A second remarkable claim is that ordinary sentence understanding involves the same kind of internal models. So if I tell you that there’s a cup on the table and you understand me, that’s in part because you have constructed an internal scale model containing entities corresponding to the cup and the table, spatially arrayed.

A difficulty for those who want to understand or to appraise mental models is that their proponents seem to have somewhat different views. The term “model” has, of course, lots of current uses—compare “model” in “model theory,” “mathematical model,” “simulation model,” or “scale model”—and the notions of mental model differ in a parallel way. Examples of these varied uses can be found in the contributions to Gentner and Stevens (1983), in Johnson-Laird (1983), and in Kahneman and Tversky (1982). To help get a grip on these differences, we can begin by making a course distinction among these theories according to whether or not mental models are supposed to be representations in formats distinct in kind (and perhaps with different representational powers) from more familiar formats (e.g., networks or propositions).1

For many mental modelers representational format is not an essential issue: These investigators have been struck by the role that world knowledge or domain-specific knowledge plays in cognitive activities like problem solving or comprehension, and they have offered some plausible hypotheses about the scope of this knowledge. For example, Hayes (1979) and many of the contributors to Gentner and Stevens (1983) are interested in explaining common-sense conceptions of physical phenomena in domains like electricity, heat, and motion. The emphasis is on the content of these beliefs, however they happen to be represented, and with the development of the beliefs with increasing expertise. However, other mental modelers take a much more literal approach, and it is really to this group that my comments apply about thought experiments and internal models. According to this group, mental models are quite unlike the usual propositions or networks and promote a kind of reasoning that is different from standard or probability systems.2

In what follows, I’ll use “literal mental model” to refer to those conceptions in which the models are unique representations in the above sense and “figurative mental model” for those in which no representational uniqueness is implied. “Figurative” in the latter phrase does not mean that the phenomena studied by these investigators are illusionary. On the contrary, these phenomena (some of which are discussed in the next section of this chapter) are of genuine psychological interest, and the case for domain-specific knowledge is persuasive. The question that I wish to raise is whether the evidence warrants more than this—whether the knowledge

Lately, a number of investigators in cognitive psychology and artificial intelligence have proposed that some puzzling facts about thinking can be explained through what they call “mental models.” People’s ability to understand connected discourse, their knowledge of science, and their skill in reasoning are all topics that mental models are supposed to elucidate. Mental models are held to be responsible for both the successes people have and the errors they make when they are faced with logical, scientific, or probabilistic questions. To further whet our curiosity, some of the proposals advanced by the mental modelers are really quite startling. An example, one claim is that when you reason about a problem, what you’re doing is manipulating in your head working models of the domain in question. If you’re reasoning about an electronic circuit, then you’re performing quite literally a thought experiment with a mental circuit. Similarly, if you’re trying to produce a conclusion for a classical syllogism—for example, the syllogism in (1)—then you’re mentally shuffling tokens corresponding to artists, beekeepers, and clerks.

(1) None of the artists are beekeepers.  
All of the beekeepers are clerks.

Therefore, ???

Mental Muddles


LANCE J. RIPS

259
needs to be packaged in the special purpose representations that the literalists posit. To answer this question, I'd like to examine some arguments that have been advanced to support literal models: that these representations are needed in order to provide a psychological semantics for natural language (Johnson-Laird 1983) and to explain probabilistic (Kahneman and Tversky 1982) and deductive reasoning (Johnson-Laird 1983; Johnson-Laird and Bara 1984). These claims are discussed in turn in the third, fourth, and fifth sections below. The point that I hope will emerge is that the advantages of mental models can all be obtained without commitment to literal models-in-the-head, and that the figurative position is therefore the more reasonable one.

Why Mental Models?

Even when we restrict our attention to literal mental models, we still find some important differences among these theories. For Johnson-Laird (1983), mental models have affinities to mathematical model theory, whereas the mental models of de Kleer and Brown (1981, 1983) and Kahneman and Tversky (1982) are more like computer simulation models. I'll return to these differences presently, but in the meantime it's worthwhile noticing a commonality that may help explain what motivates the literalists. This characteristic is that (literal) mental models are typically claimed to mirror quite directly the structure of the domain they represent. For example, a mental model of a circuit contains parts that correspond to those of the circuit: resistors where the circuit contains resistors, capacitors where the circuit contains capacitors, and so on. Because mental models are tailored in this way to the relevant domain, they are consistent with evidence from a variety of subfields in cognitive psychology that domain content often plays a large role in comprehension and reasoning.

It's unclear how much weight this notion of domain specificity will bear but it is easy to cite examples that accord with this intuition. In categorization research, for instance, investigators like Medin and Schaffer (1978) have proposed that mental representations of natural categories such as apple or trout might consist, not of abstract lexical entries, but of sets of remembered exemplars—memories of apples and trout actually encountered. A judgment about whether a new instance is a member of a category is then made by computing its similarity to these remembered instances. Similarly, a good deal of research on the comprehension of discourse suggests that one's memory of a passage includes facts that are not directly expressed by the sentences, but that are imported from one's prior knowledge of the passage's subject matter. A well-known example is that subjects presented with the sentence Two robins crouched on their nest as the hawk flew above them have great difficulty discriminating it from Two robins crouched on their nest as the hawk flew over it (Bransford, Barbary, and Franks 1972). Since there is much less confusion when "beside" is substituted for "on" in the above sentences, it seems reasonable to think that the difficulty is caused by subjects remembering descriptions of the spatial layout that the initial sentence implies, rather than by subjects remembering the sentence itself. Literalists tend to read these demonstrations as evidence that there is something wrong with propositional or network representations: because these abstract formats don't by themselves predict memory confusions, they fail to capture some significant psychological generalizations.

Further support for domain specificity comes from studies of probabilistic and deductive reasoning. Thus, Kahneman and Tversky's (1972) research suggests that estimates of the frequency and probability of an event are ordinarily calculated on the basis of intrinsic characteristics of the event rather than on the basis of abstract properties like the base rate of its occurrence or the variance of the relevant distribution, properties that are required for normatively appropriate judgments. Research on deductive reasoning has also uncovered clues that the specific subject matter or content of a set of premises can influence judgments about the soundness of a conclusion drawn from them. Since the interpretation of these results will be important later in this chapter, we need to pause to consider these findings in more detail.

I can most easily illustrate these content effects through some data that Sandra Marcus and I collected as part of an investigation of reasoning with conditional sentences (Marcus and Rips 1979; see also Rips and Marcus 1977). The experiment itself was quite simple. We gave subjects a series of arguments, each consisting of two premises and a conclusion, as shown in the top half of Table 1. The first premise was always a simple conditional and the second premise was either the antecedent or the consequent of the conditional in negated or unnegated form. If the second premise was the antecedent, the conclusion was the consequent, which again could appear negated or unnegated. If the second premise was the consequent, the conclusion was the antecedent, negated or unnegated. Subjects' task was to read the arguments and to decide whether they were logically sound. For our purposes, the main manipulation in this experiment consisted in the way that the arguments were framed. In one of our conditions, subjects saw the arguments phrased in terms of the disposition of letters and numbers on a set of cards. For instance, a modus ponens argument (Argument 1 in Table 1) would appear as in (2). The subjects were told that the sentences referred to a pack of cards that could contain any of the letters A, B, or C on the left side and any of the numbers 1, 2, or 3 on the right, but...
second condition, the same arguments were framed in terms of the actions of a pinball-type device that contained three channels along which a ball-bearing could roll and a set of three differently colored lights. In this condition, the same modus ponens inference would have appeared as in (3).

(2) If there's a B on the left side of the card, then there's a 1 on the right side.
  There's a B on the left side.
  There's a 1 on the right side.

(3) If the ball rolls left, then the red light flashes.
  The ball rolled left.
  The red light flashed.

As a subject in this experiment, the responses that you give should, of course, depend on your interpretation of the conditional. For example, if you understand the first premise as a material conditional, then you should respond that the modus ponens and modus tollens arguments are sound and that the other arguments are not (i.e., you should accept Arguments 1 and 8 of Table 1 and reject the rest). Using this kind of correspondence, we were able to classify the subjects into those who responded as if they were taking the first premise to be a material conditional, those who understood it as if it were a material biconditional, and others. In the bottom panel of Table 1, you'll see the distribution of subjects in these categories for each of the two types of content. These distributions are based on 54 undergraduate subjects, none of whom had taken a course in formal logic. The obvious fact about these data is that the distributions are different for the card and machine contexts. For the cards, Material Conditional responses outnumber Material Biconditional ones by more than four to one. For the machine, on the other hand, the proportions are exactly the same.

The point of this illustration is that the subjects were responding, not just to the conditional form of the sentence, but also to the subject matter of the problem. One might be tempted to dismiss the results of one experiment; however, content effects like these are pervasive. Results like ours had previously been obtained by Staudenmayer (1975). Marcus and I also found content effects when subjects were asked to determine the truth of conditionals, given truth values for their antecedents and consequents, and similar results have been reported by Johnson-Laird and Tagart (1969) and by Legrenzi (1970). Other investigators have obtained content effects in Wason's well-known selection task (Wason and Johnson-Laird 1972, chap.

Table 1. Argument forms (Panel A) and distribution of responses (Panel B) for the conditional reasoning task from Marcus and Rips (1979)

<table>
<thead>
<tr>
<th></th>
<th>A. 1. If A, C.</th>
<th>5. If A, C.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A.</td>
<td>C.</td>
</tr>
<tr>
<td></td>
<td>C.</td>
<td>A.</td>
</tr>
<tr>
<td>2.</td>
<td>If A, C.</td>
<td>6. If A, C.</td>
</tr>
<tr>
<td></td>
<td>A.</td>
<td>C.</td>
</tr>
<tr>
<td></td>
<td>Not C.</td>
<td>Not A.</td>
</tr>
<tr>
<td>3.</td>
<td>If A, C.</td>
<td>7. If A, C.</td>
</tr>
<tr>
<td></td>
<td>Not A.</td>
<td>Not C.</td>
</tr>
<tr>
<td></td>
<td>C.</td>
<td>A.</td>
</tr>
<tr>
<td>4.</td>
<td>If A, C.</td>
<td>8. If A, C.</td>
</tr>
<tr>
<td></td>
<td>Not A.</td>
<td>Not C.</td>
</tr>
<tr>
<td></td>
<td>Not C.</td>
<td>Not A.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Problem Content</th>
<th>Material Conditional</th>
<th>Material Biconditional</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cards</td>
<td>.296</td>
<td>.074</td>
<td>.630</td>
<td></td>
</tr>
<tr>
<td>Machine</td>
<td>.278</td>
<td>.278</td>
<td>.444</td>
<td></td>
</tr>
</tbody>
</table>

14), as well as in experiments on classical syllogisms (e.g., Wilkins 1928). Whatever its cause, the effect is clearly robust.

Literalists see these content effects as further evidence that something is wrong with propositional or network representations. According to the literal approach, reasoning isn't a matter of proof based on abstract logical formulas but of the creation and manipulation of internal analogues. In terms of the experiment that I described, a subject forms a model of a pinball machine in one condition and a model of a card in another. The subject will then agree that the sample argument about the pinball machine is sound only if every internal model (or simulation) in which the premises come out true, the conclusion is true as well. Since the pinball machine model will presumably produce different results from the card model, we should expect corresponding differences in subjects' performance. Thus, mental models handle content effects by building the content into the reasoning process itself. Content influences reasoning because reasoning is accomplished by transforming (domain-specific) content.

The foregoing discussion points to several aspects of literal mental
models that contrast them with more familiar assumptions in cognitive science. In the rest of this chapter, I'd like to focus on three of these, all of which are quite controversial and which raise some thorny issues about cognitive representation in general. The first is that mental models provide a kind of classical, referential semantics for linguistic forms. Roughly, understanding a sentence is a process that includes two steps. First, input sentences of natural language are stored in memory in the form of propositions; then, optionally, these propositions are interpreted with respect to a mental model of the referents of these sentences. This is the model-theory-in-the-head idea that I spoke of earlier and that is championed by Johnson-Laird. The second issue has to do with models as simulations or thought experiments—a point of view that's common to psychologists like Kahneman and Tversky (1982) and AI researchers like Forbus (1983) and de Kleer and Brown (1981, 1983). The emphasis here isn't on models as interpretative devices but on models as a way of reasoning or answering questions about complex systems. The third issue is closely tied to the problem of content effects that we have just looked at. The claim is that when people are faced with a deduction problem, they solve it by manipulating mental models. Thus, mental models allow us to eliminate the need to posit internal versions of rules like modus ponens or universal instantiation in favor of a procedure that consists in constructing and rearranging a model's components. What I propose to do is examine each of these issues in turn and then see what can be made of the literalists' case.

Before getting down to work, though, a preliminary comment is in order concerning the relationship between mental models and mental imagery. Like the mental modelers, advocates of imagery hold that networks and propositions do not exhaust the possible formats for cognitive representations. Moreover, both images and models are domain-specific, since the formal properties of these representations are supposed to mirror those of the objects and relations they stand for. Similarly, proponents believe that imagery is implicated in various problem-solving tasks, especially ones that have a spatial component. These points of resemblance may lead you to suppose that mental models are images and to take evidence for and against the imagery hypothesis as bearing on mental models. But although it may be possible to reconstruct models as images, there are some reasons to keep the current formulations of mental models distinct from imagery theories. For one thing, mental modelers themselves insist that models differ from both propositions and images (Johnson-Laird 1975, chap. 7). For another, images don't have the power that has been imputed to mental models. As an example, mental models (conceived as internal simulations) are supposed to be responsible for people's predictions about uncertain singular events, such as whether the marriage of a friend will end in divorce (Kahneman and Tversky 1982). It seems unlikely that conventional imagery could be behind such judgments; for although you may be able to image the end of a marriage as a kind of internal movie, the imagery in this situation doesn't seem instrumental in the prediction. The plausibility of the image is more likely to be a function of the prediction than the other way round. For reasons like these, I propose to take the arguments for mental models on their own merits rather than assimilate models to images. As for the imagery hypothesis itself, I would like to remain agnostic for purposes of this chapter. In particular, references to "perceptual representations" should be understood to mean representations derived from perception, not representations with an imaginal character.

Muddle I: Mental Models as Semantic Models

As I have already indicated, mental models are advertised as a "semantic" method that is inherently distinct from the "syntactic" method that is more usual in cognitive psychology. Whereas most cognitive psychologists believe that information is mentally represented in terms of networks or logical forms, which are syntactically described strings of internal symbols, mental models like Johnson-Laird (1983) claim that they're attempting to explicate notions like truth and reference and to incorporate these notions in a cognitive account of comprehension. The argument is that a psychological account can't do without these semantic properties, since there are purely psychological phenomena that depend on them.

At first glance, these claims of mental modelers may seem odd since they appear to violate a basic tenet of cognitive psychology: what Fodor (1980) calls the "formality constraint." This is the idea that the business of cognitive psychology is to describe computational processes that operate on mental representations and that the only computationally relevant properties of these representations are structural ones internal to the representations themselves. In particular, cognitive psychology must do without semantic notions like truth and reference that depend on the relationship between mental representations and the outside world. Most cognitive psychologists probably accept this formality constraint, though they might not put it in exactly this way. They believe, in other words, that what's cognitively relevant are representations (perceptual representations, for example, or representations derived from discourse), that people can compare these representations in various ways, but that they don't have cognitive access to the external relations these representations bear to the world. This makes it surprising that mental modelers—and especially the
cognitive psychologists among them—would appeal to semantic properties.

To determine whether literalists are correct in their demand for referential semantics within cognitive psychology, we need to take a closer look at their proposals. The basic argument (due to Johnson-Laird 1983, chap. 10) is that the usual psychological accounts of how people grasp the meaning of sentences—accounts based on network structures, semantic features, or meaning postulates—leave out facts about the referents of phrases and sentences. But an explanation of reference is needed to explicate even such basic notions as ambiguity and deductive inference. Therefore, the usual psychological explanations are no good. And in particular, what they're missing is exactly what mental models can provide, since mental models supply referents for expressions of mentalese. This conclusion clearly depends on whether Johnson-Laird is correct in asserting that the explanation of comprehension and inference requires a theory of reference, and by way of evidence he offers a number of examples that he believes make the case for incorporating reference within psychology. To see if this is right, let us examine one of these examples, the one that seems to provide the strongest support for his position. The first concerns people understand sentences with pronouns, and the second concerns inference with relational predicates.

To begin, then, consider the sentence They are handsome. This sentence seems to be ambiguous because handsome itself has two meanings: When applied to acts of various kinds, handsome means something like generous (as in Their contributions were handsome), whereas when applied to people, handsome means attractive. So by knowing the referent of they in the original sentence, you can infer the correct meaning of the adjective and hence of the sentence as a whole. Conversely, if you happen to know which meaning of handsome is intended but aren't sure of the referent of they, then you can use that fact to narrow the possibilities. If it means generous, they will have to denote acts rather than people. Now, according to Johnson-Laird, this interplay between lexical meaning and referential context can't be handled by meaning rules expressed as features or meaning postulates, since these lexical meaning rules don't have access to facts about what the word they refers to. But mental models can again come to the rescue. If you happen to know the referent of they from other linguistic context, then you can put this information into your mental model and use it in the interpretation of the sentence.

Johnson-Laird is short on the details of how mental models are supposed to effect this marriage of reference and lexical analysis. You might also legitimately wonder whether this example shows a fundamental flaw with internal meaning rules. You might feel, for example, that there is a reasonable sense in which They are handsome really is ambiguous, no matter how much a speaker or listener knows about the surrounding referential context. However, the present point is that even if you believe there is a problem, there is nothing about mental models that make them especially well suited to solve it. The relevant contextual features of such an utterance could just as easily be spelled out mentally as a logical formula or as part of a network. For instance, if They are handsome can be represented in the network shown on the right of Figure 1 (based on Anderson 1976), then it seems that we should also be able to encode the relevant contextual information (that "they" are people, say) as in the left half of this structure. The network should then be sufficient to trigger whatever inference procedures are necessary to select the correct meaning of the sentence.

As we will see, similar conclusions apply to the second example, which centers on the relation x is on y's right (or x is to the right of y). Suppose Ann, Beth, and Cathy are sitting on a bench such that Ann is on the right of Beth and Beth is on the right of Cathy. Then surely Ann is on the right of Cathy. However, Johnson-Laird points out that this conclusion does not follow if some other seating arrangement is adopted. For example, if the three were evenly spaced around a circular table, then when Ann is on Beth's right and Beth is on Cathy's right, Ann is not on Cathy's right. Plainly, too, we could alter the size of the table and the number of people seated around it so that Ann is to the right of the first k individuals but not to the right of person k + 1. What Johnson-Laird (1983, p. 241) concludes from this is that "the extent of the transitivity of the relation varies as a function of the seating arrangements up to an arbitrary number of individuals, and would accordingly require an infinite number of different meanings in order to cope with each possible extent from zero upwards."

Since an infinite number of meaning postulates for to the right of is out of
the question, meaning postulates must go. What should replace them are mental models, which can incorporate contextual information about the exact nature of the seating arrangement.

Someone who believes in meaning postulates might respond to such an argument like this: "What Johnson-Laird's example proves is not that the extent of the transitivity of [to the right of] varies—whatever that might mean—but simply that to the right of is intransitive. What he has done is provide a counterexample showing that for some $x$, $y$, and $z$, Right-of ($x,y$) and Right-of ($y,z$) but not Right-of ($x,z$), and this exactly meets the definition of an intransitive relation. The fact that it is sometimes true that Right-of ($x,z$) couldn't be more irrelevant. This being so, meaning postulates can hardly be faulted for failing to state under what circumstances, when Right-of ($x,y$) and Right-of ($y,z$) are true, Right-of ($x,z$) is true as well. Indeed, the situation is exactly the same as with any more obviously intransitive predicate—say, can win against (e.g., in tennis). It's notorious that Ann can win against Beth and Beth against Cathy, but Ann can lose to Cathy. It would be absurd to require that a lexical semantic formalism predict who would win in this third match—that's the job for a theory of tennis skill, not for a theory of meaning."

This objection looks convincing, but a further point that needs to be emphasized is that, even if the objection were mistaken, nothing about Johnson-Laird's example requires literal mental models. For instance, we could specify by means of a network or by separate propositions the position of the individuals and the directions they face (e.g., as vector coordinates). Then questions about who is to the right of whom can be solved by computations on these values. Of course, there may be empirical problems with such a propositional representation—for example, it might mispredict reaction time data when subjects are asked to decide about Right-of relations—but the round table example in itself is no argument against it.

Exactly the same deficiency recurs for all of the other examples that Johnson-Laird mentions as reasons for abandoning a psychosemantics based on meaning rules; that is, none of them call for anything more than a mechanism that can compare representations derived from discourse to representations derived from perception. Disambiguation of They are landowners requires only that comprehenders (a) have some way of representing the perceptual context in which they is specified, and (b) have a way of relating it to the representation of the sentence. Believing that Ann is (or is not) on the right of Cathy demands only that a representation of the sentence match one derived from perception or earlier discourse. The need for such a comparison process is no news, and it has been the subject of a great deal of research in cognitive psychology (e.g., Clark, Carpenter, and Just 1973). The examples provide no grounds to suppose that either perceptual or discourse representations are other than propositions or networks, and thus they are no evidence for literal mental models.

A number of philosophers have argued that a semantics for natural language is incomplete without a theory of truth to show how expressions of the language are connected to the world (Davidson 1967; Lewis 1972). Johnson-Laird's plea for a semantics based on mental models often appears to echo these ideas. For example, Johnson-Laird (1983 p. 259) complains that networks, meaning postulates, and decompositional theories "purport to capture the semantic properties and relations of expressions, but they give no account of their truth conditions," whereas mental model theory can. But it is essential to notice that mental models—even literal ones—are just representations too. Although it might be possible to construct a theory of what it means for one mental representation (e.g., a mental proposition) to be true in relation to another (e.g., a mental model), this clearly wouldn't fulfill the goal of connecting mental representations to the external world. For precisely the same reason about truth would arise with respect to the mental models. Thus, if meaning postulates and other schemes can be faulted for not dealing with referential semantics, mental models can't be criticized on the same basis. To provide a referential semantics for mental representation, one must show how to get around a number of forceful objections that appear to doom any such project from the start (e.g., Putnam 1983; Quine 1969; Stich 1983). And to provide such a semantics within cognitive psychology, one must also get around additional obstacles concerning the limits of computational theories (Pollack 1978, 1980). As far as I can see, mental models are of no help in resolving any of these difficulties.

In short, anyone interested in selling mental models is probably well advised not to base his pitch on similarity to referential semantics. Fans of reference-and-truth are unlikely to find internal models to be fair substitutes. Furthermore, whatever mental models appear to be able to do in the line of explaining psychological semantics can already be done by the old notational systems. If mental models do have an advantage, it is more likely to be in explaining reasoning, so it's to this topic that we should turn.

Muddle II: Mental Models as Simulations

A better case for mental models can be built around the way we reason about certain sorts of physical systems. There is an intuition that in thinking about systems we know well, we are able to trace mentally the causal sequence of events that take place during the systems' operation. This kind
of thinking can prove useful in correcting faults in a device or in anticipating how the device will perform in some new situation. I'm going to assume that this sort of reasoning is familiar—perhaps from program debugging or from other sorts of troubleshooting (e.g., figuring out why your doorbell isn't working). The important thing is that this sort of reasoning doesn't feel like carrying out a derivation in some sort of internal logic or probability calculus. In AI this kind of reasoning has been called "envisioning" or "running" a mental model, and it is the subject of current research by Forbus (1983) and by de Kleer and Brown (1981, 1983). For example, de Kleer and Brown have constructed programs that can answer questions about circuits by simulating their operation in a qualitative model. The question I want to raise concerns how easily this simulation idea can be extended as an account of actual human performance. How much of human reasoning can be reduced to performing a simulation in a mental model of the target domain? This question is highlighted by Kahneman and Tversky's (1982) claims that many tasks calling for prediction and evaluation rest on mental simulations. They mention in particular predicting how well two friends of yours who haven't seen each other will get along when you introduce them, assessing the likelihood that the United States will invade Saudi Arabia in the next decade, and evaluating counterfactual statements such as "He could have coped with the job situation if his child hadn't become ill." All of these tasks are supposed to be solved by running a mental model of the situation and seeing how easily such outcomes can be produced.

The trouble is that the AI research on mental models itself casts doubt on the generality of mental simulation. In the complex social-political contexts that Kahneman and Tversky discuss, mental simulation may just be too hard. In order to make mental simulation even remotely plausible, both Forbus and de Kleer and Brown limit the envisioning process to qualitative facts about the device or system in question. For example, the circuit simulations don't have access to exact values for circuit parameters but only to information about whether these parameters are going up or down. This moves the envisioning process away from simulations of the sort that an expert might perform on a computer (simulations involving the solution of large systems of differential equations) and toward psychologically plausible simulations. But lack of precise parametric values means that the performance of simulation is underdetermined. Although you don't have to solve lengthy systems of equations in your head, qualitative simulation requires you to consider multiple possible realizations of the system. The difficulty of so doing is amply demonstrated by the same AI research.

As an example, consider Forbus's program FROB. This program is able to describe and answer questions about the behavior of balls (actually point masses) bouncing around in a two-dimensional plane bounded by arbitrary line segments. In the part of the program that is relevant to us, the plane is divided into nonoverlapping regions that Forbus calls "places." The trick is to predict the places a ball can reach, given a qualitative description of its motion. For example, can the ball reach Place Number 1 if it is currently at Place Number 3 headed northeast? The obvious difficulty is that this qualitative information doesn't constrain the ball's trajectory sufficiently to allow an unambiguous prediction of its motion. The program therefore has to keep track of all possible paths through the space, which is infeasible in any but the very limited terrain. Of course, you can reduce the number of paths by re-introducing quantitative information about energy or elasticity—FROB actually does this in its nonenvisioning mode—but the process then reverts from mental simulation to solving equations in textbook mechanics.

Forbus's conclusion from this is that "although envisioning is an important technique, . . . the burden of building a complete description of possible states is too onerous outside very small domains, and is too restrictive a style to capture all of the ways people use qualitative physical knowledge." (Forbus 1983, p. 70). Similarly, de Kleer and Brown mention that "although one would intuitively expect qualitative simulations to be simpler than quantitative simulations of a given device, they turn out to be equally complex, but in a different way" (de Kleer and Brown 1983, p. 155). The moral is that unless you know a great deal about a system, even qualitative simulations are still limits, given the normal memory and processing limits that humans have to contend with. For any system that is too complex, we're thrown back on crude rules of thumb that are far from literal simulation. For example, I suspect that although most people know a little about what the parts in their car do, if anything goes wrong they have to rely on simple-minded heuristics such as "if nothing happens when I turn the key and if it's a cold day, then maybe it's the battery."

Some initial empirical evidence against mental simulation comes from an experiment conducted with DeeDee Gentner. In this study, college-age subjects were told that the experiment concerned a closed room containing a pan of water and that they would be asked questions about the relation between physical variables that describe this room. These variables were temperature of the air, temperature of the water, air pressure, evaporation rate, and relative humidity. On each trial in the experiment, subjects were presented a pair of these variables and were asked to decide whether a change in the first variable would cause a change in the second. For example, if air pressure goes up, would that cause a change in the room's relative humidity? Since each subject answered a question for each pair, it's possi-
The thing to reflect on is that if mental simulation breaks down even in this austere system, it seems hopeless when applied to systems that are as complex as the social and political ones that Kahneman and Tversky discuss. The amount of knowledge required to make the models runnable is just too great, if the AI work that I've mentioned is any guide. Defenders of mental simulations are likely to object that in these social-political contexts, people revert to very simplified models of the domain and run their simulations on these stripped-down models. However, at some point it becomes difficult to tell whether processing of this kind is literal simulation or is more like the rule-of-thumb reasoning described above. For example, take Kahneman and Tversky's (1982, p. 203) own evidence on this issue. In one of their experiments, subjects were given the following problem.

Mr. Crane and Mr. Tees were scheduled to leave the airport on different flights, at the same time. They traveled from town in the same limousine, were caught in a traffic jam, and arrived at the airport 30 minutes after the scheduled departure time of their flights.

Mr. Crane is told his flight left on time.

Mr. Tees is told that his flight was delayed, and just left 5 minutes ago.

Who is the more upset?
Mr. Crane Mr. Tees

Not unexpectedly, 96% of the subjects stated that Mr. Tees was the more upset; and according to Kahneman and Tversky, this is because the subjects tried to imagine how close each traveler came to making his flight. But assuming this is true, are subjects literally simulating a further delay in the flight (e.g., a further problem with the aircraft) or an event that would permit the limousine to get to the airport sooner (e.g., a new route that the limousine could take)? Or rather are they reasoning with simpler plausible rules—for example, reasoning that any flight that is delayed by 25 minutes could easily be delayed by 30 or more, thus permitting Mr. Tees to get on board.

Mental muddles' loose talk about mental simulations or simulation heuristics is of little use unless it can be translated into a plausible psychological mechanism that is capable of doing the simulating, where by "plausible mechanism" I mean one that is framed in terms of the ordinary cognitive vocabulary of elementary information processes—e.g., comparison and storage operations. But the best examples we have of how this translation would go—the AI research just described—puts these simulations out of the reach of all but experts in a given domain. Of course,
people do make predictions about what will happen in complex physical and social interactions. They plan actions and evaluate the probable consequences. If you like, you can refer to these projections as "simulations based on mental models." But in doing so, you forfeit the claim that "simulation" should be taken literally as a distinct type of reasoning. You have become a figurative, rather than a literal, mental modeler.

Muddle III: Models vs. Inference Rules

I mentioned at the beginning that a main attraction of mental models for psychologists is that they seem to provide a way of dealing with the effects of context in deductive reasoning tasks. The hope is that this problem can be explained if we conceive of deduction as a matter of manipulating models, which wear their content on their sleeves. Furthermore, once we have switched to mental models as a deductive method, we can eliminate the need to posit mental inference rules that carry out mental derivations. In Chapter 2 of his book, Johnson-Laird explicitly contrasts the mental-models idea with what he calls "the doctrine of mental logic," the notion that people possess internal deduction schemas. These rules might be mental analogues of the ones you find in Gentzen- or Jaskowski-style natural deduction systems. They might include rules like modus ponens, conjunction introduction, universal instantiation, and so on, that are familiar from introductory texts in predicate logic. In order to decide if an argument is sound, you apply these rules to the premises in an attempt to find a mental derivation of the conclusion. Psychological theories of this type have been proposed by Braine (1978), Osherson (1974–1976), and me (Rips 1983), but also by Johnson-Laird himself in earlier work (Johnson-Laird 1975).

What Johnson-Laird now believes is that mental logic was just a mistake—that, in short, deductive reasoning is not a matter of mental proof but of mental models.

One way in which mental models could be responsible for subjects' performance in deduction tasks is if subjects treat the tasks as thought experiments about the domain in which the problem is framed—for example, packs of cards or pinball machines in Examples (2) and (3). But the difficulty with this line is the one that we've just encountered in discussing the AI results: Literal mental simulation is implausible for most domains. It's highly unlikely, for example, that naive subjects know enough about a pinball machine's inards to be able to construct a workable mental simulation. To make matters worse, the deduction-as-simulation idea makes it impossible to explain why certain inferences seem so easy, no matter what domain they're applied to. For example, subjects' recognition of the validity of an argument of the modus ponens form does not depend on the complexity of its content (Rips and Marcus 1977). But we ought to predict such dependence if this inference had to be checked by running a mental simulation. You might counter by saying that a statistically null effect is weak grounds for rejecting the mental-simulation idea. Perhaps the simulation need only be run once per domain and the results of the simulation stored for later use. But even this possibility is ruled out by the fact that such inferences are readily carried out even for domains for which you've had no previous exposure and have no model to employ (Evans 1977; Tappin and Stadtmeyer 1973). Deduction-as-simulation explains content effects, but unfortunately it does so at the cost of being unable to explain the generality of inference.

It appears that the only way to achieve this generality is to drop mental simulation and adopt some more abstract procedure, and this is exactly what Johnson-Laird has done in the explanations he offers for deductive reasoning. His account relies not on the notion of mental models as simulations, but on the idea of mental models as a kind of semantics, the same idea we looked at earlier. To see what mental models of this sort are supposed to look like and how they eliminate inference rules, let's consider a specific example. It's unfortunately true that psychologists have devoted a great deal of effort to studying classical syllogisms, even though syllogisms comprise only a tiny subset of deductive inferences. Johnson-Laird is no exception to this trend and in fact his theory of syllogisms is the central example of mental models in his book. Let us suppose that you are the subject in an experiment in which you are given on each trial a pair of syllogistic premises and are asked to produce a conclusion that follows from these premises. On one of the trials, you receive the pair in (4), the same problem we saw in Example (1):

(4) None of the artists are beekeepers.
All of the beekeepers are clerks.

Therefore, ??

According to the theory, the first thing to do in this situation is to interpret the individual premises, and you do this by configuring a set of mental exemplars that correspond to particular artists, beekeepers, and clerks. The theory stipulates that the model for the first premise is the one in (5), and the model of the second premise is that of (6).

(5) artist
artist

(6) beekeeper

(6) $\text{beekeeper}_1 = \text{clerk}_1$
$\text{beekeeper}_2 = \text{clerk}_2$
$0 \text{clerk}_3$

In these diagrams, the exemplars are indicated by words, though the words themselves have no significance other than to differentiate exemplars of different types. To indicate identity among the exemplars, you place an equal sign between them; so the equal sign between beekeepers and clerks in (6) represents the fact that those beekeepers also happen to be clerks. The zero in front of clerk 3 indicates that this individual may or may not exist; that is, there may or may not be a clerk who is not a beekeeper. In (5) the line in the middle of the diagram indicates that these artists and beekeepers are distinct individuals. The models in (5) and (6) are those of Johnson-Laird and Bara (1984), who present a more detailed and somewhat revised version of the theory described by Johnson-Laird (1983) and Johnson-Laird and Steedman (1978).

After you've formed these representations of the premises, you have to combine them into a unified model of the problem as a whole. One way to do this is according to the diagram in (7), and it's easy to see that this is a perfectly good model in the sense that it makes both of the premises true.

(7) $\text{artist}_1$
$\text{artist}_2$
$\text{beekeeper}_1 = \text{clerk}_1$
$\text{beekeeper}_2 = \text{clerk}_2$
$0 \text{clerk}_3$

In order to generate a conclusion to the syllogism—which is your task as a subject in this experiment—you try to read from this model the relationships that hold between the outer terms, artists and clerks. The conclusions that suggest themselves on the basis of this diagram are None of the artists are clerks, None of the clerks are artists, Some of the artists are not clerks, and Some of the clerks are not artists. However, in order to be sure that these candidate conclusions are valid, you need to check that there are no alternative models of the premises that make these conclusions false. You therefore need to look for new ways of combining the premises, and one such possibility is given in (8):

(8) $\text{artist}_1$
$\text{artist}_2$
$0 \text{clerk}_3$
$\text{beekeeper}_1 = \text{clerk}_1$
$\text{beekeeper}_2 = \text{clerk}_2$

We've moved the position of the optional clerk 1, and the juxtaposition of clerk 4 and artist 1 is supposed to represent an "indeterminate" relation between them; Artist 2 may be the same person as clerk 3. According to this new model, it is no longer necessarily true that none of the artists are clerks nor that none of the clerks are artists. However, the conclusions Some of the artists are not clerks and Some of the clerks are not artists remain true in (8).

Finally, by adding a new optional clerk 1 to the first line of the diagram, we can create the model in (9), in which only the conclusion Some of the clerks are not artists holds.

(9) $\text{artist}_1$
$\text{artist}_2$
$0 \text{clerk}_3$
$\text{beekeeper}_1 = \text{clerk}_1$
$\text{beekeeper}_2 = \text{clerk}_2$

Books (1984) has noted in a critique of Johnson-Laird and Bara (1984) that even this latter conclusion is invalid, since the premises of Sylllogism (4) could be true and the conclusion false under circumstances where there are no beekeepers. However, Johnson-Laird and Bara assume that subjects treat all three classes (artists, beekeepers, and clerks) as nonempty, and under this assumption Some of the clerks are not artists is valid, since there are no further models of the premises that make this sentence false.

Johnson-Laird claims that the theory can successfully predict the difficulty people have in producing correct conclusions. The key notion is that the larger the number of models that you must consider, the greater the demand on memory, since memory is required to keep track of them. Because working memory is limited in capacity, the greater the number of models, the more likely you are to overlook one or more of them and produce an inappropriate conclusion. In the above example, you supposedly need to consider at least three models—the ones in (7)–(9). Other syllogisms require from one to three models and should vary accordingly in their solution difficulty. For reasons that I'll mention in a moment, my own view of the evidence is that Johnson-Laird simply hasn't given us enough information to be able to assess the prediction. However, we can grant that some proposal of this type is workable in principle and that it may be able to predict syllogism difficulty successfully.
The important question is whether mental models of this sort are more plausible components of a theory of deduction than mental inference rules such as modus ponens, and it seems to me that there are at least four reasons for being suspicious about this possibility. The first is that mental modelers get rid of traditional inference rules only at the expense of positing some rules of their own. Procedures are necessary to specify how the models should be used to represent the premises, how these premise models should be combined, how one combined model should be transformed into another, and how potential conclusions should be evaluated for consistency with a combined model.

To see that this isn’t a trivial task, let’s return to Syllogism (4). An important feature of Johnson-Laird’s theory is that the probability of a correct response depends crucially on which model of the premises is considered first. For example, subjects who initially consider Model (7) should have a relatively difficult time with this syllogism; for as we’ve seen, there are four potential conclusions that hold in this model, only one of which is a valid conclusion to the argument. Two additional models [i.e., Models (8) and (9)] must be considered to eliminate the invalid conclusions and ensure a correct answer. On the other hand, subjects who start with Model (8) should be somewhat more successful. There are only two potential conclusions that are true of this model, and only a single additional model—Model (9)—is needed to determine the correct one. Thus, predictions about the difficulty of this syllogism depend on whether Model (7) or Model (8) is taken up first. In fact, Johnson-Laird and Bara (1984, p. 34) mention only one principle that is relevant to determining the order in which models are constructed: “In forming an initial model, reasoners are guided by the heuristic of trying to maximize the greatest number of roles on the fewest number of individuals.” This means that, all things being equal, subjects should opt for models in which tokens are identified—for example, where a given artist also has the role of a beekeeper or a clerk (see also Johnson-Laird and Steedman 1978). However, it is not clear whether this principle can decide between Models (7) and (8) in our example. To the extent that it applies at all, it would seem to favor (8), since in this model clerk, might also be an artist. However, if this is the case, Johnson-Laird and Bara’s own prediction about the difficulty of this syllogism comes out wrong: Instead of having Difficulty Level 3, it should have Difficulty Level 2. In other words, it’s unclear what the basis is for ordering the search of possible models of premises, and this ambiguity about the procedures makes the empirical content of the theory uncertain.

A second problem is that, although the elements in a model are supposed to be objects that provide the referents for the premises and conclusions, the rules that make use of them have to respect properties that seem much like those of standard logical syntax. The equal sign in the syllogism example has to be interpreted in a way that ensures the idea that the two tokens refer to the same individual, and the line between the artists and beekeepers must be interpreted to indicate that the separated tokens cannot refer to the same individual. Because of these features, manipulation of mental models isn’t fundamentally different from manipulation of mental propositions. To the extent that the rules that operate on the models are sensitive to these logical constants, they just are inference rules. If I’m right about this, the contrast that Johnson-Laird sets up between mental logic and mental models collapses. Of course, the inference rules that mental models employ are nonstandard, and such rules may certainly be worth study. But there appears to be no principled difference between mental logic and mental models of the sort that distinguishes formal proof theory from formal model theory.

Third, although this sort of mental model is more general than the simulation approach, it sacrifices the ability to account naturally for content effects. That is, if the terms in a syllogism are interpreted simply as tokens, then it’s no more obvious why content should influence reasoning with mental models than reasoning with standard inference rules. Representations like (7)–(9) are quite abstract and drift away from the idea of domain specificity that provides the underlying motivation for mental models. So, contrary to Johnson-Laird’s claims, content effects can’t be used as an argument for the superiority of (these sorts of) mental models over inference schemata.

Fourth, what’s so bad about inference rules, anyway? In point of empirical evidence, mental logic greatly outscore mental models. Although this isn’t the place for a thorough review (see Rips 1984), theories of mental logic successfully predict the validity judgments that subjects give for a fairly wide range of propositional-logic-type arguments (Braine, Reiser, and Ruman 1984; Osberson 1974, 1975; Rips 1983); they account for reaction times (Braine, and Reiser, and Ruman 1984), as well as inter-subject differences on problems of this type (Rips and Conrad 1983); and they help explain what subjects say when they’re made to think aloud while they’re making validity decisions (Rips 1983). They even predict which lines subjects will recall when they have to remember a natural-language proof (Marcus 1982). Braine and Ruman (1983) propose a way in which mental logic could handle classical syllogisms, and Osberson (1976) has described a mental logic for monadic predicate calculus (which includes classical syllogisms as a special case), showing that it accounts for subjects’ judgments of difficulty in evaluating a broad sample of arguments. By comparison, the only tasks for which we have anything ap-
proaching a fleshed out mental-model theory are classical syllogisms and transitive inferences.

Although it has been claimed that content effects pose a problem for mental logic, this approach has at least two options for dealing with content. One obvious way is to follow the lead of logicians and absorb content into the logic by adding operators that formalize important relations such as causality and temporal precedence. Both Osherson and I have proposed mental logics of this sort (Osherson 1976; Rips 1983). The other way of dealing with content effects is to regard them as a result of additional premises that subjects bring to the task. Although subjects are told in reasoning experiments to decide whether the conclusion follows on the basis of the premises alone, they may well find this instruction difficult to follow and treat the stimulus argument as an enthymeme. For example, subjects may implicitly use rules of thumb about the likely workings of the pinball machine to eliminate some of the contingencies in our earlier example.

A more serious objection is that the inference rules posited in a mental logic (and the control procedures that apply these rules) are ad hoc—arbitrarily selected to fit the data—and hence are uninformative about the actual procedures that subjects use in reasoning. This is true insofar as the initial selection of rules is based on their ability to account for some body of preliminary data or intuitions. But the selection is also constrained by the ability of these rules to fit together into a system that can account for the infinite number of arguments that people are capable of recognizing as sound. There has also been an attempt to constrain the systems even more tightly by enforcing symmetry relations between the introduction and elimination rules for each connective or quantifier (Osherson 1977). Furthermore, as in any other psychological theory, empirical constraints are added as the system is applied to a larger range of phenomena—in this case, a larger set of logic tasks—and as I have mentioned, mental logic has already been used to predict the outcomes for a fairly reasonable sample of problems. Mental logic has not evolved into a completed scientific theory, but neither is it any more ad hoc than its competitors. In brief, although mental models could conceivably be used to banish mental proofs, it is not clear why you would want to.

In the Mental Model Mood

Researchers who subscribe to mental models have offered a number of insights into human cognitive processes. One of their basic themes is that many psychological activities are less isolated than one may have guessed: Language understanding and reasoning (at least at higher levels) proceed against a background of common-sense knowledge. This knowledge is sometimes analogous to theories in substantive domains like mechanics or economics. The positive contributions of mental models lie in working out the details of these ideas, and there has been some progress along these lines, as several of the papers in Gentner and Stevens (1983) attest. This chapter is certainly not meant to belittle these achievements but to point out some problems that can arise from taking mental models too literally. Perhaps one could summarize these problems by saying that literal versions of the mental model approach promise much more than they can deliver. They promise to give psychology a referential semantics, explain probabilistic reasoning, and eliminate psychological rules for deductive and lexical analysis. A theory that could do all this would indeed be appealing, but in the case of mental models the promises are spurious. Looking closely, we find: (a) There is no account at all of reference or truth, if we take these notions to be relations between expressions and external objects. The semantic work that models can do can be accomplished through propositional or network mechanisms. (b) For most realistic situations, probabilistic reasoning is more plausibly explained through simple rule-of-thumb heuristics rather than through simulation in mental models. (c) Mental model approaches to deduction are restricted to a small subset of inferences, are not clearly formulated, and fail in accounting for the content effects that motivate them. Moreover, "mental model" is not a unitary concept but requires ad hoc assumptions for each domain to which it has been applied. Why would anyone suppose that a representation that seems so clearly limited could carry out such a grand design?

You can put yourself into a mental model mood by concentrating on the varied roles that perceptual representations play in cognition. It's surely possible to form perceptual representations from external information and compare these representations to others that we've constructed from discourse. Perceptual representations are helpful in other ways too. For instance, if Chase and Simon (1973) are right, master chess players can store thousands of representations of board positions to guide them to a winning strategy. In certain reasoning tasks, we might also make use of perceptual representations to ease computational difficulties—keeping track of or transforming the premises. Although you might dispute the role of perceptual representations in particular kinds of problems, you would probably admit that they're sometimes useful.

Clearly, perceptual representations are important aspects of cognition, both when we have to compare sentences with the products of visual input and when we reason. And by the way, since it's a bit trumpery to have to refer to them as "perceptual representations," let's call them "mental models" instead.
This seems to be the intuition behind what literalist mental modelers believe, and up to this point it's unobjectionable. The controversy starts when you begin to take mental models as representations that are distinct in kind—formally unlike other representations. If you then endow them with imperialist ambitions over all of reasoning and psychossemantics, you get the muddles we've discussed.

Notes

This paper benefited from discussions with cognitive psychologists and philosophers at the University of Chicago, Bolt Beranek and Newman, MIT, and the University of Arizona. I'm especially grateful to Norman Brown, Rob Chametzky, Fred Conrad, Dan Garber, Dedre Gentner, Janis Handte, Mike Harsh, Reid Hastie, Gary Kahn, Philip Johnson-Laird, Linda Levine, Ron McClanahan, Greg Murphy, and Ed Smith for comments about earlier versions of this manuscript. Part of this research was supported by US Public Health Service Grant K02 MH100236 and National Science Foundation Grant BNS 80-14131.

1. I use "proposition" (as do most psychologists) to mean what philosophers prefer to call "an internal sentence token" or, perhaps more accurately, "an internal sentence token in logical form." Since it is annoying to spell this out each time, I'll stick to "proposition" throughout this chapter.

2. By "standard deduction systems" I mean the types of axiomatic or natural proof methods described in logic texts (e.g., Mendelson 1964). Of course, by calling them "standard" I don't mean to imply that they are the only correct ways to arrive at deductive conclusions.

3. I owe this point to Edward Smith. See his review (Smith 1985) for some further examples.

4. For the distinction between "true in a model" and "true," see Wallace (1974). The closest Johnson-Laird comes to confronting these problems is in Chapter 15 of his book, and it boils down to the following "truth definition": "If a discourse has complete truth conditions, it is true with respect to the world if and only if it has at least one mental model that can be mapped into the real world. If a discourse has only partial truth conditions . . . , it is false with respect to the world if it has no mental model that can be mapped into the real world . . .." (Johnson-Laird 1983, p. 442). Since there is no explicit account of what this "mapping" amounts to, this definition is not very helpful.

5. Although I'm supposing that these rules of thumb can be expressed in logical form (either as propositions or as production rules), I'm certainly not assuming that only deductive operations are involved in answering the test questions. Expressing information in logical form does not carry with it a commitment to purely deductive procedures (Israel 1984); and in the case of the questions about evaporation, it's natural to expect that inductive inference will come into play. Although I will later defend the plausibility of mental versions of deduction rules (e.g., modus ponens) as explanations of people's reasoning about deductive problems, I don't believe that the rule inferences are deductive ones.

6. Johnson-Laird (1983, pp. 71–72) contends that syllogisms play a frequent part in everyday argumentation; yet it is significant that the examples he gives of everyday syllogisms are not of the categorical form that he proceeds to investigate. The real difficulty with the psychologists' emphasis on syllogisms, however, is not so much that they are "ecologically invalid," but that the conclusions reached in studying them do not generalize easily to other forms of deductive inference (Rips 1984). There are exactly 512 categorical syllogisms but an infinite number of deductive arguments.

7. It is also worth noticing that for the subjects to be confident that they have found the correct response to a given syllogism, they must either construct enough mental models to eliminate all possible conclusions (in which case a "no valid conclusion" response is warranted) or construct all possible models of the premises to ensure that a putative conclusion is true in all of them. However, there appear to be no theoretical limits to the number of mental models that can be constructed for a premise pair, since a new model can always be found by adding another optional token, as in the construction of Model (9). Only performance factors such as external deadlines or memory size can stop this search. This implies that subjects can never be certain of the validity of a conclusion, for some not-yet-constructed model might refute it. It also means that (contrary to Johnson-Laird 1983) mental model theory is not an effective procedure for syllogisms and is left open to some of the criticisms that he levels against earlier syllogism theories.

8. Of course, these objections are directed specifically at the mental models of Johnson-Laird, and it is always possible that another such formulation will succeed where this one fails. However, I know of no current theory that is both obviously nonpropositional and that accounts naturally for content effects. For example, Erickson's (1974) theory that syllogisms are solved by the mental manipulation of Euler circles has the same problem with content effects. All Euler circles are the same, whether they represent beekeepers or bedwarmers. (Erickson, incidentally, doesn't claim to solve the content problem).

Bibliography


283 Mental Muddles
Introduction

This paper has a somewhat different focus than most of the papers for this conference. It isn't concerned with mental representation per se, but with the relation between mental representation and epistemology. I take my topic to fall under the theme of the conference, since the conference theme is the representation of knowledge and belief and these are of central interest to epistemology. I am not concerned, however, with the standard epistemological task of giving an analysis of "S knows that p." Rather, the intent is to engage in epistemological reflection on selected aspects of mental representation. I shall begin by laying out some themes from the psychological literature on certain traits of human representation. Then I shall turn to some epistemological morals and implications of these themes.

To explain the rationale for this approach, a few comments are needed about my view of epistemology, which forms the backdrop of the discussion. (For further details, see Goldman 1986). First, I conceive of epistemology as an evaluative subject, not a purely descriptive one. Broadly speaking, it evaluates practices, processes, and methods in the conduct of intellectual affairs. Affairs of the intellect transpire in different "locations," especially the sphere of the mental and the sphere of public speech. The


