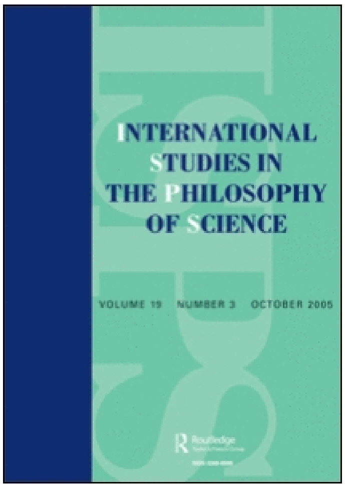


International Studies in the Philosophy of Science
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713427740

A model theory of induction

Philip N. Johnson-Laird [a]
[a] Department of Psychology, Princeton University, New Jersey, USA

To cite this Article Johnson-Laird, Philip N.(1994) 'A model theory of induction', International Studies in the Philosophy of Science, 8: 1, 5 — 29

To link to this Article: DOI: 10.1080/02698599408573474

URL: http://dx.doi.org/10.1080/02698599408573474

A model theory of induction

PHILIP N. JOHNSON-LAIRD
Department of Psychology, Princeton University, New Jersey 08544, USA

Abstract *Theories of induction in psychology and artificial intelligence assume that the process leads from observation and knowledge to the formulation of linguistic conjectures. This paper proposes instead that the process yields mental models of phenomena. It uses this hypothesis to distinguish between deduction, induction, and creative forms of thought. It shows how models could underlie inductions about specific matters. In the domain of linguistic conjectures, there are many possible inductive generalizations of a conjecture. In the domain of models, however, generalization calls for only a single operation: the addition of information to a model. If the information to be added is inconsistent with the model, then it eliminates the model as false: this operation suffices for all generalizations in a Boolean domain. Otherwise, the information that is added may have effects equivalent (a) to the replacement of an existential quantifier by a universal quantifier, or (b) to the promotion of an existential quantifier from inside to outside the scope of a universal quantifier. The latter operation is novel, and does not seem to have been used in any linguistic theory of induction. Finally, the paper describes a set of constraints on human induction, and outlines the evidence in favor of a model theory of induction.*

Introduction

Induction is part of both everyday and scientific thinking. It enables us to understand the world and to predict events. It can also mislead us. Many of the cognitive failures that have led to notable disasters are inductions that turned out to be wrong. For instance, when the car ferry, Herald of Free Enterprise, sailed from the Belgian port of Zeebrugge on the 6 March, 1987, the master made the plausible induction that the bow doors had been closed. They had always been closed in the past, and there was no evidence to the contrary. The chief officer made the same induction, as did the bosun. But, the assistant bosun, whose job it was to close the doors, was asleep in his bunk, and had not closed the doors. Shortly after leaving the harbor, the vessel capsized and sank, and 188 people drowned. Induction is indeed an important but risky business. If psychologists had a better understanding of the strengths and weakness of human inductive competence, then they might be able to help individuals to perform more skillfully and to introduce more effective measures—especially by way of advisory computer systems—to prevent inductive disasters.

Induction is also a theoretically confusing business. Some authors restrict the term to very narrow cases; others outlaw it altogether. Textbooks often define it as leading from particular premises to a general conclusion, in contrast to deduction, which they define as leading from general premises to a particular conclusion. In fact, induction can lead from particular observations to a particular conclusion—as it did in the case of the Herald of Free Enterprise, and deduction can lead from general premises to a general conclusion. The first goal of this paper is accordingly to draw a principled distinction between induction, deduction, and other forms of thought. Its second goal is to distinguish between varieties of induction. And its third goal is to outline a new theory of induction. This theory departs from the main tradition in psychology and philosophy, which treats induction as a process yielding plausible *verbal* generalizations or hypotheses. It proposes instead that induction generates *mental models* of domains. As we shall see, this distinction is not trivial, and it turns out to have some unexpected consequences.

An outline of the theory of mental models

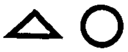
The central idea in the theory of mental models is that the process of understanding yields a model (Johnson-Laird, 1983). Unlike other proposed forms of mental representation, such as propositional representations or semantic networks, models are based on the fundamental principle that their structure corresponds to the way in which human beings conceive the structure of the world. This principle has three important corollaries:

(1) Entities are represented by corresponding tokens in mental models. Each entity is accordingly represented only once in a mental model.

(2) The properties of entities are represented by the properties of tokens representing entities.

(3) Relations among entities are represented by relations among the tokens representing entities.

Thus, a model of the assertion, "The circle is on the right of the triangle" has the following structure:



A model may be experienced as a visual image, but what matters is, not the subjective experience, but the structure of the model: entities are represented by tokens, their properties are represented by properties of the tokens, and the relations between them are represented by the relations between the tokens.

As an illustration of the theory and of its implications for the mental representation of concepts, I will consider its implementation in a program for spatial reasoning that generates models like the one above (Johnson-Laird & Byrne, 1991). The program constructs three-dimensional models on the basis of verbal assertions. It has a lexicon in which each word has a analysis of its meaning into primitive constituents, which I shall refer to as *subconcepts*. It has a grammar in which each rule has a corresponding semantic principle for forming combinations of subconcepts. As the program parses a sentence, it assembles subconcepts to form a representation of the sentence's meaning. This *propositional representation* is then used by other procedures to construct a model of a particular situation described by the sentence.

Given a noun-phrase such as "the circle", the program uses the subconcept underlying *circle* to set up a simple model:

And given the assertion:

The circle is on the right of the triangle

the parsing process combines the subconcepts underlying the words in the sentence to yield the following result, which represents the meaning of the assertion:

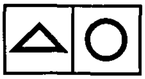
((1 0 0) (O)(△))

The meaning of the relation x *on the right of* y is a set of subconcepts that consists of values for incrementing y's Cartesian co-ordinates to find a location for x:

1 0 0

The 1 indicates that x should be located by incrementing y's value on the left-right dimension whilst holding y's values on the front-back and up-down dimensions constant, i.e. adding 0s to them.

What the program does with a propositional representation of the meaning of a sentence depends on context. If the assertion is the first in a discourse, the program uses the representation to construct a complete model within a minimal array:

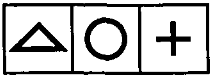


The reader will note that an assertion about the relation between two entities with distinct properties is represented by a model in which there is a relation between two entities with distinct properties.

Depending on the current state of any existing models, the program can also use the propositional representation to add an entity to a model, to combine two previously separate models, to make a valid deduction, or to make a non-monotonic inference. For example, the program can make a transitive deduction, such as:

The circle is on the right of the triangle.
The cross is on the right of the circle.
∴ The cross is on the right of the triangle.

without relying on any explicit statement of transitivity. It uses the subconcepts for *on the right of* to construct the model:



It verifies the conclusion in the model, and is unable to find an alternative model of the premises in which the conclusion is false. In summary, subconcepts combine to form propositional representations that can be used by many different procedures for constructing and manipulating models.

The concept of *on the right of* is part of a system based on the same underlying set of subconcepts:

on the right of:	1	0	0
on the left of:	− 1	0	0
in front of:	0	1	0
behind:	0	− 1	0
above:	0	0	1
below:	0	0	− 1

The theory postulates that some such system allows human reasoners to set up spatial models and to manipulate them. It must exist prior to the mastery of any particular spatial relation, and can be used to acquire new high-level concepts. For example, one might acquire the relation represented by (1 0 1), roughly *diagonally up and to the right*, if it played an important part in spatial thinking and was accordingly dignified by a single spatial term. The subconceptual system also provides individuals with an idealized taxonomy. In the real world, objects do not have to be perfectly aligned, and so a judgement of the relation between them may compare their actual co-ordinates with alternative possibilities in the taxonomy. Hence, the extension of a relation depends not just on its subconceptual analysis but also on other concepts in the same taxonomy.

The theory of mental models extends naturally to the representation of sentential connectives, such as *and*, *if*, and *or*, and quantifiers, such as *any*, and *some*. The theory posits that models represent as little as possible explicitly. Hence, the initial representation of a conditional, such as "if there is an A then there is a 2", is by the following two models:

```
[A]  2
  . . .
```

The first line represents an explicit model of the situation in which the antecedent is true, and the second line represents an implicit model of the alternative situation(s). The second model is implicit because it has no immediately available content, but it can be fleshed out to make its implicit content explicit. The square brackets around the A in the first model are an "annotation" indicating that the A has been exhaustively represented, i.e. it cannot occur in any other model (for a defense of such annotations, see Newell, 1990; Johnson-Laird & Byrne, 1991). The implicit model can be, and in certain circumstances is, fleshed out explicitly. The fleshing out can correspond to a bi-conditional, "if and only if there is an A then there is a 2:

```
 A    2
¬A   ¬2
```

where "¬" is an annotation representing negation. Alternatively, the fleshing out takes the weaker conditional form:

```
 A    2
 A   ¬2
¬A   ¬2
```

There are similar models that represent the other sentential connectives, such as *or*, *only if*, *unless* (see Johnson-Laird & Byrne, 1991, for the evidence for the psychological reality of these models).

The representation of quantifiers is also a natural extension of the theory. An assertion such as, "Some of the athletes are bakers", has the following single model:

```
a    b
a    b
a
     b
 . . .
```

where, unlike the previous diagrams, each line now represents a separate individual in the *same* model of a state of affairs: "a" denotes a representation of an athlete and "b" denotes a representation of a baker. The number of tokens representing individuals is arbitrary. The final line represents implicit individuals, who may be of some other sort. The statement, "All of the athletes are bakers", has the following initial model:

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

The square brackets represent that the athletes have been exhaustively represented (in relation to the bakers). Similar interpretations are made for other quantifiers and for assertions that contain more than one quantifier, such as "None of the Avon letters is in the same place as any of the Bury letters". Undoubtedly, the best success of the theory of mental models has been in accounting for the phenomena of the comprehension of discourse and the phenomena of deductive reasoning. The theory rejects the idea that discourse is encoded in a semantic network or in any other way that represents merely the meanings of expressions. What is represented, as experiments have corroborated (see e.g. Garnham, 1987) are referents, their properties, and the relations among them. The theory rejects the idea that deduction depends on formal rules of inference. It proposes instead that reasoners construct models of premises, draw conclusions from them, and search for alternative models of the premises that might falsify these conclusions. It makes two principal predictions about deduction: the major cause of difficulty of making deductions is the need to consider models of alternative possibilities; the most likely errors are conclusions that overlook such alternatives. These predictions have been corroborated in all the main domains of deductive reasoning, including propositional, relational, and quantificational inferences (Johnson-Laird & Byrne, 1991). We now turn to the application of the model theory to induction, and we begin by using it to help to draw a systematic distinction between induction and deduction.

Induction, deduction and semantic information

A simple way in which to distinguish induction, deduction, and other forms of thought, depends on semantic information, that is, the models of possible states of affairs that a proposition rules out as false (see Bar-Hillel & Carnap, 1964; Johnson-Laird, 1983). For example, the proposition, "The battery is dead or the voltmeter is faulty, or both", has the following three explicit models of alternative possibilities:

d f
d ¬ f
¬ d f

For simplicity, I am here using single letters to denote entities with particular properties: d represents "the battery is dead", and f represents "the voltmeter is faulty", and, as before, "¬" represents negation. Each line denotes a model of a different situation, and so the disjunction eliminates only one out of four possibilities: the situation where there is neither a dead battery nor a faulty voltmeter: ¬ d ¬ f. The categorical assertion, "The battery is dead", eliminates two models out of the four possibilities, ¬ d f, and ¬ d ¬ f, and so it has a greater information content. And the conjunction, "The battery is dead and the voltmeter is faulty", eliminates all but one of the four and so it has a still higher information content.

This notion of semantic information enables us to distinguish between different sorts of thought process. Given a set of premises and a conclusion, we can ask: what is the relation between the states of affairs that they respectively eliminate? There are clearly five possibilities (corresponding to the five possible relations between two sets):

(1) The premises and conclusion rule out exactly the same states of affairs. This is a case of deduction, as the following example makes clear. You know that the battery is dead or the voltmeter is faulty, or both. By testing the voltmeter, you observe that is not faulty. Your premises are thus:

The battery is dead or the voltmeter is faulty, or both.
The voltmeter is not faulty.

And so you infer:

∴ The voltmeter is not faulty and the battery is dead.

The conclusion follows validly from your premises, i.e. it must be true given that the premises are true. It does not increase semantic information: the premises eliminate all but one possibility:

$d \quad \neg f$

and the conclusion holds in this model too. Like any useful deduction, the conclusion makes explicit what was hitherto only implicit in the premises.

(2) The premises rule out fewer states of affairs than the conclusion, i.e. the conclusion is consistent with additional models. Here is an example. There is a single premise:

The battery is dead.

and the conclusion is:

The battery is dead or the bulb is broken, or both.

The conclusion follows validly from the premise, i.e. it must be true given that the premise is true. Logically-untrained individuals shun such deductions, however, presumably because they throw semantic information away.

(3) The premises and conclusion rule out disjoint states of affairs. This case can only occur when the conclusion contradicts the premises. For example, the premise:

The battery is dead

rules out any model containing:

$\neg d$

whereas the conclusion:

The battery is not dead

rules out any model containing:

d

Hence, the two assertions rule out disjoint states of affairs. A deduction may lead to the negation of a hypothetical assumption, but no rational process of thought leads *immediately* from a premise to its negation (though, cf. Freud, 1925).

(4) The premises and conclusion rule out overlapping states of affairs. For example, the premise:

The battery is dead

leads to the conclusion:

There is a short in the circuit.

The two propositions each rule out any situation in which both are false, namely, any model containing:

$\neg d \quad \neg s$

where "$\neg$ s" denotes "there is not a short in the circuit". Each proposition, however, also rules out independent states of affairs. The premise rules out any situation containing $\neg$ d, and so it rules out the model:

$\neg d \qquad s$

And the conclusion rules out any situation containing $\neg$ s, and so it rules out the model:

$d \quad \neg s$

The production of a conclusion that rules out situations overlapping those ruled out by the premises may be the result of a free association where one proposition leads to another that has no simple relation to it, or it may be the result of a creative thought process.

(5) The conclusion goes beyond the premises to rule out some additional state of affairs over and above what they rule out. This case includes all the traditional instances of induction, and so henceforth I shall use it to define induction: *An induction is any process of thought yielding a conclusion that increases the semantic information in its initial observations or premises.* Here is an example. Your starting point is the premises:

The battery is dead or the voltmeter is faulty, or both.
The voltmeter is faulty.

And you infer:

∴ The battery is not dead.

The conclusion does not follow validly from the premises. They eliminate all but two models:

$d \qquad f$
$\neg d \qquad f$

The conclusion increases information beyond what is in the premises because it eliminates the first of these two models. Yet, the conclusion is quite plausible and it may be true. The difference between induction and deduction is accordingly that induction increases the semantic information in the premises, whereas deduction maintains or reduces it.

An assertion has semantic information because it eliminates certain models *if* it is true. It may not be true, however. And neither deduction nor induction comes with any guarantee that their conclusions are true. If the conclusion you deduced about the battery turns out to be false, then you should revise your belief in one or other of the premises. If the conclusion you induced about the battery turns out to be false, then you should not necessarily change your mind about the truth of the premises. A valid deduction yielding a false conclusion must be based on false premises, but an induction yielding a false conclusion need not be.

Some varieties of induction

Induction occurs in three stages. The first stage is to grasp some propositions—some verbal assertions or perceptual observations. The second stage is to frame a tentative hypothesis that reaches a semantically stronger description or understanding of this information. If this conclusion follows validly from the premises and the background knowledge, then the inference is not an induction but an enthymeme, i.e. a deduction that depends on premises that are not stated explicitly (see Osherson, Smith & Shafir, 1986). The third stage, if a reasoner is prudent, is to evaluate the conclusion, and as a result to maintain, modify, or abandon it.

A common form of induction in daily life concerns a *specific* event, such as the induction made by the master of the Herald of Free Enterprise that the bow doors had been closed. Another form of induction leads to a *general* conclusion. For instance, after standing in line to no avail for just one occasion in Italy, you are likely to infer:

> In Italian bars with cashiers, you pay the cashier first and then take your receipt to the bar to make your order.

A special case of an induction is an *explanation,* though not all explanations are arrived at inductively. In the preceding case, the induction yields a mere description that makes no strong theoretical claim. But, the process may be accompanied by a search for an explanation, e.g.:

> The barmen are too busy to write bills, and so it is more efficient for customers to pay the cashier and then to use their receipts to order.

Scientific laws are general descriptions of phenomena, e.g. Kepler's third law of planetary motion describes the elliptical orbits of the planets. Scientific theories explain these regularities on the basis of more fundamental considerations, e.g. Einstein's theory of gravitation explains planetary orbits in terms of the effects of mass on the curvature of space-time. Some authors argue that induction plays no significant role in scientific thinking. Thus, Popper (1972) claims that science is based on explanatory conjectures that are open to falsification, but he offers no account of their origins. The distinction between an explanation and a corresponding description is far from clear. One view is that the explanation is a statement in a theoretical language that logically implies the description, which is a statement in an observation language. But this claim is disputed (see e.g. Harman, 1973; Thagard, 1988), and it misses the heart of the matter psychologically. You can describe a phenomenon without understanding it, but you cannot explain a phenomenon unless you have some putative understanding of it. Descriptions allow one to make a mental simulation of a phenomenon, whereas explanations allow one to take it to pieces: you may know what causes the phenomenon, what results from it, how to influence, control, initiate, or prevent it, how it relates to other phenomena or how it resembles them, how to predict its onset and course, what its internal or underlying structure is, how to diagnose unusual events, and, in science, how to relate the domain as a whole to others. Scientific explanations characteristically make use of theoretical notions that are unobservable, or that are at a lower physical level than descriptions of the phenomena. An explanation accounts for what you do not understand in terms of what you do understand: you cannot construct a model if the key explanatory concepts are not available to you. Hence, a critical distinction is whether an explanation is developed by deduction (without increasing the semantic information in the premises and background knowledge), by induction (increasing the semantic information), or by creation (with an overlap in the semantic information in the explanation and the original knowledge and premises).

The induction of a generalization could just as well be described as an induction about a concept. In the earlier example, you acquired knowledge about the concept:

Italian bars with cashiers.

These *ad hoc* concepts are clearly put together inductively out of more basic concepts, such as the concepts of cashiers, receipts, and bars (Barsalou, 1987). Adults continue to learn concepts throughout their lives. Some are acquired from knowledge by acquaintance, others from knowledge by description. You cannot acquire the full concept of a color, a wine, or a sculpture without a direct acquaintance with them, but you can learn about quarks, genes, and the unconscious, from descriptions of them.

In summary, inductions are either specific or general; and either descriptive or explanatory. Generalizations include the acquisition of *ad hoc* concepts and the formulation of conjectures to explain sets of observations, even perhaps a set containing just a single datum, such as Sir Alexander Fleming's observation of the destruction of bacteria on a culture plate—an observation that led to the discovery of penicillin. All these results are fallible, but human reasoners are usually aware of the fallibility of their inductions.

Two hypotheses about induction: common elements vs. prototypes

My goal now is to advance a new theory of induction, which accounts for specific and general inductions. To set the scene, however, I want to sketch the main lines of the only two historically important ideas about induction. The first idea is that induction is a search for what is common to a set of observations. Hence, if they all have an element in common, then this element may be critical. If the positive and negative instances of the class of observations differ just in respect of this element, then it is indeed the critical element. This idea implies that a class of events has a set of necessary conditions that are jointly sufficient to determine its instances (see Smith & Medin, 1981). It can be traced back to the British Empiricist philosophers, such as Mill (1843), and it provided the blueprint for a generation of modern psychological investigations. For example, one of the founders of Behaviourism, Clark L. Hull (1920), studied the acquisition of concepts based on common elements, and he extended his results to everyday concepts, arguing that the meaning of dog is "a characteristic more or less common to all dogs and not common to cats, dolls, and teddy-bears".

The second idea rejects common elements (e.g. Wittgenstein, 1953; de Saussure, 1960). Hence, dogs have nothing in common with one another. They tend to have four legs, fur, and the ability to bark, but these are not necessary conditions—a dog could be three-legged, bald, and mute. The criteria for doghood accordingly characterize a prototypical dog. Prototypes led a secret life in psychology (see e.g. Fisher, 1916; Bruner, Goodnow & Austin, 1956) until they emerged in the work of Rosch (e.g. 1973). She argued that real entities are mentally represented by prototypes. This idea was corroborated by the finding that not all instances of a concept are deemed to be equally representative—a terrier is a prototypical dog, but a chihuahua is not. Similarly, the time to make judgements about membership of a concept depends on the distance of the instance from the prototype (see e.g. Rips, Shoben & Smith, 1973; Hampton, 1979).

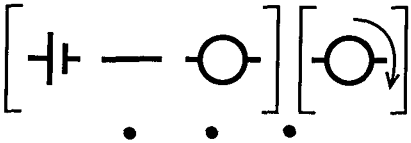
The contrast between the two ideas of induction is striking. The first idea presupposes that a general phenomenon has common elements, and the second idea rejects this presupposition in favor of prototypes. Not surprisingly, current studies of induction are in a state of flux. Students of artificial intelligence have turned the first idea into machines that manipulate explicitly structured symbols in order to produce inductive generalizations (e.g. Hunt, Marin & Stone, 1966; Winston, 1975; Quinlan, 1983; Michalski, 1984; Langley, Simon, Bradshaw & Zytkow, 1987). Connectionists have implemented a version of the second idea (e.g. Hinton, 1986; Hanson & Bauer, 1989). Psychologists have examined both ideas experimentally (see Smith & Medin, 1981). And philosophers have argued that neither idea is viable and that induction is impossible (see e.g. Fodor, 1988, and for a rebuttal, Johnson-Laird, 1983, Ch. 6). The state of the art in induction can be summarized succinctly: theories and computer programs alike represent inductive conjectures in an internal language based on a given set of concepts; they use a variety of linguistic operations for generalizing (and specializing) these conjectures; there are as yet no procedures that can rapidly and invariably converge on the correct inductive description in a language as powerful as the predicate calculus. Certainly, no adequate theory of the human inductive process exists, and this gap is a serious defect in knowledge.

A model theory of specific inductions

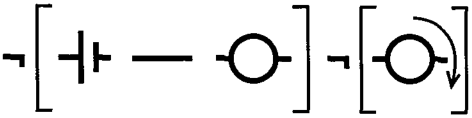
The process of induction, I am going to argue, is the addition of information to a model. In the case of specific inductions in everyday life, the process is hardly separable as a distinct mental activity: it is part of the normal business of making sense of the world. When the starter won't turn over the engine, your immediate thought is:

The battery is flat.

Your conclusion is plausible, but invalid, and so Polya (1957) has suggested that formal, but invalid, rules are the heuristic basis of such inferences. Because rules do not even appear to underlie valid inferences (see Johnston-Laird & Byrne, 1991), it is likely that specific inductions have another basis. You have models, perhaps simplistic, of the car's electrical circuitry including the battery and starter:



The three symbols in the top left-hand brackets denote a model of the battery with power, a model of the circuit conveying power to the starter, and a model of the starter as working. The symbol on the top right hand denotes a model of the starter turning over the engine. The second model consisting in the three dots is initially implicit: It is just a place-holder to allow for the fact that there is an alternative to the first model. When you observe that the starter does *not* turn over the engine, then this observation eliminates the first model and fleshes out the second model to yield:



You can now diagnose that the battery is dead, though there are other possible diagnoses: the circuit is broken, or the starter does not work. The original model might be triggered by anything in working memory that matches its explicit content, and so it can be used to make both deductions and inductions.

People are extraordinarily imaginative in building explanatory models that interrelate specific events. Tony Anderson and I demonstrated their ability in an experiment based on randomly paired events (Johnson-Laird & Anderson, 1991). In one condition the subjects received pairs of sentences taken at random from separate stories:

> John made his way to a shop which sold TV sets.
> Celia had recently had her ears pierced.

In another condition, the sentences were modified to make them co-referential:

> Celia made her way to a shop which sold TV sets.
> She had recently had her ears pierced.

The subjects' task was to explain what was going on. They readily went beyond the information given to them in order to account for what was happening. They proposed, for example, that Celia was getting reception in her ear-rings and wanted the TV shop to investigate, that she was wearing new earrings and wanted to see herself on closed circuit TV, that she had won a bet by having her ears pierced and was going to spend the money on a TV set, and so on. The subjects were almost as equally ingenious with the sentences that were not co-referential.

A critical factor in the construction of a model is, as Tversky and Kahneman (1973) have established, the availability of relevant knowledge. We investigated this aspect of specific inductions in an experiment using such premises as:

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

When we asked the subjects to say what happened, every single one replied that the old man died. But, when the experimenter responded, "Yes, that's possible but not in fact true," then the majority of subjects were able to envisage alternative models in which the old man survived. If the experimenter gave the same response to each of the subjects' subsequent ideas, then sooner or later they ran out of ideas. Yet, they tended to generate ideas in approximately the same order as one another, i.e. the sequences were reliably correlated. Hence, the availability of relevant knowledge has some consistency within the culture. The conclusions to the snake-bite problem, for instance, tend to be produced in the following order:

(1) The old man died.
(2) The poison was successfully removed, e.g. by sucking it out.
(3) The old man was immune to the poison.
(4) The poison was weak, and not deadly.
(5) The poison was blocked from entering the circulatory system, e.g. by the man's thick clothing.

Could the subjects be certain that they had exhausted all possible models of the

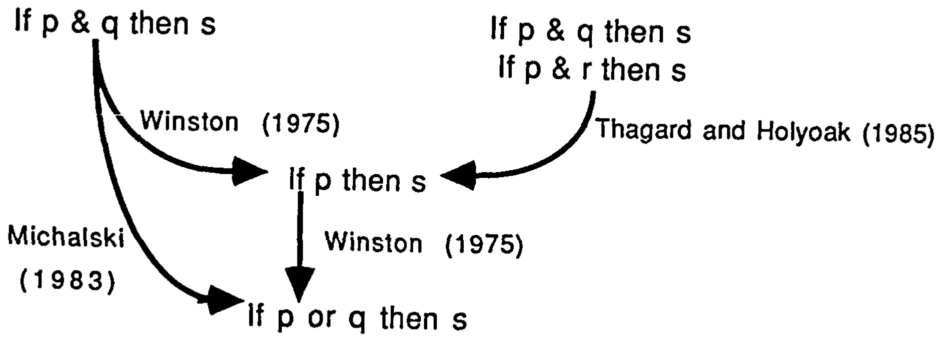


Figure 1. *Some rules of generalization used in inductive programs*

premises? Of course, not. Indeed, by the end of the experiment, their confidence in their initial conclusion had fallen reliably, even in a second group where the experimenter merely responded, "Yes, that's possible" to each idea. Specific inductions crop up so often in everyday life because people rarely have enough information to make valid deductions. Life is seldom *deductively closed.* Specific inductions are potentially unlimited, and so there may always be some other, as yet unforeseen, counterexample to a putative conclusion. A few of the subjects in the experiment produced still more baroque possibilities, such as that the old man was kept alive long enough for someone to invent an antidote. If the sequence of conclusions is to be explained in terms of rules for common sense inferences (see Collins & Michalski, 1989), then they will have to generate a sequence of plausible inferences. However, Bara, Carassa and Geminiani (1984) have shown in a computer simulation that such sequences can be generated by the manipulation of models.

Models and generalization

General inductions, according to linguistic conceptions, depend on going beyond the data in order to make a generalization. A variety of linguistic operations have been proposed to make such generalizations, and Fig. 1 shows just some of them. It is natural to wonder how many different linguistic operations of generalization there are. We can begin to answer this question by considering the following possibilities. A sufficient condition for a concept can be stated in a conditional assertion, e.g.:

If it is a square then it is an instance of the concept.

A necessary condition for a concept can also be stated in a conditional assertion, e.g.:

If it is an instance of the concept then it is a square.

Sufficient conditions for a concept, C, can be generalized by the following operations:

(1) Dropping a conjunct from the antecedent:
If A & B then C becomes *If A then C*
(2) Adding an inclusive disjunction to the antecedent:
If A then C becomes *If A or B then C*

Necessary conditions for a concept, C, can be generalized by the following operations:

(3) Adding a conjunct to the consequent:

If C then A becomes *If C then (A & B)*

(4) Dropping an inclusive disjunct from the consequent:
If C then (A or B) becomes *If C then A*

Both conditionals can be generalized by adding the respective converse so as to state necessary and sufficient conditions:

(5) Adding the converse conditional:
If A then C become *If and only if A then C*
If C then A

These transformations are applicable to inductive hypotheses in general. Hence, for example, the step from:

If something is ice, then it is water.

to:

If something is ice, then it is water and it is frozen.

is a generalization (based on operation 3).

The five operations above by no means exhaust the set of possible generalizations. Consider, for example, the set of all four possible models that can be built from the two propositions: "it's a square", and "it's a positive instance of the concept," and their negations:

□ + ve
□ − ve
¬□ + ve
¬□ − ve

where " + ve" symbolizes "it's a positive instance of the concept", and " − ve" symbolizes "it's a negative instance of the concept". The number of relevant propositions, n, is two in this case, the number of possible models 2^n, and the number of possible subsets of them is $2^{(2^n)}$, including both the set as a whole and the empty set. With any set of models based on n propositions, then a hypothesis such as:

If it's a square, then it's a positive instance of the concept

eliminates a quarter of them. We can now ask how many logically distinct propositions are generalizations of this hypothesis, i.e. how many eliminate the same models plus at least one additional model. The answer equals the number of different sets of models that exclude the same quarter of possible models as the original hypothesis minus two cases: the empty set of models (which corresponds to a self-contradiction) and the set excluded by the original hypothesis itself.

$$2^{(2^n - (0.25)(2^n))} - 2$$

In general, given a hypothesis, H, that rules out a proportion, I(H), of possible models, the number of possible generalizations of H is equal to:

$$2^{(2^n - (I(H))(2^n))} - 2$$

Unless a hypothesis has a very high information content, which rules out a large proportion of models, then the formula shows that the number of its possible generalizations increases exponentially with the number, n, of potentially relevant propositions. Any simple search procedure based on eliminating putative hypotheses will not be computationally tractable: it will be unable to examine all possible generalizations in a reasonable time. Many inductive programs have been designed without taking this problem into account. They are viable only because the domain of generalization has been kept artificially small. The programmer rather than the machine has determined the members of the set of relevant propositions.

Although there are many possible operations of linguistic generalization, the situation is very different if induction is based instead on mental models. Only one operation is needed for the five generalizations above and indeed for all possible generalizations in a Boolean domain. It is the operation of adding information to a model with the effect of eliminating it. For example, to revert to the five operations above, the generalization of dropping a conjunct from an antecedent is equivalent to eliminating a model. Thus, an assertion of the form:

If A & B then C

corresponds to a set of models that includes:

A $\neg$B $\neg$C

When this model is eliminated, the resulting set is equivalent to:

If A then C

and the tautology, B or not-B.

The operation of eliminating a model—by adding information that contradicts it—suffices for any generalization, because generalization is nothing more than the elimination of possible states of affairs. The resulting set of models can then be described by a parsimonious proposition. Although the operation obviates the need to choose among an indefinite number of different forms of linguistic generalization, it does not affect the intractability of the search. The problem now is to determine which models to eliminate. And, as ever, the number of possibilities to be considered increases exponentially with the number of models representing the initial situation.

Models and the operations of generalization with quantifiers

Some inductive programs, such as INDUCE 1.2 (Michalski, 1983), operate in a domain that allows quantification over individuals, i.e. with a version of the predicate calculus. Where quantifiers range over infinitely many individuals, it is impossible to calculate semantic information on the basis of cardinalities, but it is still possible to maintain a partial rank order of generalization: one assertion is a generalization of another if it eliminates certain states of affairs over and above those eliminated by the other assertion. Once again, we can ask: how many operations of generalization are necessary?

The answer, once again, is that the only operation that we need is the archetypal one that adds information to models so as to eliminate otherwise possible states of affairs. Tokens can be added to the model in order to generalize the step from a finite number of observations to a universal claim. You observe that some entities of a particular sort have a property in common:

Electrons emitted in radioactive decay damage the body.
Positrons emitted in radioactive decay damage the body.
Photons emitted in radioactive decay damage the body.

These initial observations support the following model:

```
p    d
p    d
p    d
  . . .
```

where "p" symbolizes a particle and "d" damage to the body. Information can be added to the model to indicate that all such particles have the same property:

```
[p]    d
[p]    d
[p]    d
  . . .
```

where the square brackets indicate that the set of particles is now exhaustively represented in the model. This model rules out the possibility of any particles emitted in radioactive decay that do not damage the body. This operation on models corresponds to a linguistic operation that leads from an initial observation:

Some particles emitted in radioactive decay damage the body.

to the conclusion:

Any particles emitted in radioactive decay damage the body.

Some authors refer to this operation as "instance-based" generalization (Thagard and Holyoak, 1985) or as "turning constants into variables" (Michalski, 1983, p. 107). There seems to be little to choose between the operation on models and the linguistic operation. However, the operation on models turns out to yield other forms of linguistic generalization.

Information can be added to a model to represent a new property of existing individuals. If you have established that certain individuals have one property, then you can make a generalization that they satisfy another. You observe, for example, bees with an unusually potent sting:

```
[s]
[s]
[s]
. . .
```

where "s" denotes a bee with a potent sting, and the set is exhaustively represented. You conjecture that the cause of the sting is a certain mutation:

```
[m]  [s]
[m]  [s]
[m]  [s]
  . . .
```

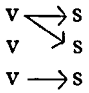
Likewise, new relations can be added to hold between existing entities in a model. For example, a model might represent the relations among, say, a finite set of viruses and a finite set of symptoms. The semantically weakest case is as follows:

```
v    s
v    s
v ——→ s
```

where there is one definite causal relation, signified by the arrow, but nothing is known about the relations, positive or negative, between the remaining pairwise combinations of viruses and symptoms. You can describe this model in the following terms:

At least one virus causes at least one of the symptoms.

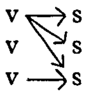
By the addition of further causal relations the model may be transformed into the following one:



You can describe a model of this sort as follows:

Each of the symptoms is caused by at least one of the viruses.

Hence, the effect is still equivalent to the linguistic operation of replacing an existential quantifier ("at least one") in the previous description by a universal quantifier ("each"). The addition of a further causal link, however, yields a still stronger model.



You can describe a model of this sort in the following terms:

At least one of the viruses causes each of the diseases.

In the predicate calculus, the linguistic effect of the operator is now to promote an existential quantifier from inside to outside the scope of a universal quantifier:

$$\forall s \; \exists v \; v \text{ causes } s \Rightarrow \exists v \; \forall s \; v \text{ causes } s$$

No such rule, however, appears to be have been proposed by any current inductive theory. The model theory has therefore led us to the discovery of a new form of linguistic generalization.

The operation of adding information to models enables us to generalize from the weakest possible model to the strongest possible one in which each of the viruses causes each of the symptoms. Hence, the addition of information to models suffices for all possible generalizations in those everyday domains that can be described by the predicate calculus. It replaces the need for a battery of various linguistic operations.

How can induction be constrained?

The burden of the argument so far is simple: induction is a search for a model that is consistent with observation and background knowledge. Generalization calls for only one operation, but the search is intractable because of the impossibility of examining all of its possible effects. The way to cut the problem down to a tractable size is, not to search blindly by trial and error, but to use constraints to guide the search (Newell & Simon, 1972). Three constraints can be used in any domain and may be built into the inductive mechanism itself: specificity, availability, and parsimony.

Specificity is a powerful constraint on induction. It is always helpful to frame the most specific hypothesis consistent with observation and background knowledge, that is, the hypothesis that admits the fewest possible instances of a concept.[1] This constraint is essential when you can observe only positive instances of a concept. For example, if you encounter a patient infected with a new virus, and this individual has a fever, a sore throat, and a rash, then the most specific hypothesis about the signs of the disease is:

$$\text{fever} \cap \text{sore throat} \cap \text{rash}$$

where $\cap$ denotes the intersection of sets. If you now encounter another patient with the same viral infection, who has a fever and a rash, but no sore throat, you will realize that your initial hypothesis was too specific. You can generalize it to one consistent with the evidence:

$$\text{fever} \cap \text{rash}$$

Suppose, however, that you had started off with the more general inclusive disjunction:

$$\text{fever} \cup \text{sore throat} \cup \text{rash}$$

where $\cup$ denotes the union of sets. Although this conjecture is consistent with the data, it is too general, and so it would remain unaffected by your encounter with the second patient. If the only genuine sign of the disease is the rash, then you would never discover it from positive examples alone, because your hypothesis would accommodate all of them. Hence, when you are trying to induce a concept from positive instances, you must follow the specificity constraint. Your hypothesis may admit too few instances, but if so, sooner or later, you will encounter a positive instance that will allow you to correct it.

This principle has been proposed by Berwick (1986) in terms of what he calls the "subset" principle, which he derives from a theorem in formal learning theory due to Angluin (1978). In elucidating children's acquisition of syntax, phonology, and concepts—domains in which they are likely to encounter primarily positive instances—Berwick argues that the instances that are described by a current inductive hypothesis should be as few as possible. If they are a proper subset of the actual set of instances, then children can correct their inductive hypothesis from encounters with further positive instances. But, if the current hypothesis embraces all the members of the actual set and more, then it will be impossible for positive instances to refute the hypothesis. What Angluin (1978) proved was that positive instances could be used to identify a language in the limit, i.e. converge upon its grammar without the need for subsequent modifications (see Gold, 1967), provided that the candidate hypotheses about the grammar could be ordered so that each progressively more general hypothesis includes items that are not included in its predecessor. The inductive system can then start with the most specific hypothesis, and it will move to a more general one whenever it encounters a positive instance that falls outside its current hypothesis.

Availability is another general constraint on induction. It arises from the machinery that underlies the retrieval of pertinent knowledge. Some information comes to mind more readily than other information, as we saw in the case of the specific induction about the snake bite. The availability of information, as Tversky and Kahneman (1973) have shown, can bias judgement of the likelihood of an event. It also underlies the "mutability" of an event—the ease with which one can envisage a counterfactual scenario in which the event does *not* occur (see Tversky & Kahneman, 1982; Kahneman & Miller, 1986). Availability is a form of bias, but bias is what is needed to deal with the intractable nature of induction.

Parsimony is a matter of fewer concepts in fewer combinations. It can be defined only with respect to a given set of concepts and a system in which to combine them. Hence, it is easily defined for the propositional calculus, and there are programs guaranteed in principle to deliver maximally parsimonious descriptions of models within this domain (see Johnson-Laird, 1990). What complicates parsimony is that the presumption of a conceptual system begs the question. There is unlikely to be any procedure for determining absolute parsimony. Its role in induction therefore seems to be limited to comparisons among alternative theories using the same concepts.

The most important constraint on induction I have left until last for reasons that will become clear. It is the use of existing knowledge. A rich theory of the domain will cut down the number of possible inductions; it may also allow an individual to generalize on the strength of only a single instance. This idea underlies so-called "explanation-based learning" in which a program uses its background knowledge of the domain to deduce why a particular instance is a member of a concept (see e.g. DeJong & Mooney, 1986; Mitchell, Keller & Kedar-Cabelli, 1986). Another source of knowledge is a helpful teacher. A teacher who cannot describe a concept may still be able to arrange for a suitable sequence of instances to be presented to pupils. This pedagogical technique cuts down the search space and enables limited inductive mechanisms to acquire concepts (Winston, 1975). The constraints of theory are so important that they often override the pure inductive process: one ignores counterexamples to the theory. The German scientist and aphorist, Georg Lichtenberg (1742–1799), remarked: "One should not take note of contradictory experiences until there are enough of them to make constructing a new system worthwhile". The molecular biologist James Watson has similarly observed that no *good* model ever accounts for all the facts because some data are bound to be misleading if not plain wrong (cited in Crick, 1988, p. 60). This methodological prescription appears to be observed automatically by young children seeking to acquire knowledge. Karmiloff-Smith and Inhelder (1974/5) have observed that children learning how to balance beams ignore counterexamples to their current hypotheses. Such neglect of evidence implies that induction plays only a limited role in the development of explanations. An explanation does not increase the semantic information in the observations, but rather eliminates possibilities that only overlap with those that the evidence eliminates. According to my earlier analysis, the process is therefore not inductive, but creative.

The design of the human inductive system

Although studies in psychology and artificial intelligence have been revealing, no-one has described a feasible program for human induction. What I want to consider finally are some of the design characteristics that any plausible theory must embody. If nothing else, these characteristics show why no existing algorithm is adequate for human induction. The agenda is set by the theory of mental models and the underlying subconcepts from which they are constructed. This theory implies that there are three sources of concepts.

The first source is evolution. What must be genetically endowed for induction to be possible are the following basic components:

(1) A set of subconcepts. These subconcepts include those for entities, properties, and relations, that apply to the perceptual world, to bodily states and emotions, and to mental domains including deontic and epistemic states (facts, possibilities, counterfactual states, impossibilities). They are the ultimate components out of which all induc-

TABLE 1. Some examples illustrating the ontological and epistemological classification of concepts

	The epistemological dimension		
The ontological dimension	Analytical concepts	Natural kinds	Artefacts
Entities			
Objects:	Triangle	Dog	Table
Substances:	Space	Water	Food
Properties	Straight	Alive	Expensive
Relations	Causes	Sees	Owns

tions are constructed, and they are used in the construction and manipulation of mental models. It is of little use to define one concept, such as woman, in terms of other high-level concepts, such as adult, human, female. The concept must depend on subconcepts that can be used to construct models of the world. What is needed is an analysis of the satisfaction conditions of our most basic ideas and their interrelations. These conditions enable us to envisage the world and in certain circumstances to verify our imaginings.

(2) A set of methods for combining concepts. These methods include composition, i.e. a method that allows one subconcept to call upon another, and recursion, i.e. a method that allows a set of subconcepts to be used in a loop that is executed for a certain number of times (see Boolos & Jeffrey, 1989). These combinations interrelate subconcepts to form concepts, and they interrelate concepts to form new high-level concepts or inductive conjectures.

(3) A set of inductive mechanisms. It is these mechanisms that make possible induction of concepts and generalizations.

The second source of concepts is *knowledge by compilation.* The process depends on an inductive mechanism that assembles concepts (and their taxonomic interrelations) out of the set of innate subconcepts and combinations. Verbal instruction alone is no use here: there is no substitute for the construction of models of the world—its entities, properties, and relations. Ultimately, the repeated construction of models, as I suggested in the case of spatial relations, such as *diagonally up and to the right,* enables the relevant concept to be compiled into subconcepts.

Concepts constructed from subconcepts are heterogeneous: some are *analytic* in that they have necessary and sufficient conditions; others such as *natural kinds* and *artefacts* are open-ended, prototypical, and depend on default values. Table 1 gives examples of these three main sorts of concepts for entities, properties, and relations. This heterogeneity has consequences for the mechanism that constructs new concepts. Those concepts with necessary and sufficient conditions might be induced by a variant of the method that builds decision trees (Quinlan, 1983), but this method runs into difficulties with concepts that depend on prototypes. They might be acquired by a program that constructs hierarchies of clusterings in which instances are grouped together in ways that are not all or none (e.g. Pearl, 1986; Fisher, 1987; Gennari, Langley & Fisher, 1990)

Neither decision-tree nor connectionist programs correspond precisely to the required inductive mechanism. Its input, as I have argued, is a sequence of models, and its output is a heterogeneous set of concepts. This heterogeneity suggests that the mechanism builds up a hierarchy of defaults, exceptions, and necessary conditions. The mechanism must also be able to acquire concepts of objects, properties, relations, and quantification. Although there are proposals that are syntactically powerful enough to cope with these demands, no satisfactory program for inducing the corresponding concepts yet exists.

The third source of both concepts and conjectures is *knowledge by composition*. The process depends on a mechanism that comes into play only after some high-level concepts have been assembled out of subconcepts. Its results take the form of inductive generalizations and *ad hoc* concepts, which it composes out of existing high-level concepts. For example, you can acquire the *ad hoc* concept "araeostyle" from the following definition:

> An *araeostyle* building has equi-distant columns with a distance between them of at least four times the diameter of the columns.

Many machine programs are static in that they construct new concepts out of a fixed basic set that depends on the user's specification of the inductive problem (see e.g. Hunt, Marin & Stone, 1966; Mitchell, 1977). The human inductive mechanism, however, is evolutionary. It constructs new concepts from those that it has previously constructed; and, unlike most existing programs, it can construct novel concepts in order to frame an inductive hypothesis. It can also deploy compositional principles in the construction of concepts such as araeostyle. Existing programs can induce some compositional concepts using explicitly structured representations (see Anderson, 1975; Power & Longuet-Higgins, 1978; Selfridge, 1986), but they cannot yet induce concepts that depend on recursion. Although connectionist systems can acquire rudimentary conceptual systems, and also rudimentary syntactic rules (see e.g. Hanson & Kegl, 1987), they are not presently able to learn sufficiently powerful concepts to emulate human competence.

Knowledge by composition, as in the case of araeostyle, saves much time and trouble, but it is superficial. The shift from novice to expert in any conceptual domain appears to depend on knowledge by compilation. Only then can a concept be immediately used to construct models or to check that the concept is satisfied in a perceptual model. The induction of generalizations similarly depends on the use of experience to add information to models of the relevant domain.

The case for models

Theorists tend to think of induction as yielding linguistic generalizations (see e.g. the contributions in Kodratoff & Michalski, 1990). Hence, a major question has been to find the right language in which to represent concepts and conjectures. There has been much debate amongst the proponents of different mental languages, such as semantic networks, production systems, and versions of the predicate calculus. Yet, as I have argued, to think of the results of induction as linguistic representations may be a vast mistake. It may not do justice to human thinking. The purpose of induction is to make sense of the world, either by enabling individuals to predict or to categorize more efficiently or, better still, to understand phenomena. The mind builds models, and the structure of models is the basis of human conceptions of the structure of the world. The products of induction may therefore be models, either ones that simulate phenomena (descriptive inductions) or else ones constructed from more elemental subconcepts (explanatory inductions). After such models have been induced, they can, if necessary, be used to formulate verbal generalizations.

One advantage of models is that the inductive mechanism needs, in principle, only one operation of generalization: the addition of information to models. This operation is equivalent to quite diverse effects on linguistic hypotheses. When it leads to the elimination of a model, it is equivalent to adding the negation of the description of that model to the current verbal hypothesis. It can have the effect of a so-called universal generalization, which introduces a universal quantifier in place of an existential. And it can have the effect of promoting an existential quantifier from inside to outside the scope of a universal quantifier.

Another advantage of models is that they embody knowledge in a way that naturally constrains inductive search. They maintain semantic information, they ensure internal consistency, and they are parsimonious because each entity is represented only once. They can also focus attention on the critical parts of the phenomena. An instructive example is provided by Novak's (1977) program for solving textbook problems in physics. It initially represents problems in a semantic network, but this representation contains too much information, and so the program extracts from it a model of the situation that is used to identify the points where forces have to balance.

One final advantage of models is that they elucidate the clues about induction that have emerged from the psychological laboratory. Because of the limited processing capacity of working memory, models represent only certain information explicitly and the rest implicitly. One consequence is that people fall into error, and the evidence shows they make the same sorts of error in both deduction and induction. Thus, in deduction, they concentrate on what is explicit in their models, and so, for example, they often fail to make certain deductions. In induction, they likewise focus on what is explicit in their models, and so seldom seek anything other than evidence that might corroborate their inductive conjectures. They eschew negative instances, and encounter them only when they arise indirectly as a result of following up alternative hypotheses (see e.g. Bruner *et al.*, 1956; Wason, 1960; Klayman & Ha, 1987). In deduction, they are markedly influenced by the way in which a problem is framed: what a model represents explicitly depends on what is explicitly asserted, and so individuals often have difficulty in grasping that two assertions have the same truth conditions, e.g. "Only the bakers are athletes" and "All the athletes are bakers" (see Johnson-Laird & Bryne, 1991). In induction, there are equally marked effects of how a problem is framed (see Hogarth, 1982). In both deduction and induction, disjunctive alternatives cause difficulties. They call for more than one model to be constructed, whereas reasoners are much better able to cope with a single model and thus have a natural preference to work with conjunctions. Disjunctive informative even appears to block straightforward decisions. For example, many people who choose a vacation if they pass an exam, or if they fail it, do not choose it when the outcome of the exam is unknown (see Shafir & Tversky, 1991; Tversky & Shafir, 1991). The very preference of a "common element" analysis of concepts is just another manifestation of the same phenomenon. Similarly, a single model of a *process* underlies the natural tendency to overgeneralize. Once children learn, for example, how to form the regular past tense, their tendency to generate "go-ed" supplants their previous grasp of "went" as a separate lexical item. Finally, knowledge appears to play exactly the same part in both deduction and induction. It biases the process to yield more credible conclusions.

Conclusions

The principal contrast in induction is between specific and general inductions. Specific inductions are part of comprehension: you flesh out your model of a discourse or the world with additional information that is automatically provided by your general knowledge. General inductions yield new models, which can also enrich your conceptual repertoire. The human inductive mechanism that carries out these tasks appears to embody five design characteristics:

(1) Its ultimate constituents are a set of subconcepts and conceptual combinations that are powerful enough to construct any mental model.

(2) It can induce heterogeneous concepts of objects, properties, and relations from knowledge by acquaintance. The repeated construction of models leads to the compilation of concepts into subconcepts, including necessary subconcepts and those that specify default values.

(3) It can construct novel concepts from knowledge by composition, assembling them according to principles that generate recursively embedded structures.

(4) It can construct a model of a domain either *ab initio* or by adding information to an existing set of models in accordance with evidence.

(5) It is guided by constraints. It takes into account available knowledge; it formulates the most specific generalizations consistent with the data and background knowledge; and, perhaps, it seeks the simplest possible conjecture consistent with the evidence.

The price of tractable induction is imperfection. We often concentrate on the triumphs of induction and the minor imperfections that yield clues to the nature of its mechanism. We overlook its catastrophes—the fads of pseudo-science, the superstitions of daily life, and the disastrous generalizations that lead to such events as the sinking of the Herald of Free Enterprise. The origin of these errors is in the human inductive mechanism: its heuristics are what students of other cultures refer to as "magical thinking". And the pressure of working memory capacity often puts too much emphasis on what is explicitly represented in a model. Theorists, this theory argues, are bound to focus on what is explicitly represented in their models. The reader is invited to reflect on the recursive consequences of this claim for the present theory.

Acknowledgements

I am grateful to friends and colleagues who over the years have worked with me on the theory of mental models: Tony Anderson, Bruno Bara, Monica Bucciarelli, Alan Garnham, Vittorio Girotto, Mark Keane, Maria Sonino Legrenzi, Paolo Legrenzi, Jane Oakhill, Mick Power, Patrizia Tabossi, Walter Schaeken. I am especially grateful to Ruth M. J. Byrne who helped in the formulation of the theory as it applies to deduction (see Johnson-Laird & Byrne, 1991). Part of the material in this paper is based on part of one of three MacEachran Memorial Lectures, which I delivered in October, 1990, in the Department of Psychology at the University of Alberta, Edmonton (see Johnson-Laird, 1992). Finally, my thanks to the editors of RISESST for soliciting this article.

Note

[1] The most specific description of a set is one that admits the fewest possible instances. The most specific proposition about a phenomenon is the one that rules out as false the fewest possible states of affairs. The difference reflects the difference between sets (which have conditions of satisfaction) and propositions (which are true or false).

References

ANDERSON, J.R. (1975) Computer simulation of a language acquisition system—a first report. In SOLSO, R.L. (Ed.) *Information Processing and Cognition* (Hillsdale, NJ, Lawrence Erlbaum Associates).

ANGLUIN, D. (1978) Inductive inference of formal languages from positive data. *Information and Control*, 45, pp. 117–135.

BARA, B.G., CARASSA, A.G. & GEMINIANI, G.C. (1984) Inference processes in everyday reasoning. In PLANDER, D. (Ed.) *Artificial Intelligence and Information-Control Systems of Robots* (Amsterdam, Elsevier).

BAR-HILLEL, Y. & CARNAP, R. (1964) An outline of a theory of semantic information. In BAR-HILLEL, Y. *Language and Information* (Reading, MA, Addison-Wesley).

BARSALOU, L.W. (1987) The instability of graded structure: implications for the nature of concepts. In NEISSER, U. (Ed.) *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization* (Cambridge, Cambridge University Press).

BERWICK, R.C. (1986) Learning from positive-only examples: The Subset principle and three case studies. In MICHALSKI, R.S., CARBONELL, J.G. & MITCHELL, T.M. (eds) *Machine Learning: An Artificial Intelligence Approach, Vol. II* (Los Altos, CA, Morgan Kaufmann).

BOOLOS, G. & JEFFREY R. (1989) *Computability and Logic.* Third Edition. (Cambridge, Cambridge University Press).

BRUNER, J.S., GOODNOW, J.J. & AUSTIN, G.A. (1956) *A Study of Thinking* (New York, Wiley).

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

CRICK, F. (1988) *What Mad Pursuit* (New York, Basic Books).

DEJONG, G.F. & MOONEY, R. (1986) Explanation-based learning: An alternative view. *Machine Learning*, 1, pp. 145–176.

DE SAUSSURE, F. (1960) *Course in General Linguistics* (London, Peter Owen).

FISHER, D. (1987) Knowledge acquistiion via incremental conceptual clustering. *Machine Learning*, 2, pp. 139–172.

FISHER, S.C. (1916) The process of generalizing abstraction; and its product, the general concept. *Psychological Monographs*, 21, No. 2, whole number 90.

FODOR, J.A. (1980) Fixation of belief and concept acquisition. In PIATTELLI-PALMARINI, M. (Ed.), *Language and Learning: The Debate between Jean Piaget and Noam Chomsky* (Cambridge, MA, Harvard University Press).

FREUD, S. (1925) Negation. *Complete Psychological Works of Sigmund Freud, Vol. 19: The Ego and the Id and Other Works.* (trans J. Strachey) (London, Hogarth).

GARNHAM, A. (1987) *Mental Models as Representations of Discourse and Text* (Chichester, Ellis Horwood).

GENNARI, J.H., LANGLEY, P. & FISHER, D. (1990) Models of incremental concept formation. In CARBONELL, J.G. (Ed.), *Machine Learning: Paradigms and Methods.* (Cambridge, MA, MIT Press). Reprinted from *Artificial Intelligence*, 40, 1989.

GOLD, E.M. (1967) Language identification in the limit. *Information and Control*, 16, pp. 447–474.

HAMPTON, J.A. (1979) Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 18, pp. 441–461.

HANSSON, S.J. & BAUER, M. (1989) Conceptual clustering, categorization, and polymorphy. *Machine Learning*, 3, pp. 343–372.

HANSON, S.J. & KEGL, J. (1987) PARSNIP: A connectionist network that learns natural language grammar from exposure to natural language sentences. *Proceedings of the Ninth Annual Conference of the Cognitive Science Society*, Seattle, pp. 106–119 (Hillsdale, NJ, Lawrence Erlbaum Associates).

HARMAN, G. (1973) *Thought* (Princeton, NJ, Princeton University Press).

HINTON, G.E. (1986) Learning distributed representations of concepts. In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society* (Hillsdale, NJ, Lawrence Erlbaum Associates).

HOGARTH, R. (Ed.) (1982) *New Directions for Methodology of Social and Behavioral Science, No. 11: Question Framing and Response Consistency* (San Francisco, Jossey-Bass).

HULL, C.L. (1920) Quantitative aspects of the evolution of concepts. *Psychological Monographs*, 28, whole number 123.

HUNT, E.B., MARIN, J. & STONE, P.T. (1966) *Experiments in Induction* (New York, Academic Press).

JOHNSON-LAIRD, P.N. (1983) *Mental Models*. (Cambridge, MA, Harvard University Press).

JOHNSON-LAIRD, P.N. (1990) *Propositional reasoning: an algorithm for deriving parsimonious conclusions*. Unpublished MS, Department of Psychology, Princeton University.

JOHNSON-LAIRD, P.N. (1992) *Human and Machine Thinking* (Hillsdale, NJ, Lawrence Erlbaum Associates).

JOHNSON-LAIRD, P.N. & ANDERSON, T. (1989) Common-sense inference. Unpublished MS, Princeton University, New Jersey.

JOHNSON-LAIRD, P.N. & BYRNE, R.M.J. (1991) *Deduction* (Hillsdale, NJ, Lawrence Erlbaum Associates).

KAHNEMAN, D. & MILLER, D. (1986) Norm theory: Comparing reality to its alternatives. *Psychological Review*, 93, pp. 136–153.

KARMILOFF-SMITH, A. & INHELDER, B. (1974/5) 'If you want to get ahead, get a theory'. *Cognition*, 3, pp. 195–212.

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

KODRATOFF, Y. & MICHALSKI, R.S. (1990) *Machine Learning: An Artificial Intelligence Approach. Vol. III* (San Mateo, CA, Morgan Kaufmann).

LANGLEY, P., SIMON, H.A., BRADSHAW, G.L. & ZYTKOW, J.M. (1987) *Scientific Discovery* (Cambridge, MA, MIT Press).

MICHALSKI, R.S. (1983) A theory and methodology of inductive learning. In MICHALSKI, R.S., CARBONELL, J.G. & MITCHELL, T.M. (eds) *Machine Learning: An Artificial Intelligence Approach* (Los Altos, CA, Morgan Kaufmann).

MILL, J.S. (1843/1950) *A System of Logic, Ratiocinative and Inductive* (Toronto, University of Toronto Press).

MITCHELL, T.M. (1977) Version spaces: A candidate elimination approach to rule learning. *Fifth International Joint Conference on Artificial Intelligence* (Cambridge, MA), pp. 305–310.

MITCHELL, T.M., KELLER, R. & KEDAR-CABELLI, S. (1986) Explanation-based generalization: A unifying view. *Machine Learning*, 1, pp. 47–80.

NEWELL, A. (1990) *Unified Theories of Cognition* (Cambridge, MA, Harvard University Press).

NEWELL, A. & SIMON, H.A. (1972) *Human Problem Solving* (Englewood Cliffs, Prentice-Hall).

NOVAK, G.S. (1977) Representations of knowledge in a program for solving physics problems. *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, pp. 286–291.

OSHERSON, D.N., SMITH, E.E. & SHAFIR, E. (1986) Some origins of belief. *Cognition*, 24, pp. 197–224.

PEARL, J. (1986) Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, 29, pp. 241–288.

POLYA, G. (1957) *How to Solve It*. Second edition (New York, Doubleday).

POPPER, K.R. (1972) *Objective Knowledge* (Oxford, Clarendon).

POWER, R.J.D. & LONGUET-HIGGINS, H.C. (1978) Learning to count: A computational model of language acquisition. *Proceedings of the Royal Society (London) B*, 200, pp. 391–417.

QUINLAN, R. (1983) Learning efficient classification procedures and their application to chess end games. In MICHALSKI, R.S., CARBONELL, J.G. & MITCHELL, T.M. (eds), *Machine Learning: An Artificial Intelligence Approach* (Los Altos, CA, Morgan Kaufmann).

RIPS, L.J., SHOBEN, E.J. & SMITH, E.E. (1973) Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior*, 12, pp. 1–20.

ROSCH, E. (1973) Natural categories. *Cognitive Psychology*, 4, pp. 328–350.

SELFRIDGE, M. (1986) A computer model of child language learning. *Artificial Intelligence*, 29, pp. 171–216.

SHAFIR, R. & TVERSKY, A. (1991) *Thinking through uncertainty: Nonconsequential reasoning and choice*. Unpublished MS, Department of Psychology, Princeton University.

SMITH, E.E. & MEDIN, D.L. (1981) *Categories and Concepts* (Cambridge, MA, Harvard University Press).

THAGARD, P. (1988) *Computational Philosophy of Science* (Cambridge, MA, Bradford Books, MIT Press).

THAGARD, P. & HOLYOAK, K.J. (1985) Discovering the wave theory of sound. *Proceedings of the Ninth International Joint Conference on Artificial Intelligence* (Los Altos, CA, Morgan Kaufmann).

TVERSKY, A. & KAHNEMAN, D. (1973) Availability: a heuristic for judging frequency and probability. *Cognitive Psychology*, 4, pp. 207–232.

TVERSKY, A. & KAHNEMAN, D. (1982) The simulation heuristic. In KAHNEMAN, D., SLOVIC P. & TVERSKY, A. (eds), *Judgement under Uncertainty: Heuristics and Biases* (Cambridge, Cambridge University Press).

TVERSKY, A. & SHAFIR, E. (1991) *The disjunction effect in choice under uncertainty*. Unpublished MS, Department of Psychology, Stanford University.

WASON, P.C. (1960) On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12, pp. 129–140.

WINSTON, P.H. (1975) Learning structural descriptions from examples. In WINSTON, P.H. (Ed.), *The Psychology of Computer Vision* (New York, McGraw-Hill).

WITTGENSTEIN, L. (1953) *Philosophical Investigations* (New York, Macmillan).