contribution to the truth conditions of sentences. Having done this, they will implicitly have mastered the word's logical properties.

Discourse Theory:

The basic problem about **story grammars** is that we do not have effective procedures for categorizing parts of discourse into the basic "syntactic" and "semantic" categories of such grammars such as "setting", "event", "reaction", "causes", "motivates", "initiates", "episode", "internal response", etc. Hence story grammars do not possess any great explanatory value: they rely upon the semantic intuitions of the grammarian and their use of context-free syntactic rules is not clearly motivated.

A necessary and sufficient condition for the coherence of discourse is the possibility to construct a single mental model from it, says Johnson-Laird. This proposal may be circular, however, as long as it is not clear that people are unable to construct incoherent mental models. The possibility of constructing a single mental model of a discourse depends on the principal factors of co-reference and consistency. Each sentence in a discourse must refer, implicitly or explicitly, to an entity referred to (or introduced) in another sentence, since this is a precondition of representing the discourse in a single integrated mental model. The coherence of prose depends primarily on its pattern of co-reference. The properties and relations ascribed to referents must be consistent, i.e., compatible with one another and free from contradiction.

Plausibility is different from coherence, since a discourse can be coherent yet highly implausible. Plausibility depends on the possibility of interpreting the discourse in an appropriate temporal, spatial, causal, and intentional framework. In the construction of mental models, subjects make use of cues about both coherence and plausibility. Scripts are probably used in many cases in order to judge the plausibility of discourse. But we are also able to understand discourse about events that are not stereotyped. More work is needed in order to understand the rapid retrieval of relevant information underlying the plausibility of discourse. So we need theories of relevance and plausibility.

Many aspects of reference seem to depend on the distinction between propositional representations and mental models:

(a) Definite and indefinite descriptions. A prototypical shop visited by a customer, for instance, contains at least one assistant. When the first sentence introduces the shop, therefore, the second sentence may safely refer to "the assistant" without this token having been explicitly introduced before.
(b) Referential and attributive (definite) descriptions. Such descriptions clearly demonstrate that discourse has at least two contexts: one for the speaker and one for the listener. The use of a definite description can be referential for the speaker but attributive for the listener, or vice versa.
(c) Pronouns. There are at least five seemingly different uses of pronouns: 1. deictic, 2. anaphoric (and cataphoric), 3. following a quantifier and behaving like a bound variable, 4. what Johnson-Laird calls "Evans-pronouns", 5. pronouns replacing earlier expressions in a sentence. Johnson-Laird claims that the behaviour of pronouns can be explained on the assumption that discourse has two levels of representation. In all cases, the referents of pronouns are recovered on the basis of the mental model constructed of the discourse using syntactic, perceptual, and all the other kinds of cues, as appropriate.
What is not evident, however, is how much of an argument (a)-(c) constitute for the assumption that discourse comprehension happens in two stages, the first being the construction of a "propositional" representation.

Discourse is the communication of a single mental model between the participants. A description of a single state of affairs is represented by a single mental model even if the description is incomplete or indeterminate. Grice's (1975) remarks on conversation can be seen as stating rules that facilitate the communication of a single mental model between the discourse participants.

3. Mental Models in General.

According to the theory under consideration, there are at least three different kinds of mental representation: propositional representations, mental models, and images. These kinds are functionally and structurally different and it can be experimentally determined which kind of representation a person is using on a specific occasion. Mental representations may differ widely in their content, as do models of different systems, models of different tasks, discourse models of all kinds of topics, and so on. But there is no evidence that they differ in representational format or in the processes that construct and manipulate them.

Images: There is a continuing debate as to whether mental images are either (a) epiphenomenal, i.e., do not contain any new information over and above the information contained in the propositional representations that encode them and which also encode the input from perception (Fodor, Pylyshyn, and many others), or are (b) mental representations that do contain more information (which is in analogue form) than what has been propositionally encoded and which have to be mentally scanned in order to retrieve that information (Shepard, Kosslyn, and many others). Johnson-Laird argues that, in one sense of "propositional representation", it is a physiological (machine-code-like) claim that there are only propositional mental representations. In another sense, the conflict is real. Images exist and are just as high-level representations as beliefs, and hence surely can be cognitively penetrated by beliefs (cf. Pylyshyn 1983). Images correspond to views of mental models from a particular point of view, that is, images correspond to perception of the world and are based on mental models of the world akin to those 3-D, object-centered models constructed in the course of, stored from, and used in, visual perception (Marr 1982).

Propositional representations are mental representations of the sense or meaning of verbally expressible propositions. Such mental representations are strings of symbols which "resemble" natural language and which thus have a vocabulary related to that of natural language, an arbitrary syntax, which will probably remain unknown, and a semantics. A propositional representation represents a function from states of affairs to truth values. It should be clear by now that Johnson-Laird's notion of propositional representations is not very informative.

Mental models have an extremely general area of application. This follows from the above, simple tripartition of mental representations into three categories. All knowledge of the world depends on the ability to construct mental models, claims Johnson-Laird. Mental models are structural analogues of the world, that is, they have a structure which is analogous to the structure of states of affairs (objects, events, processes, actions, etc.) in the world as perceived or conceived. The semantics of the mental language (the propositional representations) maps propositional representations into mental models of real or imaginary worlds, i.e., propositional representations are interpreted with respect to mental models. Mental models enable people to make inferences and predictions, to understand phenomena, to decide what action to take and to control its execution, and to experience events by proxy; they allow language to be used to create representations comparable to those deriving from direct acquaintance with the world; and they relate words to the world by way of conception and perception. There are no complete mental models for any empirical phenomena. Mental models are radically incomplete.

Most of the following **assumptions** should be straightforward from what has already been said:

1. Mental models, and the machinery for constructing and interpreting them, are computable.
2. A mental model must be finite in size and cannot directly represent an infinite domain. Though finite in size, mental models are capable of representing an infinite number of different possibilities. The theory of mental models is compatible with a model-theoretic semantics for finite domains. Mental models, like images, are highly specific (so there cannot be, e.g., a mental model of a triangle in general). This implies that very often during inference and reasoning, the use of mental models is accompanied by a characteristic representativity problem concerning how this particular mental model manages, as intended, to represent a much more general class of entities. If some description is provided then the mental model constructed on its basis is a representative sample from the set of possible models satisfying the description.

3. A mental model is constructed from tokens arranged in a particular structure to represent states of affairs. The structure of a (correct) mental model corresponds to the structure of the situation that it represents.
4. A description of a single state of affairs is represented by a single mental model even if the description is incomplete or indeterminate.
5. Mental models can directly represent indeterminacies if and only if their use is not computationally intractable, i.e., iff there is not an exponential growth in complexity.
6. All conceptual primitives, that is, the conceptual primitives from which all mental models are constructed, are innate. Johnson-Laird *seems* to be speaking here only of procedural primitives, i.e., the primitives underlying perceptual experiences, motor abilities, and cognitive skills.
7. There is a finite set of conceptual primitives that give rise to a corresponding set of semantic fields (like shape, colour, person, kinship, motion, perception, cogitation, emotion, bodily action, possession, communication - the furniture of the world) which are reflected in the lexicon by a large number of words sharing a common concept at the core of their meanings; and there is a further finite set of concepts, or 'semantic operators' (like time, space, possibility, permissibility, causation, intention - relations between the furniture of the world), that occur in every semantic field serving to build up more complex concepts out of the underlying primitives. Nearly all complex concepts corresponding to words can be constructed from simpler concepts by the operation of composition.
8. The structures of mental models are identical to the structures of the states of affairs, whether perceived or conceived, that the models represent. So mental models differ from truth tables, Euler circles, Venn diagrams, semantic networks, predicate logic formalisms, (and Johnson-Laird's own illustrations), which all have structures that are not identical to the states of affairs they represent. Furthermore, semantic networks and predicate logic formalisms need to be interpreted: they have no machinery for assigning a truth value to an assertion.

Mental models owe their origin to the evolution of perceptual ability in organisms with nervous systems. Perception provides us with our richest model of the world. A primary source of mental models - three-dimensional kinematic models of the world - is perception. As remarked above, a basic example of mental models are those (not very well known, for the time being) 3-D, object-centered models constructed in the course of, stored from, and used in visual perception. The use of mental models in interpreting language and in making inferences is a natural extension of their perceptual function: if perception of the world is model-based then discourse about the world must be model-based too. The major constraints on mental models derive from the perceived and conceived structure of the world, from the conceptual relations governing ontology, and from the need to maintain a system free from contradictions.

With increasing expertise in a particular domain, people develop richer mental models of that domain. Johnson-Laird's theory of mental representation does not include any distinction between the types of representation involved in skill-, rule-, and knowledge-based performance, respectively. Input-output rules for system behaviour, task-action models of system handling, and models used in diagnosing system malfunction all seem to be mental models. They are different, and some are richer than others, but they are all considered mental models of the systems involved. Typically, in scientific domains, novices reason qualitatively on the basis of mental models simulating objects, events, and processes in real time, whereas experts use more abstract mental models representing abstract properties and relations and able to support quantitative reasoning. Mental models of systems and domains can be useful for, e.g., many purposes of ordinary life, although they are incomplete or inaccurate. Examples of inaccuracies are the many cases of wrong naive physics where people's models lead to the prediction of physically impossible sequences of events.

Typology of mental models (admittedly informal and tentative):

(a) Physical models which represent perceptible situations but which cannot represent either abstract relations or anything other than determinate physical descriptions:

1. Simple relational models: static "frames" consisting of finite sets of tokens, relations between the tokens, and properties representing physical entities, their relations and their properties, as in standard examples of syllogistic premises.
2. Spatial models in which the only relations between tokens are spatial and are represented in 2-D or 3-D.
3. Temporal models consisting of a sequence of spatial "frames" that occur in a temporal order corresponding to the temporal order of events.
4. Kinematic models are psychologically continuous temporal models.
5. Dynamic models are kinematic models in which there are representations of causal relations between the depicted events.
6. Images produced by visual imagination but otherwise similar to Marr's viewer-centered 2 1/2-D sketches. This is clearly putting too much weight on Marr's contribution (cf. the section on images above). The distinction between images and mental models, it seems, is primarily meant to block the objection that the mental model theory requires that all thought be pictorial. Johnson-Laird's reply is that we may be handling mental models (in, e.g., discourse comprehension) even if we are not aware of doing any imagery. Mental model theory, that is, does not represent a revival of the "imageless thought" controversy early this century. Essential though it may be to make this point, it does nothing to substantiate Marr's distinction between 2 1/2-D, viewer-centered representations of objects and 3-D, object-centered representations of objects for which Marr does not offer much psychological evidence. The primary "cash value" of the distinction between mental models and images is the distinction between representations using imagery and representations doing without imagery, and that distinction has nothing to say about the nature of the mental models active when we are not using imagery.

(b) Conceptual models:

Conjunction is represented through co-presence within a model.
Negation is represented through some kind of annotation on models or parts of models indicating non-existence of what is represented.
Disjunction is represented through relations between models or parts of models indicating that one or more of the models or model-parts exist.
Conditionals are handled through constructing a scenario in which the antecedent is realized and then interpreting the - separately constructed - model of the consequent with respect to the particular type of conditional used.
Quantifiers are always represented by finite sets of mental tokens.
Identity and non-identity are represented symbolically.
Uncertainty as to the existence of entities (etc.) of a particular type is symbolically represented.

Mental models do not contain variables. The assumption seems to be that information provided through the use of any kind of variables gets interpreted using default information about the discourse domain.

1. Monadic models represent assertions about individuals, their properties, and identities between them.
2. Relational models introduce a finite number of relations between tokens in the monadic model.
3. Meta-linguistic models contain tokens corresponding to linguistic expressions, certain abstract relations between them (like refers to, means, is true), and elements in mental models of any type.
4. Set-theoretic models contain a finite number of tokens directly representing sets and possibly also finite sets of associated tokens designating the abstract properties of a set, and a finite set of relations between the tokens designating sets.

The machinery for embedding one mental model within another can account for the semantics of propositional attitudes. A propositional attitude is a relation (of belief, hope, or thought, etc.) between an individual and that individual's mental model of the relevant state of affairs.

This typology of mental models is not very informative. The reason is simply that since mental models are supposed to be used in representing everything, a typology of mental models will have to be a typology of everything. Philosophers have been trying to do just that through millenia and from that point of view Johnson-Laird's list is just one more attempted sketch.

Mental models may represent true situations (courses of events or scenarios, etc.), possible situations, or imaginary situations. A constructed mental model may be imagined more or less vividly with lots of prototypical detail or it may not be imagined at all. When used in reasoning mental models may or may not be accompanied by imagery and/or propositional representations.

Johnson-Laird argues (1983 p. 425) that it is unlikely ever to be discovered how tokens representing entities are represented in the mind. Similarly, the recursive procedures used in the construction of mental models from propositional representations and the procedures used in the construction of an image of the mental model from the mental model, are ineffable (p. 446) or "tacit". The syntax of the mental language will probably never be known. The structure of the concepts on which cognition depends is not open to conscious inspection. One might express the hope that this position is overly pessimistic. The tripartition of mental representations into images, propositional representations, and mental models may be spurious, and there may be no mental language and hence no syntax of this language to discover. The same may be true of the construction of mental models from propositional representations. But if there are mental models at all having some of the core properties ascribed to them by the theory, then it might be possible to come closer to understanding their psychological implementation and the processes operating on them. One basic approach in this endeavour, according to the theory itself, is through an increased understanding of vision and other sensory modalities. Johnson-Laird (1989) says so much himself, namely that mental model theory is incomplete and that too little is known about 3-D model formation in vision, about the construction of discourse models and models of the truth conditions of expressions, and about model-based reasoning.

Truth: According to the Discourse Representation Theory of Hans Kamp and others, a text represented in a discourse model is true if and only if there is a mapping of the individuals and events in the discourse model into the real world model in a way that preserves their respective properties and the relations between them. Discourse models mediate between language and model-structure in order to provide a fuller account of the truth conditions of connected discourse. Kamp's early discourse models remain, however, abstract idealizations, argues Johnson-Laird (1983). They are, for instance, formulated so that they never have to be revised in the light of subsequent information in the discourse. Defining truth for mental models requires the combination of Kamp's notion of an appropriate mapping and the idea that a mental model is a representative sample from an infinite set of possible models. Any member of the set could in principle be generated by using the recursive procedures for revising the representative sample. A mental model represents the extension or reference of an assertion or a discourse, i.e., the situation it describes; and the recursive machinery for revising the model represents together with the initial linguistic representation the intension or meaning of the assertion or the discourse, i.e., the set of all possible situations it could describe. A discourse is true if and only if there is at least one mental model of it which can be mapped (or "embedded") into the real world model in a way that preserves the content of the mental model. This can be established by, e.g., visual perception. "Embedding" means that the same individuals with the same properties and relations are preserved from one model to the other.

Some evidence of the distinction between propositional (or linguistic) representations and mental models:

- Subjects tend to form mental models of spatially determinate descriptions but not of spatially indeterminate descriptions consistent with more than one spatial layout, and they remember the gist of the former much better than of the latter. On the other hand, subjects remember the exact wording of spatially indeterminate descriptions better than that of determinate descriptions. This may be because they construct a mental model of the latter but refrain from constructing a finished mental model of the former in favour of committing an indeterminate description *verbatim* to memory as soon as they encounter some indeterminacy which would otherwise require the construction of a number of alternative possible mental models. Alternatively, subjects may choose to represent a particular one among the mental models made possible by the indeterminacy on the risk that it is wrong and subsequently has to be revised (if subjectively possible by then), or they may construct some sort of hybrid between propositional representations and mental models - an (propositionally) "annotated" mental model;

- mental models are more easily remembered than propositions, perhaps because they are more structured and elaborated and require a greater amount of processing to construct. If a lengthy description of some entity does not give rise to a single mental model then the description and its contents are soon forgotten;

- mental models do not preserve the propositional representations on which they have been based and thus subjects tend to confuse those representations with propositions inferrable from them. Similarly, subjects tend to confuse propositions having slightly different meanings as long as they enable the construction of the same mental model. Also, subjects tend to confuse different verbalisations of one and the same proposition;

- where two expressions with different meanings occur in contexts in which they refer to one and the same individual, subjects tend to confuse these expressions;

- much work on anaphoric reference (Hans Kamp and others, Discourse Representation Theory) argues that a representation of anaphoric referents separate from their linguistic representation is needed. This is a step towards positing mental models for the understanding of discourse.

Whereas a number of the findings and theoretical developments just mentioned are rather robust and constitute a strong argument for taking mental model theory seriously, they constitute only the weakest of arguments for the distinction between linguistic surface structure, on one hand, and propositional representations (the "mental language") on the other.

Consciousness:

1. The mind employs different levels of organisation.
2. Mental processes at each level take context into account.
3. Processing at different levels is not autonomous, but interactive.
4. Mental processes occur in parallel in a hierarchy of parallel processors.
5. On top of the hierarchy there is an operating system working serially, which monitors and to some extent controls the lower-level processors. At the second level down there are (interactive) processors for perceiving, understanding, acting, remembering, communicating, and thinking.
6. The contents of consciousness are the current values of parameters governing the high-level computations of the operating system.
7. The operating system can receive values from lower-level processors, but it cannot inspect the internal operations of these processors. So these operations are necessarily unconscious, whereas the values received by the operating system (i.e., intentional states or propositional attitudes) are conscious. The specific contents of consciousness ("qualia") are ineffable as is the way in which we exercise mental skills such as learning or inference, or the underlying nature and mechanism of mental representations. One is aware of what is being represented and of whether it is being perceived or imagined, but not of the inherent nature of the representation itself. The system employs parallel taxonomic systems to handle input and output. Since these systems are inaccessible to consciousness, so is the structure of the concepts on which cognition depends.
8. The division between conscious and unconscious processes is a consequence of parallelism which ensures rapid input/output operation and graceful degradation.

These points have very little directly to do with mental model theory.

4. Discussion of the Theory.

(a) Some characteristics of mental models:

It is a common experience that a new and interesting theory turns out to be rather meager when we look into its details. Mental model theory is such a theory. It has a central, intuitively appealing core and it is an early attempt to formulate a framework for a cognitive semantics. The remainder is mainly a large research agenda or programme. The core is that a mental model resembles a perception or conception of a situation or an event, i.e., that a basic class of mental models are much closer to the models of the world created by perception than they are to the abstract, formal, and in a specific sense syntactic apparatus of formal logic and much of contemporary linguistics. Mental models basically are models of the real thing, of the world perceived and thought about. Mental models are structural analogues of the world, that is, they have a structure which is analogous to the structure of states of affairs (objects, events, processes, actions, etc.) in the world as perceived or conceived. A mental model is constructed from tokens arranged in a particular structure to represent states of affairs. There are no complete mental models for any empirical phenomena. Mental models are radically incomplete. A mental model must be finite in size. There is no problem, unsolvable in principle, about relating mental models created through reasoning or language comprehension to the world reasoned about or dealt with in discourse in order to try to verify or falsify statements and conclusions.

Some of the basic questions raised by mental model theory, therefore, are: (1) how mental models theory hooks up with our knowledge of perception. This question, as we saw, is for the future. (2) How mental models relate to language. Here, the theory has a lot to say, the core being the presentation of the principles of procedural semantics (pp. 15-6 above) and the idea that we need to postulate an ontological layer of mental representation in between language and its formal representation, on the one hand, and perception, on the other. Again, how the principles of procedural semantics are mentally implemented is a question for the future. However, the status of contemporary formal semantics in the theory is entirely unclear. A number of points like the following are being made, but they do not add up to a coherent picture:

- a model goes beyond the literal meaning of discourse, because it embodies inferences, instantiations, and references, so that the meaning of the sentence is not recoverable from the model;
- a mental model is a single representative sample from the (possibly indefinitely large) set of models satisfying an assertion;
- the processes that construct, extend, evaluate, and revise mental models, unlike the interpretation functions of model- theoretic semantics, cannot be treated in an abstract way. These processes are naturally suited to explain the ubiquity of nonmonotoniticity in human reasoning and discourse comprehension;
- the use of mental models in interpreting language and in making inferences is a natural extension of their perceptual function: if perception of the world is model-based then discourse about the world must be model-based too.

(b) The tripartition of mental representations into images, propositional representations, and mental models:

We have seen this tripartition to gradually evaporate during the discussion above. If a distinction must be made between images and mental models in terms of types of mental representations rather than in terms of the imagery/imageless thought distinction, this is for the future to do. How mental models relate to imagery is, as we saw, unresolved since Marr's theory cannot be relied upon for an authoritative answer.

As for the distinction between propositional representations and mental models, the issue seems to be: why do we need to assume that the mind builds up a propositional representation of the sentence meaning of linguistic input *in addition to* (a) preserving the linguistic surface formulation in short-term memory and (b) creating a mental model representation of "discourse significance" from it using the mental lexicon, context, and background knowledge ? There is no elaboration to be found of the notion of a "propositional representation" relating it to linguistic surface information, on the one hand, and to the representations of formal semantics on the other.

What formal semantics has been doing for quite some time is to develop increasingly adequate formal representations of the contents of propositions through unpacking the information contained in discourse in the form of an extended first order logic representation using parsing and the lexicon compositionally. It is true that this does not go all the way to representing the truth conditions of utterances, a representation of which also requires the use of contextual information and general knowledge. Somehow, all of this information has to be psychologically represented. What is unclear is whether anything like the processes described by formal semantics take place in the mind or whether formal semantics provides a theory of competence unrelated to mental processing. It is not obvious that the information processed during discourse comprehension is being simultaneously represented at two distinct levels, that of predicate logic (or something still closer to the surface forms of sentences - "a symbolic propositional representation close to the surface form of the sentence" - whatever that may be) and that of mental models. This is especially not obvious if we assume, with Johnson-Laird, that the process of passing from propositional representations to mental models may occur clause by clause, or constituent by constituent, rather than at the level of complete sentences. On the other hand, it does seem clear that the *information* spelled out in formal semantical representations of discourse is being used by the discourse participants. According to mental model theory, the construction of propositional representations involves mental parsing, which we know very little about. Another remnant from formal linguistics in the theory is that the syntax of the mental language is "arbitrary". The existence of linguistic universals and the growing number of arguments supporting a closer relationship between syntax and semantics go against this assumption. And given the uncertainty about the two-stage model of mental model construction, discussing the mechanisms of mental parsing and its construction of a propositional interpretation becomes a highly speculative enterprise anyway.

Parsimony, then, would seem to suggest that we avoid the assumption of two distinct constructions of elaborate representations of discourse, one "propositional" and one at the level of mental models.

Mental model theory is one among several attempts during the 1980's at "psychologising" model-theoretic semantics. Other attempts include Situation Semantics (Barwise and Perry 1983) with its notion of partially represented situations and, even more so, Kamp's Discourse Representation Theory (Kamp 1981) which includes a rudimentary discourse model in addition to a model of the world. Thus, mental model theory is not alone in raising the two-stage problem just discussed.

(c) The mental lexicon:

Closely related to the two-stage problem is the question about the contents of the mental lexicon. Johnson-Laird presents a rudimentary theory of the mental lexicon: it exists; it is taxonomically organised, possibly in a form which involves meaning postulates relating, e.g., fuzzy opposites like "tall" and "short"; it contains prototypes or stereotypes, i.e., schemata of characteristic instances defined by default values; semantic fields having a more complex structure than class inclusion, like, e.g., those of spatial prepositions;

simple and complex concepts related by the operation of composition; as well as symbolic procedures used in mental model construction. A more detailed theory of the inventory and use of the mental lexicon is of crucial importance to mental model theory.

The relationship between the two-stage model of discourse comprehension and the mental lexicon poses the following problem. If the two-stage model is right, does that mean that we have to assume two mental lexicons as well, one containing symbolic representations and one containing (sometimes) analogue elements for use in mental model construction ? This, perhaps, does not add to the plausibility of the two-stage model. Be this as it may, if there is some basic truth to mental model theory, there probably has to be, in long-term memory, an inventory of analogue and possibly other elements and complex structures for use in mental model construction. A central part of this inventory consists of models somehow abstracted from or otherwise derived from perception. Such elements, when they are analogue rather than symbolic, are mental model-like. Lakoff (1987) offers some early ideas on them.

But whereas mental models, according to the theory at hand are constructed on the fly during discourse comprehension, these mental models are relatively permanent structures in long-term memory. We should therefore distinguish between (relatively) *permanent* mental models and *temporary* mental models. This distinction gains in importance if we look to the HCI mental models literature (e.g. Gentner and Stevens 1983). This literature primarily studies the novice-to-expert development and use of complex mental models of systems such as various types of artefacts, computers, and complex systems controlled through computers. Such models are permanent in the above sense, they are constructed from experience and perception, but they are not elementary. So another distinction imposes itself, namely between (relatively) elementary and (relatively) complex mental models. The first are conceptual in nature, the latter are rather models of systems in a wide sense. Since system mental models can be described and communicated, what starts out as a temporary mental model constructed through discourse comprehension may subsequently become a permanent model in long-term memory (for further discussion see Wilson and Rutherford 1989, Bernsen 1991). Permanent mental models may include structures resembling frames, scripts, schemata, scenarios, naive or folk theories, etc., and the relationship between these notions, which have been proposed in many different areas of knowledge representation and mental models needs further study. Johnson-Laird's main contribution is to propose a process-oriented or computational format for the construction and use of temporary mental models beyond what has so far been proposed on behalf of active data structures such as frames and schemata (Wilson and Rutherford 1989).

(d) Mental models and rules:

Mental model theory does not assume rules of inference of any sort, either formal or content-specific, but instead assumes that reasoning depends on the manipulation of mental models. Part of the explanation for this seems to be that Johnson-Laird's version of mental model theory has been developed exclusively from the study of knowledge-based reasoning and discourse comprehension and in opposition to formal approaches. Even in these domains, as briefly mentioned already, content-specific, rule-based approaches have been proposed and the relationship between these and mental model theory needs clarification. But if we move on to taking into account the entire spectrum of human cognitive performance, then rule-based behaviour becomes much more prominent than recognized by mental model theory even though the theory is claimed to cover it all. As we saw, mental models were said to enable people to make inferences and predictions, to understand phenomena, to decide what action to take and to control its execution. Johnson-Laird's mental model theory implies that no distinction is needed between different types of mental representation involved in skill-, rule-, and knowledge-based performance, respectively (for this "SKR" framework see, e.g., Rasmussen 1990). Given the proven value of the SKR framework and in particular of the existence of wide areas of human rule-based performance, this would seem at best only partially true. We need an account of the relationship between mental models and the SKR framework.

(e) Conceptual note:

Words like "analogue", "analogical", and "symbolic" are used in different ways in the literature to describe representations. The ensuing confusion is substantial and deep-rooted and certainly cannot be resolved here. But some steps toward a clarification of the issues involved are needed in order to understand what mental models could be.
(1) It is a fact that a schematic drawing of a man and the English word "man" bear a fundamentally different relationship to what they represent. The drawing resembles real men and can be matched against perceived objects whereas the word does not and cannot. So there is a clear sense in which the drawing is an analogical representation of men whereas the word symbolically represents men. For obvious reasons, let us call a drawing an *external analogical representation* and the word an *external symbolic representation.*
(2) It is probably a fact that infants, before they learn a particular natural language and as a condition for being able to do so, learn to categorize objects in the environment including, say, men or persons, and to recognize them when they perceive them. This ability depends on abilities to store and use (in recognition, mental imagery, etc.) representations of objects. We may call these representations which are somehow derived from perception, *internal perceptual representations*. Later, when the child learns a particular natural language, words in the lexicon get associated with perceptual representations. For instance, "man" gets associated with the perceptual representation of men. To do this, the child presumably manages to create *internal symbolic representations* and associate them with internal perceptual representations.
(3) Now, the confusion, including the confusion surrounding the notion of mental models, starts to arise when we ask whether internal perceptual representations are analogue (or analogical) or symbolical. They are internal *perceptual* representations all right, and these are different from internal *symbolic* representations, but the point is that this does not answer the question whether or not internal perceptual representations are analogical. Nor is the question answered by insisting, as Johnson-Laird would seem to be doing, that mental models make essential use of internal perceptual representations rather than internal symbolic representations.
(4) Mental images could be characterised as a kind of internal analogical representations. As we have seen, Johnson-Laird hypothesizes that images are high-level representations which correspond to "views" of mental models from a particular point of view and to perception of the world, and are based on mental models of the world. This opens up the following chain of reasoning: if mental images are high-level (as opposed to the "machine code" of machines or brains) analogical representations and are views of mental models, then the corresponding mental models have to be analogical as well. These mental models are internal perceptual representations. It follows that internal perceptual representations are analogical representations. Is this conclusion true ? It seems that we don't know yet. But then, we don't know either if mental images are "views" of mental models from a particular point of view.

(5) I suppose this is at it should be at present. A large class of mental models essentially depend on internal perceptual representations rather than on internal symbolic representations. We are tempted to regard internal perceptual representations as being (high-level and) analogical rather than symbolic, but we don't know whether this is true or not.

Conclusion

Mental model theory, in the version discussed here, is more of an intuitively plausible question and of a general framework of research than of an answer and it makes no sense, at this stage, to try to list the vast number of further issues on its research agenda.

Bibliography.

Barwise, J. and Perry, J.: *Situations and Attitudes.* Cambridge MA: MIT Press 1983.

Bernsen, N.O.: Mental models in human-computer interaction (in preparation).

Byrne, R.M.J.: Suppressing valid inferences with conditionals. *Cognition* 31: 61-83, 1989.

Collins, A.M. and Quillian, M.R.: How to make a language user. In E. Tulving and W. Donaldson (eds.): *Organization and Memory.* New York: Academic Press 1972.

Erickson, J.R.: A set analysis theory of behavior in formal syllogistic reasoning tasks. In R. Solso (ed.): *Loyola Symposium on Cognition* Vol. 2, Hillsdale NJ: Erlbaum 1974.

Fodor, J.A, Garrett, M.F., Walker, E.C.T., and Parkes, C.H.: Against definitions. *Cognition* 8, 263-367, 1980

Gentner, D. and Stevens, A.L. (eds.): *Mental Models.* Hillsdale NJ, Erlbaum1983.

Grice, P.: Logic and conversation. In P. Cole and J.L. Morgan (eds.): *Studies in Syntax* Vol. 3: *Speech Acts.* New York: Academic Press 1975.

Guyote, M.J. and Sternberg, R.J.: A transitive-chain theory of syllogistic reasoning. *Technical Report No. 5,* Dept. of Psychology, Yale University 1978.

Johnson-Laird, P.: *Mental Models. Towards a Cognitive Science of Language, Inference, and Consciousness.* Cambridge UK: Cambridge University Press1983.

Johnson-Laird, P.: Mental Models. In M. Posner (ed.): *Foundations of Cognitive Science.* Cambridge MA, MIT Press 1989.

Kamp, J.A.W.: A theory of truth and semantic representation. In J.A.G. Groenendijk, T. Janssen, and M. Stokhof (eds.): *Formal Methods in the Study of Language.* Amsterdam: Mathematical Center Tracts 1981.

Katz, J.J. and Fodor, J.A.: The structure of semantic theory. *Language* 39, 170-210, 1963.

Lakoff, G.: *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind.* Chicago: University of Chicago Press 1987.

Marcus, M.P.: A computational account of some constraints on language. In A.K. Joshi, B.L. Webber, and I.A. Sag (eds.): *Elements of Discourse Understanding.* Cambridge UK: Cambridge University Press 1981.

Marr, D.: *Vision: A Computational Investigation in the Human Representation of Visual Information.* San Francisco: Freeman 1982.

Newell, A.: Unified theories of cognition. The William James lectures. Psychology Dept., Harvard University, Cambridge MA 1987.

Pylyshyn, Z.W.: *Computation and Cognition: Toward a Foundation for Cognitive Science.* Cambridge MA: MIT Press 1984.

Rasmussen, J.: Mental models and the control of action in complex environments. In Ackermann, D. and Tauber, M.J. (eds.): *Mental Models and Human-Computer Interaction 1.* Amsterdam: North-Holland 1990.

Wason, P.C. and Johnson-Laird, P.: *Psychology of Reasoning: Structure and Content.* Cambridge MA: Harvard University Press 1972.

Wilson, J.R., and Rutherford, A.: Mental Models: Theory and Application in Human Factors. *Human Factors* 31 (6), 617-34, 1989.