Cognition, 50 (1994) 189–209
0010-0277/94 Elsevier Science B.V.

Mental models and probabilistic thinking

Philip N. Johnson-Laird*
Department of Psychology, Princeton University, Green Hall, Princeton, NJ 08544, USA

Abstract

This paper outlines the theory of reasoning based on mental models, and then shows how this theory might be extended to deal with probabilistic thinking. The same explanatory framework accommodates deduction and induction: there are both deductive and inductive inferences that yield probabilistic conclusions. The framework yields a theoretical conception of strength of inference, that is, a theory of what the strength of an inference is objectively: it equals the proportion of possible states of affairs consistent with the premises in which the conclusion is true, that is, the probability that the conclusion is true given that the premises are true. Since there are infinitely many possible states of affairs consistent with any set of premises, the paper then characterizes how individuals estimate the strength of an argument. They construct mental models, which each correspond to an infinite set of possibilities (or, in some cases, a finite set of infinite sets of possibilities). The construction of models is guided by knowledge and beliefs, including lay conceptions of such matters as the "law of large numbers". The paper illustrates how this theory can account for phenomena of probabilistic reasoning.

1. Introduction

Everyone from Aristotle to aboriginals engages in probabilistic thinking, whether or not they know anything of the probability calculus. Someone tells you:

*Fax (609) 258 1113, e-mail phil@clarity.princeton.edu

The author is grateful to the James S. McDonnell Foundation for support. He thanks Jacques Mehler for soliciting this paper (and for all his work on 50 volumes of *Cognition*!). He also thanks Ruth Byrne for her help in developing the model theory of deduction, Eldar Shafir for many friendly discussions and arguments about the fundamental nature of probabilistic thinking, and for his critique of the present paper. Malcolm Bauer, Jonathan Evans and Alan Garnham also kindly criticized the paper. All these individuals have tried to correct the erroneous thoughts it embodies. Thanks also to many friends – too numerous to mention – for their work on mental models.

SSDI 0010-0277(93)00594-W

There was a severe frost last night.

and you are likely to infer:

The vines will probably not have survived it.

basing the inference on your knowledge of the effects of frost. These inferences are typical and ubiquitous. They are part of a universal human competence, which does not necessarily depend on any overt mastery of numbers or quantitative measures. Aristotle's notion of probability, for instance, amounts to the following two ideas: a probability is a thing that happens for the most part, and conclusions that state what is probable must be drawn from premises that do the same (see Rhetoric, I, 1357a). Such ideas are crude in comparison to Pascal's conception of probability, but they correspond to the level of competence a psychological theory should initially aspire to explain.

Of course many people do encounter the probability calculus at school. Few master it, as a simple test with adults shows:

There are two events, which each have a probability of a half. What is the probability that both occur?

Many people respond: a quarter. The appropriate "therapy" for such errors is to invite the individual first to imagine that A is a coin landing heads and B is the same coin landing tails, that is, $p(A \& B) = 0$, and then to imagine that A is a coin landing heads and B is a coin landing with the date uppermost, where date and head are on the same side, that is, $p(A \& B) = 0.5$. At this point, most people begin to grasp that there is no definite answer to the question above – joint probabilities are a function of the dependence of one event on the other.

Cognitive psychologists have discovered many phenomena of probabilistic thinking, principally that individuals do not follow the propositional calculus in assessing probabilities, and that they appear to rely on a variety of heuristics in making judgements about probabilities. A classic demonstration is Tversky and Kahneman's (1983) phenomenon of the "conjunction fallacy", that is, a violation of the elementary principle that $p(A \& B) \leq p(B)$. For example, subjects judge that a woman who is described as 31 years old, liberal and outspoken, is more likely to be a feminist bankteller than a bankteller. Indeed, we are all likely to go wrong in thinking about probabilities: the calculus is a branch of mathematics that few people completely master.

Theorists relate probability to induction, and they talk of both inductive inference and inductive argument. The two expressions bring out the point that the informal arguments of everyday life, which occur in conversation, newspaper

editorials and scientific papers, are often based on inductive inferences. The strength of such arguments depends on the relation between the premises and the conclusion. But the nature of this relation is deeply puzzling – so puzzling that many theorists have abandoned logic altogether in favor of other idiosyncratic methods of assessing informal arguments (see, for example, Toulmin, 1958; the movement for "informal logic and critical thinking", e.g. Fisher, 1988; and "neural net" models, e.g. Thagard, 1989). Cognitive psychologists do not know how people make probabilistic inferences: they have yet to develop a computable account of the mental processes underlying such reasoning.

For this celebratory volume of *Cognition*, the editor solicited papers summarizing their author's contributions to the field. The present paper, however, looks forward as much as it looks back. Its aim is to show how probabilistic thinking could be based on mental models – an approach that is unlikely to surprise assiduous readers of the journal (see, for example, Byrne, 1989; Johnson-Laird & Bara, 1984; Oakhill, Johnson-Laird, & Garnham, 1989). In pursuing the editor's instructions, part 2 of the paper reviews the theory of mental models in a self-contained way. Part 3 outlines a theoretical conception of strength of inference, that is, a theory of *what* objectively the strength of an inference or argument depends on. This abstract account provides the agenda for what the mind attempts to compute in thinking probabilistically (a theory at the "computational" level; Marr, 1982). However, as we shall see, it is impossible for a finite device, such as the human brain, to carry out a direct assessment of the strength of an inference except in certain limiting cases. Part 4 accordingly describes a theory of how the mind attempts to *estimate* the strength of inferences (a theory at the "algorithmic" level). Part 5 shows how this algorithmic theory accounts for phenomena of probabilistic thinking and how it relates to the heuristic approach. Part 6 contrasts the model approach with theories based on rules of inference, and shows how one conception of rules can be reconciled with mental models.

2. Reasoning and mental models

Mental models were originally proposed as a programmatic basis for thinking (Craik, 1943). More recently, the theory was developed to account for verbal comprehension: understanding of discourse leads to a *model* of the situation under discussion, that is, a representation akin to the result of perceiving or imagining the situation. Such models are derived from syntactically structured expressions in a mental language, which are constructed as sentences are parsed (see Garnham, 1987; Johnson-Laird, 1983). Among the key properties of models is that their structure corresponds to the structure of what they represent (like a visual image), and thus that individual entities are represented just once in a model. The theory of mental models has also been developed to explain deductive

reasoning (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991). Here, the underlying idea is that reasoning depends on constructing a model (or set of models) based on the premises and general knowledge, formulating a conclusion that is true in the model(s) and that makes explicit something only implicit in the premises, and then checking the validity of the conclusion by searching for alternative models of the premises in which it is false. If there are no such counterexamples, then the conclusion is deductively valid, that is, it must be true given that the premises are true. Thus, the first stage of deduction corresponds to the normal process of verbal comprehension, the second stage corresponds to the normal process of formulating a useful and parsimonious description, and only the third stage is peculiar to reasoning. To characterize any particular domain of deduction, for example reasoning based on temporal relations such as "before", "after" and "while", or sentential connectives such as "not", "if", "and" and "or", it is necessary to account for how the meanings of the relevant terms give rise to models. The general reasoning principles, as outlined above, then automatically apply to the domain. In fact, the appropriate semantics has been outlined for temporal relations, spatial relations, sentential connectives and quantifiers (such as "all", "none" and "some"), and all of these domains can be handled according to five representational principles:

(1) Each entity is represented by an individual token in a model, its properties are represented by properties of the token, and the relations between entities are represented by the relations between tokens. Thus, a model of the assertion "The circle is on the right of the triangle" has the following spatial structure:

△ ○

which may be experienced as a visual image, though what matters is not so much the subjective experience as the structure of the model. To the extent that individuals grasp the truth conditions of propositions containing abstract concepts, such as friendship, ownership and justice, they must be able to envisage situations that satisfy them, that is, to form mental models of these situations (see Johnson-Laird, 1983, Ch. 15).

(2) Alternative possibilities can be represented by alternative models. Thus, the assertion "Either there is a triangle or there is a circle, but not both" requires two alternative models, which each correspond to separate possibilities:

△

○

(3) The negation of atomic propositions can be represented by a propositional annotation. Thus, the assertion "There is not a triangle" is represented by the following sort of model:

¬△

where "¬" is an annotation standing for negation (for a defence of such annotations, see Polk & Newell, 1988; and Johnson-Laird & Byrne, 1991, pp. 130–1). Of course, the nature of the mental symbol corresponding to negation is unknown. The principal purpose of the annotation is to ensure that models are not formed containing both an element and its negation. Thus, the only way to combine the disjunctive models above with the model of "There is not a triangle" is to eliminate the first model, leaving only the second model and its new negated element:

¬△ ○

It follows that there is a circle. As this example shows, deductions can be made without the need for formal rules of inference of the sort postulated in "natural deduction systems" (see, for example, Rips, 1983; Braine, Reiser & Rumain, 1984), such as, in this case, the formal rule for disjunction:

A or B
not A
∴ B

(4) Information can be represented implicitly in order to reduce the load on working memory. An explicit representation makes information immediately available to other processes, whereas an implicit information encodes the information in a way that is not immediately accessible. Individuals and situations are represented implicitly by a propositional annotation that works in concert with an annotation for what has been represented exhaustively. Thus, the proper initial representation of the disjunction "Either there is a triangle or there is a circle, but not both" indicates that for the cases in which triangles occur, and the cases in which circles occur, have been exhaustively represented, as shown by the square brackets:

[△]

[○]

This set of models implicitly represents the fact that circles cannot occur in the first model and triangles cannot occur in the second model, because circles are exhaustively represented in the second model and triangles are exhaustively represented in the first model. Thus, a completely explicit set of models can be constructed by fleshing out the initial models to produce the set:

△ ¬○

¬△ ○

where there is no longer any need for square brackets because all the elements in the models have been exhaustively represented. The key to understanding implicit information is accordingly the process of fleshing out models explicitly, which is governed by two principles: first, when an element has been exhaustively represented (as shown by square brackets) in one or more models, add its negation to any other models; second, when a proposition has not been exhaustively represented, then add both it and its negation to separate models formed by fleshing out any model in which it does not occur. (Only the first principle is needed to flesh out the models of the disjunction above.)

(5) The epistemic status of a model can be represented by a propositional annotation; for example, a model represents a real possibility, a counterfactual state of affairs, or a deontic state.

A model that does not contain propositional annotations, that is, a model based on the first two assumptions above, represents a *set* of possible states of affairs, which contains an infinite number of possibilities (Barwise, 1993). Hence, the model above of the assertion "The circle is on the right of the triangle" corresponds to infinitely many possibilities; for example, the model is not specific about the distance apart of the two shapes. Any potential counterexample to a conclusion must be consistent with the premises, but the model itself does not enable the premises to be uniquely reconstructed. Hence, in verbal reasoning, there must be an independent record of the premises, which is assumed to be the linguistic representation from which the models are constructed. This record also allows the inferential system to ascertain just which aspects of the world the model represents; for example, a given model may, or may not, represent the distances apart of objects, but inspection of the model alone does not determine whether it represents distance. Experimental evidence bears out the psychological reality of both linguistic representations and mental models (see Johnson-Laird, 1983).

Models with propositional annotations compress sets of states of affairs in a still more powerful way: a single model now represents a finite *set* of alternative sets of situations. This aspect of mental models plays a crucial role in the account of syllogistic reasoning and reasoning with multiple quantifiers. For example, syllogistic premises of the form:

All the A are B
All the B are C

call for one model in which the number of As is small but arbitrary:

[[a] b] c
[[a] b] c
...

As are exhaustively represented in relation to Bs, Bs are exhaustively represented in relation to Cs, Cs are not exhaustively related, and the three dots designate implicit individuals of some other sort. This single model supports the conclusion:

All the A are C

and there are no counterexamples. The initial model, however, corresponds to eight distinct sets of possibilities depending on how the implicit individuals are fleshed out explicitly. There may, or may not, be individuals of each of the three following sorts:

individuals who are not-a, not-b, not-c
individuals who are not-a, not-b but c
individuals who are not-a, but b and c

These three binary contrasts accordingly yield eight alternatives, and each of them is consistent with an indefinite number of possibilities depending on the actual numbers of individuals of the different sorts (see also Garnham, 1993). In short, eight distinct potentially infinite sets have been compressed into a single model, which is used for the inference.

The theory of reasoning based on mental models makes three principal predictions. First, the greater the number of models that an inference calls for, the harder the task will be. This prediction calls for a theoretical account of the models postulated for a particular domain. Such accounts typically depend on independently motivated psycholinguistic principles; for example, negative assertions bring to mind the affirmative propositions that are denied (Wason, 1965). Second, erroneous conclusions will tend to be consistent with the premises rather than inconsistent with them. Reasoners will err because they construct some of the models of the premises – typically, just one model of them – and overlook other possible models. This prediction can be tested without knowing the detailed models postulated by the theory: it is necessary only to determine whether or not erroneous conclusions are consistent with the premises. Third, knowledge can influence the *process* of deductive reasoning: subjects will search more assiduously for alternative models when a putative conclusion is unbelievable than when it is believable. The first two of these predictions have been corroborated experimentally for all the main domains of deduction (for a review, see Johnson-Laird & Byrne, 1991, and for a reply to commentators, see Johnson-Laird & Byrne, 1993). The third prediction has been corroborated in the only domain in which it has so far been tested, namely, syllogistic reasoning (see, for example, Oakhill, Johnson-Laird, & Garnham, 1989). In contrast, theories of deduction

based on formal rules of inference exist only for spatial reasoning and reasoning based on sentential connectives (e.g., Rips, 1983; Braine, Reiser, & Rumain, 1984). Where the model theory and the formal rule theories make opposing predictions, the evidence so far has corroborated the model theory.

3. The strength of an inference

By definition, inductive arguments are logically invalid; that is, their premises could be true but their conclusions false. Yet such arguments differ in their strength – some are highly convincing, others are not. These differences are an important clue to the psychology of inference. However, one needs to distinguish between the strength of an argument – the degree to which its premises, if true, support the conclusion, and the degree to which the conclusion is likely to be true in any case. An argument can be strong but its conclusion improbable because the argument is based on improbable premises. Hence, the probability of the premises is distinct from the strength of the argument. In principle, the probability of a conclusion should depend on both the probability of the premises and the strength of the argument. But, as we shall see, individuals are liable to neglect the second of these components.

Osherson, Smith, and Shafir (1986) in a ground-breaking analysis of induction explored a variety of accounts of inferential strength that boil down to three main hypotheses: (1) an inference is strong if, given an implicit assumption, schema or causal scenario, it is logically valid; that is, the inference is an enthymeme (cf. Aristotle); (2) an inference is strong if it corresponds to a deduction in reverse, such as argument from specific facts to a generalization of them (cf. Hempel, 1965); and (3) an inference is strong if the predicates (or arguments) in premises and conclusion are similar (cf. Kahneman & Tversky, 1972). Each hypothesis has it advantages and disadvantages, but their strong points can be captured in the following analysis, which we will develop in two stages. First, the present section of the paper will specify an abstract characterization of the objective strength of an argument – *what* in theory has to be computed in order to determine the strength of an inference (the theory at the "computational" level). Second, the next section of the paper will specify *how* in practice the mind attempts to assess the strength of an argument (the theory at the "algorithmic" level).

The relation between premises and conclusion in inductive inference is a semantic one, and it can be characterized abstractly by adopting the semantic approach to logic (see, for example, Barwise & Etchemendy, 1989). An assertion such as "The circle is on the right of the triangle" is, as we have seen, true in infinitely many different situations; that is, the distance apart of the two shapes can differ, as can their respective sizes, shapes, textures and so on. Yet in all of

these different states the circle is on the right of the triangle. Philosophers sometimes refer to these different states as "possible worlds" and argue that an assertion is true in infinitely many possible worlds. We leave to one side the issue of whether or not possible worlds are countably infinite. The underlying theory has led to a powerful, though controversial, account of the semantics of natural language (see, for example, Montague, 1974).

Armed with the notion of possible states of affairs, we can define the notion of the strength of an inference in the following terms: a set of premises, including implicit premises provided by general and contextual knowledge, lend *strength* to a conclusion according to two principles:

(1) The conclusion is true in at least one of the possible states of affairs in which the premises are true; that is, the conclusion is at least consistent with the premises. If there is no such state of affairs, then the conclusion is inconsistent with the premises: the inference has no strength whatsoever, and indeed there is valid argument in favor of the negation of the conclusion.

(2) Possible states of affairs in which the premises are true but the conclusion false (i.e., counterexamples) weaken the argument. If there are no counterexamples, then the argument is maximally strong – the conclusion follows validly from the premises. If there are counterexamples, then the strength of the argument equals the proportion of states of affairs consistent with the premises in which the conclusion is also true.

This account has a number of advantages. First, it embraces deduction and induction within the same framework. What underlies deduction is the semantic principle of validity: an argument is valid if its conclusion is true in any state of affairs in which its premises are true. An induction increases semantic information and so its conclusion must be false in possible cases in which its premises are true. Hence, inductions *are* reverse deductions, but they are the reverse of deductions that throw semantic information away.

Second, the probability of any one distinct possible state of affairs (possible world) is infinitesimal, and so it is reasonable to assume that possible states of affairs are close to equi-possible. It follows that a method of integrating the area of a subset of states of affairs provides an extensional foundation for probabilities. The strength of an inference is accordingly equivalent to the probability of the conclusion given the premises. It is 1 in the case of a valid deduction, 0 in the case of a conclusion that is inconsistent with the premises, and an intermediate value for inductions. The two abstract principles, however, are not equivalent to the probability calculus: as we shall see, the human inferential system can attempt to assess the relevant proportions without necessarily using the probability calculus. Likewise, the principles have no strong implications for the correct interpretation of probability, which is a matter for self-conscious philosophical reflection. The

principles are compatible with interpretations in terms of actuarial frequencies of events, equi-possibilities based on physical symmetry, and subjective degrees of belief (cf. Ramsay, 1926; Hintikka's, 1962, analysis of beliefs in terms of possibility; and for an alternative conception, see Shafer & Tversky's, 1985, discussion of "belief functions"). Hence, an argument (or a probability) may concern either a set of events or a unique event. Individuals who are innumerate may not assign a numerical degree of certainty to their conclusion, and even numerate individuals may not have a tacit mental number representing their degree of belief. Individuals' beliefs do differ in subjective strength, but it does not follow that such differences call for a mental representation of numerical probabilities. An alternative conception of "degrees of belief" might be based on analogue representations (cf. Hintzman, Nozawa, & Irmscher, 1982), or on a system that permitted only partial rankings of strengths, such as one that recorded the relative ease of constructing different classes of models.

Third, the account is compatible with semantic information. The semantic information conveyed by a proposition, A, equals $1 - p(\mathrm{A})$, where $p(\mathrm{A})$ denotes the probability of A (Bar-Hillel & Carnap, 1964; Johnson-Laird, 1983). If A is complex proposition containing conjunctions, disjunctions, etc., its probability can be computed in the usual way according to the probability calculus. Hence, as argued elsewhere (Johnson-Laird, 1993), we can distinguish between deduction and induction on the basis of semantic information, that is, the possible states of affairs that a proposition rules out as false. Deduction does not increase semantic information; that is, the conclusion of a valid deduction rules out the same possibilities as the premises or else fewer possibilities, and so the conclusion must be true given that the premises are true. Induction increases semantic information; that is, the conclusion of an induction goes beyond the premises (including those tacit premises provided by general knowledge) by ruling out at least some additional possibility over and above the states of affairs that they rule out. This account captures all the standard cases of induction, such as the generalization from a finite set of observations to a universal claim (for a similar view, see Ramsay, 1926).

Fourth, the account is compatible with everyday reasoning and argumentation. One feature of such informal argumentation is that it typically introduces both a case for a conclusion and a case against it – a procedure that is so unlike a logical proof that many theorists have supposed that logic is useless in the analysis of everyday reasoning (e.g, Toulmin, 1958). The *strength* of an argument, however, can be straightforwardly analyzed in the terms described above: informal argumentation is typically a species of induction, which may veer at one end into deduction and at the other end into a creative process in which one or more premises are abandoned. Thus, a case for a conclusion may depend on several inductive arguments of differing strength.

The obvious disadvantage of the account is that it is completely impractical. No

one can consider all the infinitely many states of affairs consistent with a set of premises. No one can integrate all those states of affairs in which the conclusion is true and all those states of affairs in which it is false. Inference with quantifiers has no general decision procedure; that is, proofs for valid theorems can always be found in principle, but demonstrations of invalidity may get lost in the "space" of possible derivations. Inference with sentential connectives has a decision procedure, but the formulation of parsimonious conclusions that maintain semantic information is not computationally tractable; that is, as premises contain more atomic propositions, it takes exponentially longer to generate such conclusions (given that NP $\neq$ P). So how does this account translate into a psychological mechanism for assessing the strength of an argument? It is this problem that the theory of mental models is designed to solve.

4. Mental models and estimates of inferential strength

Philosophers have tried to relate probability and induction at a deep level (see, for example, Carnap, 1950), but as far as cognitive psychology is concerned they are overlapping rather than identical enterprises: there are probabilistic inferences that are not inductive, and there are inductive inferences that are not probabilistic. Here, for example, is a piece of probabilistic reasoning that is deductive:

The probability of heads is 0.5.
The probability of the date uppermost given heads is 1.
The probability of the date uppermost given tails is 0.
Hence, the probability of the date uppermost is 0.5.

This deduction makes explicit what is implicit in the premises, and it does not increase their semantic information. A more mundane example is as follows:

If you park illegally within the walls of Siena, you will probably have your car towed.
Phil has parked illegally within the walls of Siena.
Phil will probably have his car towed.

This inference is also a valid deduction. Conversely, many inductive inferences are not probabilistic; that is, they lead to conclusions that people hold to be valid. For example, the engineers in charge at Chernobyl inferred initially that the explosion had not destroyed the reactor (Medvedev, 1990). Such an event was unthinkable from their previous experience, and they had no evidence to suppose that it had occurred. They were certain that the reactor was intact, and their

conviction was one of the factors that led to the delay in evacuating the inhabitants of the nearby town. Of course people do make probabilistic inductions, and it is necessary to explain their basis as well as the basis for probabilistic deductions. To understand the application of the model theory to the assessment of strength, it will be helpful to consider first how it accounts for deductions based on probabilities.

Critics sometimes claim that models can be used only to represent alternative states of affairs that are treated as equally likely. In fact, there is no reason to suppose that when individuals construct or compare models they take each model to be equally likely. To illustrate the point, consider an example of a deduction leading to a probabilistic conclusion:

Kropotkin is an anarchist.
Most anarchists are bourgeois.
∴ Probably, Kropotkin is bourgeois.

The quantifier "most" calls for a model that represents a proportion (see Johnson-Laird, 1983, p. 137). Thus, a model of the second premise takes the form:

```
[a]   b
[a]   b
[a]   b
[a]
  ...
```

where the set of anarchists is exhaustively represented; that is, anarchists cannot occur in fleshing out the implicit model designated by the three dots. When the information in the first premise is added to this model, one possible model is:

```
k   [a]   b
    [a]   b
    [a]   b
    [a]
    ...
```

in which Kropotkin is bourgeois. Another possible model is:

```
    [a]   b
    [a]   b
    [a]   b
k   [a]
    ...
```

in which Kropotkin is *not* bourgeois. Following Aristotle, assertions of the form: *probably S*, can be treated as equivalent to: *in most possible states of affairs*, *S*. And in most possible states of affairs as assessed from models of the premises, Kropotkin is bourgeois. Hence, the inferential system needs to keep track of the relative frequency with which the two sorts of models occur. It will detect the greater frequency of models in which it Kropotkin is bourgeois, and so it will deduce:

Probably, Kropotkin is bourgeois.

Individuals who are capable of one-to-one mappings but who have no access to cardinal or ordinal numbers will still be able to make this inference. They have merely to map each model in which S occurs one-to-one with each model in which S does not occur, and, if there is a residue, it corresponds to the more probable category. Likewise, there are many ways in principle in which to estimate the relative frequencies of the two sorts of model – from random sampling with replacement to systematic explorations of the "space" of possible models. The only difference in induction is that information that goes beyond the premises (including those in tacit knowledge) is added to models on the basis of various constraints (see Johnson-Laird, 1983).

The *strength* of an inference depends, as we have seen, on the relative proportions of two sorts of possible states of affairs consistent with the premises: those in which the conclusion is true and those in which it is false. Reasoners can estimate these proportions by constructing models of the premises and attending to the proportions with which the two sorts of models come to mind, and perhaps to the relative ease of constructing them. For example, given that Evelyn fell (without a parachute) from an airplane flying at a height of 2000 feet, then most individuals have a prior knowledge that Evelyn is likely to be killed, but naive individuals who encounter such a case for the first time can infer the conclusion. The inference is strong, but not irrefutable. They may be able to imagine cases to the contrary; for example, Evelyn falls into a large haystack, or a deep snow drift. But, in constructing models (of sets of possibilities), those in which Evelyn is killed will occur much more often than those in which Evelyn survives – just as models in which Kropotkin is bourgeois outnumber those in which he is not. Insofar as individuals share available knowledge, their assessments of probabilities should be consistent.

This account is compatible with the idea of estimating likelihoods in terms of scenarios, which was proposed by Tversky and Kahneman (1973, p. 229), and it forms a bridge between the model theory and the heuristic approach to judgements of probability. Estimates of the relative proportions of the two sorts of models – those in which a conclusion is true and those in which it is false – will

be rudimentary, biased and governed by heuristics. In assessing outcomes dependent on sequences of events, models must allow for alternative courses of events. They then resemble so-called "event trees", which Shafer (1993) argues provide a philosophical foundation to probability and its relations to causality. Disjunctive alternatives, however, are a source of difficulty both in deduction (see, for example, Johnson-Laird & Byrne, 1991) and in choice (see, for example, Shafir & Tversky, 1992).

5. Some empirical consequences of the theory

The strength of an argument depends on the relation between the premises and the conclusion, and, in particular, on the proportion of possibilities compatible with the premises in which the conclusion is true. This relation is *not* in general a formal or syntactic one, but a semantic one. It takes work to estimate the strength of relation, and the theory yields a number of predictions about making and assessing inductive inferences. The main predictions of the theory are as follows:

First, arguments – especially in daily life – do not wear their logical status on their sleeves, and so individuals will tend to approach deductive and inductive arguments alike. They will tend to confuse an inductive conclusion, that is, one that could be true given the premises, with a deductive conclusion, that is, one that must be true given the premises. They will tend to construct one or two models, draw a conclusion, and be uncertain about whether it follows of necessity.

Second, envisioning models, which each correspond to a class of possibilities, is a crude method, and, because of the limited processing capacity of working memory, many models are likely never to be envisaged at all. The process will be affected by several constraints. In particular, individuals are likely to seek the most specific conclusion consistent with the premises (see Johnson-Laird, 1993), they are likely to seek parsimonious conclusions (see Johnson-Laird & Byrne, 1991), and they are likely to be constrained by the availability of relevant knowledge (Tversky & Kahneman, 1973). The model theory postulates a mechanism for making knowledge progressively available. Reasoners begin by trying to form a model of the current situation, and the retrieval of relevant knowledge is easier if they can form a single model containing all the relevant entities. Once they have formed an initial model, knowledge becomes available to them in a systematic way. They manipulate the spatial or physical aspects of the situation; that is, they manipulate the model directly by procedures corresponding to such changes. Next, they make more abstract conceptual manipulations; for example, they consider the properties of superordinate concepts of entities in the model. Finally, they make still more abstract inferences based on introducing

relations retrieved from models of analogous situations (cf. Gentner, 1983). Consider the following illustration:

> Arthur's wallet was stolen from him in the restaurant. The person charged with the offense was outside the restaurant at the time of the robbery. What follows?

Reasoners are likely to build an initial model of Arthur inside the restaurant when his wallet is stolen and the suspect outside the restaurant at that time. They will infer that the suspect is innocent. They may then be able to envisage the following sort of sequence of ideas from their knowledge about the kinds of things in the model:

(1) Physical and spatial manipulations:
    The suspect leant through the window to steal the wallet.
    The suspect stole the wallet as Arthur was entering the restaurant, or ran in and out of the restaurant very quickly (creative inferences that, in fact, are contrary to the premises).
(2) Conceptual manipulations:
    The suspect had an accomplice – a waiter, perhaps – who carried out the crime (theft is a crime, and many crimes are committed by accomplices).
(3) Analogical thinking
    The suspect used a radio-controlled robot to sneak up behind Arthur to take the wallet (by analogy with the use of robots in other "hazardous" tasks).

In short, the model theory predicts that reasoners begin by focusing on the initial explicit properties of their model of a situation, and then they attempt to move away from them, first by conceptual operations, and then by introducing analogies from other domains. It is important to emphasize that the order of the three sorts of operations is not inflexible, and that particular problems may elicit a different order of operations. Nevertheless, there should be a general trend in moving away from explicit models to implicit possibilities.

Third, reasoners are also likely to be guided by other heuristics, which have been extensively explored by Tversky and Kahneman, and their colleagues. These heuristics can be traced back to Hume's seminal analysis of the connection between ideas: "there appear to be only three principles of connexion between ideas, namely, *Resemblance*, *Contiguity* in time or place, and *Cause* or *Effect*" (Hume, 1748, Sec. III). Hence, semantic similarity between the premises and the conclusion, and the causal cohesiveness between them, will influence probabilistic judgements. Such factors may even replace extensional estimates based on models.

Fourth, individuals should be inferential satisficers; that is, if they reach a credible (or desirable) conclusion, or succeed in constructing a model in which such a conclusion is true, they are likely to accept it, and to overlook models that are counterexamples. Conversely, if they reach an incredible (or undesirable) conclusion, they are likely to search harder for a model of the premises in which it is false. This propensity to satisfice will in turn lead them to be overconfident in their conclusions, especially in the case of arguments that do have alternative models in which the conclusion is false. Individuals are indeed often overconfident in their inductive judgements, and Gigerenzer, Hoffrage, and Kleinbölting (1991) have propounded a theory of "probabilistic mental models" to account for this phenomenon. These are long-term representations of probabilistic cues and their validities (represented in the form of conditional probabilities). These authors propose that individuals use the single cue with the strongest validity and do not aggregate multiple cues, and that their confidence derives from the validity of this cue. They report corroboratory evidence from their experiments on the phenomenon of overconfidence; that is, rated confidence tends to be higher than the actual percentage of correct answers. As Griffin and Tversky (1992) point out, however, overconfidence is greater with harder questions and this factor provides an alternative account of Gigerenzer et al.'s results. In contrast, the model theory proposes that the propensity to satisfice should lead subjects to overlook models in the case of multiple-model problems, and so they should tend to be more confident than justified in the case of harder problems. Overconfidence in inductive inference occurred in an unpublished study by Johnson-Laird and Anderson, in which subjects were asked to draw initial conclusions from such premises as:

> The old man was bitten by a poisonous snake. There was no known antidote available.

They tend initially to infer that the old man died. Their confidence in such conclusions was moderately high. They were then asked whether there were any other possibilities and they usually succeeded in thinking of two or three. When they could go no further, they were asked to rate again their initial conclusions, and showed a reliable decline in confidence. Hence, by their own lights, they were initially overconfident, though by the end of the experiment they may have been underconfident as a result of bringing to mind remote scenarios. With easier one-model problems, the error and its correlated overconfidence cannot occur. But should subjects be underconfident in such cases, as is sometimes observed? One factor that may be responsible for the effect in repeated-measure designs is the subjects' uncertainty about whether or not there might be other models in a one-model case.

Finally, individuals are likely to focus on what is explicit in their initial models and thus be susceptible to various "focusing effects" (see Legrenzi, Girotto, &

Johnson-Laird, 1993). These effects include difficulty in isolating genuinely diagnostic data (see, for example, Beyth-Marom & Fischhoff, 1983; Doherty, Mynatt, Tweney, & Schiavo, 1979), testing hypotheses in terms of their positive instances (Evans, 1989; Klayman & Ha, 1987), neglect of base rates in certain circumstances (Tversky & Kahneman, 1982), and effects of how problems in deductive and inductive reasoning are framed (e.g., Johnson-Laird & Byrne, 1989; Tversky & Kahneman, 1981). Focusing is also likely to lead to too great a reliance on the credibility of premises (and conclusion) and too little on the strength of the argument, that is, the relation between the premises and conclusion. Reasoners will build an initial model that makes explicit the case for a conclusion, and then fail to adjust their estimates of its likelihood by taking into account alternative models (see also Griffin & Tversky, 1992, for an analogous view). Conversely, any factor that makes it easier for individuals to flesh out explicit models of the premises should improve performance.

6. Rules for probabilistic thinking

An obvious potential basis for probabilistic reasoning is the use of rules of inference, such as:

If q & r then s (with probability p)
$\therefore$ If q then s (with probability p')

Numerous AI programs include rules of this sort (see, for example, Holland, Holyoak, Nisbett, & Thagard, 1986; Michalski, 1983; Winston, 1975). The most plausible psychological version of this idea is due to Collins and Michalski (1989). They argue that individuals construct mental models on the basis of rules of inference, and that these rules have numerical parameters for such matters as degree of certainty. They have not tried to formalize all patterns of plausible inference, but rather some patterns of inference that make up a core system of deductions, analogies and inductions. They admit that it is difficult to use standard psychological techniques to test their theory, which is intended to account only for people's answers to questions. It does not make any predictions about the differences in difficulty between various sorts of inference, and, as they point out (p. 7), it does not address the issue of whether people make systematic errors. Hence, their main proposed test consists in trying to match protocols of arguments against the proposed forms of rules. Pennington and Hastie (1993) report success in matching these patterns to informal inferences of subjects playing the part of trial jurors. But, as Collins and Michalski mention, one danger is that subjects' protocols are merely rationalizations for answers arrived at by other means. In sum, AI rule systems for induction have not yet received decisive corroboration.

In contrast, another sort of rule theory has much more empirical support. This theory appeals to the idea that individuals have a tacit knowledge of such rules as the "law of large numbers" (see Nisbett, 1993; Smith, Langston, & Nisbett, 1992). Individuals apply the rules to novel materials, mention them in justifying their responses, benefit from training with them, and sometimes overextend their use of them. The rules in AI programs are formal and can be applied to the representation of the abstract logical form of premises. The law of large numbers, however, is *not* a formal rule of inference. It can be paraphrased as follows:

> The larger the sample from a population the smaller its mean is likely to diverge from the population mean.

Aristotle would not have grasped such notions as *sample*, *mean* and *population*, but he would have been more surprised by a coin coming up heads ten times in a row than a coin coming up heads three times in a row. He would thus have had a tacit grasp of the law that he could make use of in certain circumstances. The law has a rich semantic content that goes well beyond the language of logical constants, and it is doubtful whether it could be applied to the logical form of premises. On the contrary, it is likely to be applied only when one has grasped the content of a problem, that is, constructed a model that makes explicit that it calls for an estimate based on an example.

Individuals are likely to hold many other general principles as part of their beliefs about probability. For instance, certain devices produce different outcomes on the basis of chance, that is, at approximately equal rates and in unpredictable ways; if a sample from such a device is deviant, things are likely to even up in the long run (gambler's fallacy). Such principles differ in generality and validity, but they underlie the construction of many probabilistic judgements. The fact that individuals can be taught correct laws and that they sometimes err in over-extending them tells us nothing about the mental format of the laws. They may take the form of schemas or content-specific rules of inference, but they could be represented declaratively. Likewise, how they enter into the process of thinking – the details of the computations themselves – is also unknown. There is, however, no reason to oppose them to mental models. They seem likely to work together in tandem, just as conceptual knowledge must underlie the construction of models.

7. Conclusions

The principle thesis of the present paper is that general knowledge and beliefs, along with descriptions of situations, lead to mental models that are used to assess probabilities. Most cognitive scientists agree that humans construct mental

representations; many may suspect that the model theory merely uses the words "mental model" where "mental representation" would do. So, what force, if any, is there to the claim that individuals think probabilistically by manipulating models? The answer, which has been outlined here, is twofold. First, the representational principles of models allow sets of possibilities to be considered in a highly compressed way, and even in certain cases sets of sets of possibilities. Hence, it is feasible to assess probability by estimating possible states of affairs within a general framework that embraces deduction, induction and probabilistic thinking. This framework provides an extensional foundation of probability theory that is not committed *a priori* to either a frequency or degrees-of-belief interpretation, which are both equally feasible on this foundation. Second, the model theory makes a number of predictions based on the distinction between explicit and implicit information, and on the processing limitations of working memory. Such predictions, as the study of deduction has shown, are distinct from those made by theories that postulate only representations of the logical form of assertions.

References

Aristotle (1984). *The complete works of Aristotle*, edited by J. Barnes, 2 vols. Princeton: Princeton University Press.

Bar-Hillel, Y., & Carnap, R. (1964). An outline of a theory of semantic information. In Y. Bar-Hillel, (Ed.), *Language and information*. Reading, MA: Addison-Wesley.

Barwise, J. (1993). Everyday reasoning and logical inference. *Behavioral and Brain Sciences*, *16*, 337–338.

Barwise, J., & Etchemendy, J. (1989). Model-theoretic semantics. In M.I. Posner (Ed.), *Foundations of cognitive science*. Cambridge, MA: MIT Press.

Beyth-Marom, R., & Fischhoff, B. (1983). Diagnosticity and pseudodiagnosticity. *Journal of Personality and Social Psychology*, *45*, 1185–1197.

Braine, M.D.S., Reiser, B.J., & Rumain, B. (1984). Some empirical justification for a theory of natural propositional logic. In *The psychology of learning and motivation*, (Vol. 18). New York: Academic Press.

Byrne, R.M.J. (1989). Suppressing valid inferences with conditionals. *Cognition*, *31*, 61–83.

Carnap, R. (1950). *Logical foundations of probability*. Chicago: Chicago University Press.

Collins, A.M., & Michalski, R. (1989). The logic of plausible reasoning: A core theory. *Cognitive Science*, *13* 1–49.

Craik, K. (1943). *The nature of explanation*. Cambridge, UK: Cambridge University Press.

Doherty, M.E., Mynatt, C.R., Tweney, R.D., & Schiavo, M.D. (1979). Pseudodiagnosticity. *Acta Psychologica*, *43*, 11–21.

Evans, J.St.B.T. (1989). *Bias in human reasoning: Causes and consequences*. London: Erlbaum.

Fisher, A. (1988). *The logic of real arguments*. Cambridge, UK: Cambridge University Press.

Garnham, A. (1987). *Mental models as representations of discourse and text*. Chichester: Ellis Horwood.

Garnham, A. (1993). A number of questions about a question of number. *Behavioral and Brain Sciences*, *16*, 350–351.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*, 155–170.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98*, 506–528.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology, 24*, 411–435.

Hempel, C. (1965). *Aspects of scientific explanation*. New York: Macmillan.

Hintikka, J. (1962). *Knowledge and belief: An introduction to the logic of the two notions*. Ithaca: Cornell University Press.

Hintzman, D.L., Nozawa, G., & Irmscher, M. (1982). Frequency as a nonpropositional attribute of memory. *Journal of Verbal Learning and Verbal Behavior, 21*, 127–141.

Holland, J.H., Holyoak, K.J., Nisbett, R.E., & Thagard, P. (1986). *Induction: Processes of inference, learning, and discovery*. Cambridge, MA: MIT Press.

Hume, D. (1748/1988). *An enquiry concerning human understanding*. La Salle, IL: Open Court.

Johnson-Laird, P.N. (1983). *Mental models: Towards a cognitive science of language, inference and consciousness*. Cambridge, UK: Cambridge University Press.

Johnson-Laird, P.N. (1993). *Human and machine thinking*. Hillsdale, NJ: Erlbaum.

Johnson-Laird, P.N., & Bara, B. (1984). Syllogistic inference. *Cognition, 16*, 1–61.

Johnson-Laird, P.N., & Byrne, R.M.J. (1989). *Only* reasoning. *Journal of Memory and Language, 28*, 313–330.

Johnson-Laird, P.N., & Byrne, R.M.J. (1991). *Deduction*. Hillsdale, NJ: Erlbaum.

Johnson-Laird, P.N., & Byrne, R.M.J. (1993). Authors' response [to multiple commentaries on *Deduction*]: Mental models or formal rules? *Behavioral and Brain Sciences, 16*, 368–376.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*, 430–454.

Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation and information in hypothesis testing. *Psychological Review, 94*, 211–228.

Legrenzi, P., Girotto, V., & Johnson-Laird, P.N. (1993). Focussing in reasoning and decision making. *Cognition, 49*, 37–66.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: W.H. Freeman.

Medvedev, Z.A. (1990). *The legacy of Chernobyl*. New York: Norton.

Michalski, R.S. (1983). A theory and methodology of inductive learning. In R.S. Michalski, J.G. Carbonell, & T.M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach*. Los Altos, CA: Morgan Kaufmann.

Montague, R. (1974). *Formal philosophy: Selected papers*. New Haven: Yale University Press.

Nisbett, R.E. (Ed.) (1993). *Rules for reasoning*. Hillsdale, NJ: Erlbaum.

Oakhill, J.V., Johnson-Laird, P.N. & Garnham, A. (1989). Believability and syllogistic reasoning. *Cognition, 31*, 117–140.

Osherson, D.N., Smith, E.E. & Shafir, E. (1986). Some origins of belief. *Cognition, 24*, 197–224.

Pennington, N., & Hastie, R. (1993). Reasoning in explanation-based decision making. *Cognition, 49*, 123–163.

Polk, T.A., & Newell, A. (1988). Modeling human syllogistic reasoning in Soar. In *Tenth Annual Conference of the Cognitive Science Society* (pp. 181–187). Hillsdale, NJ: Erlbaum.

Ramsay, F.P. (1926/1990). Truth and probability. In D.H. Mellor, (Ed.), *F.P. Ramsay: Philosophical papers*. Cambridge, UK: Cambridge University Press.

Rips, L.J. (1983). Cognitive processes in propositional reasoning. *Psychological Review, 90*, 38–71.

Shafer, G. (1993). *Using probability to understand causality*. Unpublished MS, Rutgers University.

Shafer G., & Tversky, A. (1985). Languages and designs for probability judgment. *Cognitive Science, 9*, 309–339.

Shafir, E., & Tversky, A. (1992). Thinking through uncertainty: Nonconsequential reasoning and choice. *Cognitive Psychology, 24*, 449–474.

Smith, E.E., Langston, C., & Nisbett, R.E. (1992). The case for rules in reasoning. *Cognitive Science, 16*, 1–40.

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences, 12*, 435–502.

Toulmin, S.E. (1958). *The uses of argument*. Cambridge, UK: Cambridge University Press.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*, 207–232.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*, 453–458.

Tversky, A., & Kahneman, D. (1982). Evidential impact of base rates. In D. Kahneman, P. Slovic, & A. Tversky, (Eds.), *Judgments under uncertainty: Heuristics and biases*. New York: Cambridge University Press.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–315.

Wason, P.C. (1965). The contexts of plausible denial. *Journal of Verbal Learning and Verbal Behavior*, *4*, 7–11.

Winston, P.H. (1975). Learning structural descriptions from examples. In P.H. Winston, (Ed.), *The psychology of computer vision*. New York: McGraw-Hill.