2 Models of Epistemic States

2.1 Program

A cornerstone concept in an epistemological theory is that of an epistemic state or state of belief. In this chapter I present a number of *models* of epistemic states that have been used within different areas of philosophy, computer science, psychology, and linguistics. I do not aim at presenting a complete list of models; I confine myself to the models that allow me to say something relevant about their *dynamics*, that is, how the states change as a result of epistemic inputs. This chapter, however, is restricted to a presentation of the *static* features of some models of epistemic states as well as the rationality criteria that are used to motivate them. The dynamics of different kinds of models are treated in chapters 3–6.

2.2 Belief Sets

A simple way of modeling the epistemic state of an individual is to represent it by a *set* of sentences. The intended interpretation of such a set is that it consists of exactly those sentences that the individual *accepts* in the modeled state of belief. Instead of saying that the individual accepts the sentence, we can also say that the individual *believes it to be true* or *regards it as certain*.

This kind of model is *linguistic* in the sense that it presupposes an object language **L** from which the sentences in the set are to be selected. For the most part I leave the details of **L** open, assuming only that **L** contains expressions for the standard sentential connectives. These connectives are symbolized as follows:

negation: $-$
conjunction: $\&$
disjunction: $\vee$
material implication: $\rightarrow$

As sentential variables I use A, B, C, ... It is also convenient to introduce symbols for two sentential constants:

truth: $\top$
falsity: $\bot$

Note that these sentences are used only as "ideal points"; they imply nothing about the correspondence between a state of belief and the external world, and they do not represent truth values.

In later chapters the same formulas will be used to denote *propositions*, that is, the "content" of sentences. However, it will be clear from the context what the symbols denote. And, because it is assumed that logically equivalent sentences cannot be distinguished within an epistemic state, there is little risk of equivocation.

Modeling epistemic states by sets of sentences is connected with a typology of epistemic attitudes that for any sentence A allows only the following three possibilities:

A is *accepted*,
A is *rejected*,
A is *indetermined*, that is, A is neither accepted nor rejected.

This typology can be reduced to two items by requiring A to be rejected iff $-A$ is accepted (compare this with section 2.3).

Not every set of sentences can be used to represent *rational* epistemic states. Sets of sentences that may be rationally held by an individual are called *belief sets*. In order to determine which sets of sentences constitute belief sets, I focus on two rationality criteria:

(2.1) The set of accepted sentences should be *consistent*.

(2.2) *Logical consequences* of what is accepted should also be accepted.

Both these criteria presuppose a *logic* for the language **L** in order to determine what is meant by "consistent" and "logical consequence." Both criteria are supported by the view of states of belief as idealized equilibria. But the criteria can also be motivated by more pragmatic considerations—inconsistent sets of belief are of no help when seeking guidance for how to act, and, in order to use one's knowledge effectively, one must be able to draw the consequences of the information one has on a topic.[1] It is clear, however, that the two criteria are not realistic as descriptions of individuals' actual sets of belief. In particular, this is the case for the requirement of including logical consequences—because of the limitations of our mental powers, we often do not see all the consequences of what we accept. I still believe, however, that the criterion is useful, at least as an ideal of rationality.

Because the concept of acceptance is central in the account of epistemic states and epistemic changes developed here, I want to make some remarks on my use of the concept. First, acceptance is a broader concept than belief; it also includes such attitudes as assuming, presupposing, and positing. To accept a proposition is to *treat* it as true in one way or another.[2] For example, in debates one often hypothetically assumes that one does not believe the proposition under discussion in order not to beg the question. In connection with an analysis of conditional sentences in chapter 7, assumed acceptance plays a central role. This position entails that acceptance is relative to context: A person may accept something in one context but reject it or leave it indetermined in another.

Second, one must distinguish the acceptance of a sentence from the *awareness* of this acceptance.[3] Because of limited calculation ability and memory failure, humans are not aware of all the consequences of their beliefs. Harman (1986, pp. 12–14) makes a distinction between *explicit* belief, which means that one's belief involves an explicit mental representation whose content is the content of the belief, and *implicit* belief, which is a belief that is derivable from the explicit beliefs. Because epistemic states are conceived of as rational equilibrium states, the beliefs that are accepted in these states include all implicit beliefs.[4]

Finally, accepting a proposition A in an epistemic state K entails *full belief* in the sense that in K there is no doubt that A is false (Levi expresses this by saying that the negation of A is not a serious possibility in K). This means that I take all accepted propositions to have maximal probability, that is, probability 1. The consequences of this position are elaborated in section 2.7 and in chapter 5. In contrast to this, some authors (for example, de Finetti and Carnap) allow that a proposition with maximal probability need not be accepted as true or, equivalently, that some propositions with zero probability are not inconsistent with the accepted beliefs.[5] However, even if on my account acceptance entails full belief, acceptance does not entail *infallibility* in the sense that an accepted belief will never be doubted. Because an agent is always influenced by different epistemic inputs, sometimes these inputs will contradict or in other ways undermine the accepted beliefs. Then, as Peirce puts it: "The scientific spirit requires a man to be at all times ready to dump his whole cartload of beliefs, the moment experience is against them" (Peirce 1932, pp. 46–47).[6]

After this digression on the notion of acceptance, I now return to the description of belief sets. The rationality criteria (2.1) and (2.2) can be seen as minimal requirements on representations of epistemic states. Belief sets are here defined for weakly specified languages. If more is known or assumed about the structure of the object language, it may be possible to formulate other, more specific rationality criteria.

I assume that the language **L** is governed by a logic that is identified with its consequence relation $\vdash$; that is, a sentence A is logically valid iff it is a consequence of the empty set. The relation $\vdash$ is assumed to satisfy the following conditions:

($\vdash$1) If A is a truth-functional tautology, then $\vdash A$.

($\vdash$2) Modus ponens. That is, if $\vdash A \rightarrow B$ and $\vdash A$, then $\vdash B$.

($\vdash$3) Not $\vdash \bot$. That is, $\vdash$ is consistent.

It follows that $\vdash$ contains classical propositional logic. Furthermore it is assumed that $\vdash$ satisfies the deduction theorem that is, that $\vdash A \rightarrow B$ iff $A \vdash B$), and that it is compact; in other words, if A is a logical consequence of some set X, then A is a consequence of some finite subset of X). I note explicitly further assumptions about the logic as they are used. We can then formulate the two rationality criteria (2.1) and (2.2) more precisely in the following definition:

(Def BS) A set K of sentences is a (nonabsurd) *belief set* iff (i) $\bot$ is not a logical consequence of the sentences in K and (ii) if $K \vdash B$, then $B \in K$.

This means that technically a belief set is what logicians normally call a *theory*. However, the interpretation here in terms of subjective beliefs is somewhat different from the standard interpretation of a theory.

It is easy to show that (Def BS) is equivalent to Stalnaker's (1984, pp. 81–82) definition of an acceptance state. He uses the following three conditions:

(2.3) If A is a member of a set of accepted sentences and A entails B, then B is a member of that set.

(2.4) If A and B are each members of a set of accepted sentences, then $A \& B$ is a member of that set.

(2.5) If A is a member of a set of accepted sentences, then $-A$ is not a member of that set.

Stalnaker then argues that each of these conditions applied to belief is motivated by his pragmatic picture.

The set of all logical consequences of a set K, that is, $\{A: K \vdash A\}$, is denoted $Cn(K)$ and is called the *consequence set* of K. A basic and useful fact about this set is that $B \in Cn(K \cup \{A\})$ iff $(A \rightarrow B) \in Cn(K)$. From (Def BS) it follows that all belief sets satisfy the following condition, which is one way of expressing the idea that epistemic states are in equilibrium:

(Cn) $K = Cn(K)$.

It turns out to be convenient for technical reasons to regard the set **L** of all sentences as a belief set. This set is called the *absurd* belief set and is denoted $K_\perp$. Note that it follows from condition (ii) in (Def BS) that the set of logically valid sentences is included in every belief set. This set, which may also be denoted $Cn(\varnothing)$, is thus the smallest belief set.

An important feature of belief sets is that they need not be *maximal* in the sense that for every sentence A either A belongs to the belief set or $-A$ belongs to it. The epistemic interpretation of this is that an individual is normally not *omniscient*. However, in some of the technical contexts to follow, maximal belief sets are of interest. Such belief sets correspond to omniscience with respect to what is expressible in **L**. In some contexts maximal belief sets have been called "state descriptions" and even "possible worlds."

In some of the applications in part II a richer language than what has been assumed so far is needed. In standard propositional languages only finite disjunctions and conjunctions are defined, but sometimes infinite disjunctions and conjunctions are useful. Let us say that the language **L** is *complete* iff, for every sequence $(A_i)_{i \in I}$, where I is an index set, there exist sentences $\bigcup_{i \in I}(A_i)$ and $\bigcap_{i \in I}(A_i)$ in **L** representing the disjunction and conjunction, respectively, of the sentences in $(A_i)_{i \in I}$. (In what follows the index set I is suppressed from the notation.) Furthermore, let us say that the relation $\vdash$ of logical consequence is complete iff it satisfies the following conditions:

($\vdash$4) For all A_i, $A_i \vdash \bigcup(A_i)$.

($\vdash$4′) For all A_i, $\bigcap(A_i) \vdash A_i$.

($\vdash$5) If $A_i \vdash B$ for all A_i, then $\bigcup(A_i) \vdash B$.

($\vdash$5′) If $C \vdash A_i$ for all A_i, then $C \vdash \bigcap(A_i)$.

In fact, it follows from de Morgan's laws that ($\vdash$4) is equivalent to ($\vdash$4′) and that ($\vdash$5) is equivalent to ($\vdash$5′). The notion of a complete logic is closely related to complete Boolean algebras.[7]

Finally, let us say that a belief set is complete iff it is closed under a complete logic. An important feature of a complete belief set K is that the conjunction of all the sentences in K is a sentence, which also is in K. Let us denote this conjunction $\bigcap K$ and call it the *determiner* for K. The name is motivated by the fact that $\bigcap K$ determines K in the sense that, for any sentence A, $A \in K$ iff $\bigcap K \vdash A$; that is, everything that is believed in K is a consequence of the single sentence $\bigcap K$.

2.3 Ellis's Belief Systems

Ellis (1976, 1979) has put forward a type of model of epistemic states that is closely related to belief sets. According to Ellis, a *belief system* is in the simplest case a set of assignments of the values T, F, and X to the sentences of a language **L**. $T(A)$ denotes the conviction (or "firm belief") that A is true, $F(A)$ the conviction that A is false, and $X(A)$ the absence of any firm belief concerning A.[8] This type of model is thus a direct way of representing a set of epistemic attitudes toward the sentence in a language **L**.

Ellis writes down a number of rationality criteria for belief systems, which he calls *acceptability* criteria (in contrast to truth conditions). The criteria are dependent on the structure of the language that is considered. In the simplest case, when **L** has the syntactic structure of the propositional calculus, the conditions are of the form '$T(A \& B)$ occurs in a belief system only if neither $F(A)$ nor $F(B)$ occur' and '$T(A \rightarrow B)$ occurs in a belief system only if $T(A)$ and $F(B)$ do not both occur'. As regards the status of these rationality criteria, Ellis says that they are "as much linguistic competence requirements as rationality laws, since they serve to define the connectives and operators [of the language **L**]" (Ellis 1979, p. 8). The ideal of rationality is satisfied mainly because of Ellis's requirement that belief systems be *completable* through every extension of **L**. This means that, for every language that syntactically includes **L**, there is some way of replacing the X evaluations that occur in a given belief system by T or F evaluations without violating any of the remaining acceptability criteria. Completability as an ideal of rationality is a consequence of "the requirement that a rational belief system be defensible against all internal criticism" (Ellis 1979, p. 9), that is, reductio ad absurdum arguments. This requirement is, of course, the same as the requirement of consistency [condition (i) in (Def BS)] in section 2.2.

Ellis states without proof that the acceptability criteria for the propositional language entail that a sentence A is a theorem of the classical propositional calculus iff $F(A)$ does not occur in any rational belief system [for a proof, see van Fraassen (1980a)].

In order to compare Ellis's belief systems with the belief sets of section 2.2, it should be noted that it is not required that $T(A \vee -A)$ occur in a belief system or that, if both $T(A)$ and $T(A \rightarrow B)$ occur, that $T(B)$ also occur. Ellis calls a belief system *strictly rational* if it satisfies these two additional requirements. In such a system one accepts as true all logical consequences of already accepted sentences. The two additional requirements thus correspond to the requirement of deductive closure [condition (ii) of (Def BS)]. Ellis seems to think that these requirements are too strong; he says that "strictly rational belief systems are really only for the gods, and they have no need for logic anyway. Therefore, I prefer the weaker concept of rationality" (Ellis 1979, p. 32). Against this it can be said that strict rationality in Ellis's sense is an idealization for certain, but so are the acceptability criteria governing rational belief systems. As mentioned earlier, Ellis also works with the equilibrium model of epistemic states. Because these criteria can be simplified if we confine ourselves to strict rationality [they can essentially be replaced by (Def BS)], there is much to be gained by such an idealization.

I now show that a strictly rational belief system in Ellis's sense is essentially identical with a belief set. Let us first assume that a nonabsurd belief set K satisfying (Def BS) is given and that the assumed logic is classical propositional logic. Define a set B_K of T, F, and X assignments in the following way: For any sentence A, A is assigned T iff $A \in K$, A is assigned F iff $-A \in K$, and A is assigned X otherwise. Then it is easy to show that B_K is a strictly rational belief system in Ellis's sense.

Conversely, suppose that a strictly rational belief system B is given. Define a set K_B of sentences as follows: $A \in K_B$ iff $T(A) \in B$. Then it is easy to show that K_B is a nonabsurd belief set governed by standard propositional logic.

These simple results show that belief sets and Ellis's strictly rational belief systems are equivalent ways of modeling epistemic states. When applying sets of sentences as models of epistemic states in what follows, I restrict myself to deductively closed sets. I prefer the belief set representation because it is more perspicuous.

I have given here only an outline of how Ellis develops, in terms of acceptability criteria, an alternative to the standard truth-condition semantics for the simplest propositional language. Ellis extends this approach to other classical and modal logics and logics of conditionals. He states completeness results for about ten different systems of logic. As regards conditional sentences, a semantical theory based on acceptability criteria that is closely related to Ellis's approach is developed in chapter 7.

2.4 Models Based on Possible Worlds

An obvious objection to using sets of sentences as models of epistemic states is that the *objects* of belief are normally not sentences but rather the *content* of sentences, that is, propositions. The characterization of propositions that has been most popular among philosophers during recent years is to identify them with *sets of possible worlds.* (An alternative construction of propositions is presented in chapter 6.) The basic semantic idea connecting sentences with propositions is then that a sentence expresses a given proposition iff it is true in exactly those possible worlds that constitute the set of worlds representing the proposition.

This set-theoretical approach to propositions makes it possible to describe some standard operations on propositions in a straightforward manner: Let $[A]$ and $[B]$ denote propositions, that is, sets of possible worlds. The conjunction of two propositions $[A]$ and $[B]$ interpreted as sets of possible worlds is represented by $[A] \cap [B]$, disjunction by $[A] \cup [B]$, and the negation of a proposition $[A]$ is $[\top] - [A] = W - [A]$. Here W is the set of all possible worlds, which also may be taken as the proposition "truth" $[\top]$ (corresponding to the sentence $\top$). We can also use the representation to introduce a relation of logical consequence between propositions: $[A]$ is a logical consequence of a set of propositions S iff the intersection of S is a subset of $[A]$. As is easily seen, the logic generated by this definition of sentential connectives includes classical propositional logic. The logic of propositions is discussed further in chapter 6.

By taking beliefs to be beliefs in propositions, we can then model an epistemic state by a set $[K]$ of possible worlds. This kind of model is called a *possible worlds model.* The epistemic interpretation of $[K]$ is that it is the *narrowest* set of possible worlds in which the individual is certain to find the actual world or, in other words, the largest set of possible worlds that is compatible with the individual's convictions.[9] Harper (1976) calls the set $[K]$ the individual's "acceptance context."

The interpretation of the set $[K]$ entails that what you accept as known in a given epistemic state is exactly what is true in all worlds in $[K]$. In a sense the use of possible worlds is a way of describing what you do *not* know. The more you learn, the fewer possible worlds are compatible with what you know. For example, I do not know whether the person who just called me on the telephone has blue or brown eyes; I do not know whether there are bats in the tower of the cathedral; and I do not know whether the train I am waiting for will be on time. Each of these possibilities demarcates a set of worlds that are compatible with what I accept as knowledge.

The central acceptability criterion is that a proposition $[A]$ is accepted in $[K]$ iff $[K]$ is a *subset* of the set of worlds representing $[A]$. Similarly $[A]$ is rejected iff $[K]$ and the set representing $[A]$ are *disjoint*. From these criteria it follows immediately that the set of accepted propositions is closed under logical consequences. The criterion that an epistemic state be consistent corresponds to the requirement that the set $[K]$ be nonempty.

Following the seminal work of Hintikka (1962), *epistemic logic* has developed as a special branch of intensional logic.[10] In this kind of logic the object languages are augmented by epistemic *operators*, so that, for example, "a knows that B" is expressed in these languages by the formula K_aB. A formal semantics for these operators is then developed in terms of possible worlds (the corresponding notion in Hintikka's writings is a model set). In contrast to this, my strategy is to "epistemize" the whole semantics, in the sense that I locate the epistemological machinery in the belief systems rather than in the object language. This does not mean that I have any aversion to epistemic logic—on the contrary. However, because I believe that the study of epistemic operators in a formal or natural language is not of primary concern for understanding the dynamics of knowledge and belief, I have chosen to keep the object language as simple as possible.

There is a close correspondence between belief sets and possible worlds models of epistemic states. An *interpretation function* I is a mapping from sentences of the language **L** to sets of possible worlds from W that satisfies the following conditions:

(I1) $I(-A) = W - I(A)$.

(I2) $I(A \& B) = I(A) \cap I(B)$.

(I3) $I(A \vee B) = I(A) \cup I(B)$.

(I4) $I(A \rightarrow B) = (W - I(A)) \cup I(B)$.

(I5) $I(\top) = W$.

(I6) $I(\bot) = \varnothing$.

Let the language **L** and an interpretation function I for **L** into W be given. It is easy to show that, if K is a nonabsurd belief set and if I assigns some sentence in K a nonempty set $I(A) = [A]$, then the intersection $[K]$ of all the sets $I(A)$ for $A \in K$ is nonempty. This intersection can then be taken as a possible worlds model. In this model all sentences in K (or rather their propositional interpretations) are accepted, and no other *sentences* from **L** are accepted. However, there may be *propositions*, that is, subsets of W, that are accepted in $[K]$, but these are not the interpretations of any sentences in K. This is possible simply because not all subsets of W need be interpretations of any sentences from **L**.

Conversely, given any nonempty possible worlds model $[K]$ together with an interpretation function I, we can define a set K of sentences as follows: $A \in K$ iff $[K] \subseteq I(A)$. It is easy to show that K is a nonabsurd belief set. Here it can be noted that, if we assume that all propositions have a name in **L**, that is, for every subset W' of W there is some sentence A in **L** such that $I(A) = W'$, then the belief set generated in this way is complete, that is, closed under infinite disjunctions and conjunctions. In particular, there is then a determiner $\bigcap K$ for the belief set such that $I(\bigcap K) = [K]$. And in this case there will be a one-to-one mapping from sentences in **L** to propositions in the power set of W.

2.5 Spohn's Generalized Possible Worlds Models

A possible worlds model (as well as a belief set) gives a crude representation of a subject's beliefs because we can express only the most elementary epistemic attitudes. This is a consequence of the fact that the set $[K]$, representing the epistemic state, divides the possible worlds into only two classes. For these models there is no possibility of expressing any degree of plausibility of different possible worlds or propositions. The best-known way of modeling *degrees* of belief is to introduce probabilities defined over a language or a class of propositions. This is the topic of the next section. In this section I present Spohn's (1987) ordinal conditional functions, which are a different way of introducing degrees of belief. Isaac Levi has pointed out that Shackle's (1961) measure of "potential surprise" is closely related to Spohn's construction.

An *ordinal conditional function*, according to Spohn, is a function k from a given set W of possible worlds into the class of ordinals such that some possible worlds are assigned the smallest ordinal 0. Intuitively k represents a *plausibility grading* of the possible worlds. The worlds that are assigned the smallest ordinals are the most plausible, according to the beliefs of the individual.

The plausibility ordering of possible worlds can be extended to an ordering of propositions, (sets of possible worlds), by requiring that the ordinal assigned to a proposition A be the *smallest* ordinal assigned to the worlds included in A; that is, $k(A) = \min\{k(w)\colon w \in A\}$. This definition entails that the plausibility ordering of propositions have the following two properties:

(2.6) For all propositions A, either $k(A) = 0$ or $k(-A) = 0$.

(2.7) For all nonempty propositions A and B, $k(A \cup B) = \min\{k(A), k(B)\}$.

We can now identify the set $[K]$ of the previous section by the set of the most plausible possible worlds, that is, the set of worlds w such that $k(w) = 0$. Following this, we can introduce the basic acceptability criterion: A proposition A is *accepted* in the epistemic state represented by the ordinal conditional function k iff $k(-A) > 0$. That this definition is the natural one follows from the fact that $k(A) = 0$ means that A and $[K]$ have some world in common; that is, A is not believed false in $[K]$. So if $k(-A) > 0$, this means that all worlds in $[K]$ must belong to A.

An important feature of ordinal conditional functions is that it makes sense to talk of greater or lesser plausibility or firmness of belief, relative to some function k. We can distinguish several cases: If both A and B are accepted [that is, if $k(-A) > 0$ and $k(-B) > 0$], we can say that A is *believed more firmly than* B iff $k(-A) > k(-B)$, that is, if the most plausible worlds outside A are less plausible than the most plausible worlds outside B. There are other cases where A is more plausible than B. First, the case where A is accepted and B is not, that is, where $k(-A) > k(-B) = 0$. Second, there is the case where A is not believed false but B is; that is, $k(A) = 0 < k(B)$. Finally, we have the case where both A and B are believed false but A less firmly so, that is, $0 < k(A) < k(B)$. This leads us to the following definition: A is *more plausible than* B relative to k iff $k(-A) > k(-B)$ or $k(A) < k(B)$.

We thus see that representing epistemic states by ordinal conditional functions makes it possible to introduce more interesting epistemic attitudes, to wit, "believed more firmly than" and "more plausible than," beside the standard "accepted," "rejected," and "kept in suspense." However, the full forces of the ordinal conditional functions appears only in the next chapter, when I turn to changes of epistemic states.

2.6 An Example from Artificial Intelligence: Doyle's Truth Maintenance System

So far the models of epistemic states that have been presented have been taken from the philosophical literature. There are, however, other areas of research where models of epistemic states are important. In general terms it can be said that such models are important for cognitive science. In this section I briefly comment on the use of semantic networks as models of epistemic states and then present a particularly interesting example from artificial intelligence (AI), to wit, Doyle's (1979) truth maintenance system.

Probably the most common models of epistemic states used in cognitive science are those called *semantic networks.* A semantic network typically consists of a set of *nodes* representing some objects of belief and, connecting the nodes, a set of *links* representing relations between the nodes. The networks are then complemented by some implicit or explicit interpretation rules that make it possible to extract beliefs and epistemic attitudes. Changing a semantic network consists in adding or deleting nodes or links.

Different semantic networks have different types of objects as nodes and different kinds of relations as links. In fact, the diversity is so large that it is difficult to see what the various networks have in common. It seems that any kind of object can serve as a node in the networks and that any type of relation or connection between nodes can be used as a link between nodes. This diversity seems to undermine the claims that semantic networks represent epistemic states, and it raises the question of what they have to do with semantics. In his excellent methodological article, Woods (1975) admits that "we must begin with the realization that there is currently no 'theory' of semantic networks" (p. 36). As a preliminary to such a theory, Woods formulates requirements for an adequate notation for semantic networks and explicit interpretation rules for such a notation. My aim here is not a presentation of his ideas but a brief outline of semantic networks as models of epistemic states.

Turning now to artificial intelligence, the philosophically most important problem seems to be the frame problem. In general terms, this problem can be defined as the problem of finding a (basically epistemic) model that permits changing and complex information about the world to be represented in an adequate and efficient way.[11] A solution to the frame problem would be a way of describing the *form* of the world (or a substantial part of it) that would enable us to translate efficiently our knowledge and beliefs about the world into a computer program. It should be noted that the frame problem is not a question of content but one of form: The simple belief sets presented in section 2.2 are perfectly capable of representing the blocks worlds studied in AI, but these models are badly suited for describing *actions* that may be performed in such worlds and the *changes* of the world the actions bring about.

Doyle's (1979) truth maintenance system (TMS) is an attempt to model changes of belief within an AI setting and, as such, is of direct relevance for the frame problem. As Doyle remarks (p. 232), the name "truth maintenance system" not only sounds like Orwellian Newspeak, but is also a misnomer, because what is maintained is the consistency of beliefs and reasons for belief. Doyle (1983) later changed the name to "reason maintenance system." In a broad sense TMS can be said to be a semantic network model, but its belief structure and its techniques for handling changes of belief are more sophisticated than in other semantic network models.

There are two basic types of entities in TMS: *nodes* representing propositional beliefs and *justifications* representing reasons for beliefs. These justifications may be other beliefs from which the current belief is derived. A node may be *in* or *out*, which corresponds to the epistemic attitudes of accepting and not accepting the belief represented by the node. As should be expected, if a certain belief is out in the system, this does not entail that its negation is in. On the other hand, as a rationality requirement, if both a belief and its negation are in, then the system will start a *revision* of the set of nodes and their justifications in order to reestablish consistency.

A justification consists of a pair of lists: an inlist and an outlist. A node is in if and only if it has some justification (there may be several for the same node), the inlist of which contains only nodes that are in and the outlist of which contains only nodes that are out. A particular type of justification, called "nonmonotonic justification," is used to make tentative guesses within the system. For example, a belief in A can be justified simply by the fact that the belief in $-A$ is out. Beliefs that are justified in this way are called *assumptions*. This technique gives us a way of representing commonsense "default" expectations. It also leads to *nonmonotonic* reasoning in the following sense: If belief in A is justified only by the absence of any justification for $-A$, then a later *addition* of a justification for $-A$ will lead to a *retraction* of the belief in A. The general problem of non-monotonic reasoning and retraction of beliefs is analyzed in chapter 3.

The basic concepts of TMS are best illustrated by an example:

Node	Justification		Status
	Inlist	Outlist	
(N1) Oscar is not guilty of defamation.	(N2)	(N3)	in
(N2) The accused should have the benefit of the doubt.	–	–	in
(N3) Oscar called the queen a harlot.	(N4), (N5)	–	out
(N4) It may be assumed that the witness's report is correct.	–	–	in
(N5) The witness says he heard Oscar call the queen a harlot	–	–	out

In this situation (N1) is in because (N2) is in and (N3) is out. Node (N3) is out because not both of (N4) and (N5) are in. If (N5) changes status to in (this may be assumed to be beyond the control of the system), (N3) will become in and consequently assumption (N1) is out.

Apart from the representations of nodes and justifications as presented here, TMS contains techniques for handling various problems that arise when the system of beliefs is adjusted to accommodate the addition of a new node or justification. In particular, when a contradiction is found, the system uses a form of backtracking to find the fundamental assumptions that directly or indirectly give support to the contradiction. One of these assumptions is chosen as the culprit and is given the status out. This process sometimes needs to be iterated, but it is beyond the scope of this chapter to give a full description of the mechanics of TMS.

However, the TMS representation of beliefs is not without epistemological problems. The example can be used to illustrate some of the drawbacks of TMS. In the handling of beliefs and justifications, TMS takes no notice of what the nodes happen to stand for. The sentences that I have added to the node names are not interpreted in any way by the system. This means that much of the *logic* of propositions is lost in the TMS representation of beliefs. All forms of logical inferences that are to be used by the system have to be reintroduced as special systems of justifications.

Doyle discusses conditional proofs, but the process for handling such inferences seems extremely complex.

Furthermore, as regards the frame problem, TMS is not an ideal solution because it leaves much of the work to the programmer. The programmer produces the nodes and their justifications; she has to organize the information in levels, and she also has to decide on how contradictions are to be engineered. In short, the programmer may end up in a task that is no easier than describing the relevant beliefs and their connections in a belief set.

I find TMS to be an interesting way of modeling epistemic states. In particular, the idea of including justifications for the beliefs held seems fruitful. This aspect of belief systems seems to be neglected in other models of epistemic states. An exception, however, is Spohn (1983a), who tries to define the notion of one belief being a *reason* for another within his model of states of belief.

Harman (1986, ch. 4) introduces a distinction between two competing theories of belief revision that is helpful here: the *foundations theory*, which holds that one needs to keep track of one's original justifications for belief, and the *coherence theory*, which holds that one need not.

> The foundations theory holds that some of one's beliefs "depend on" others for their current justification; these other beliefs may depend on still others, until one gets to foundational beliefs that do not depend on any further beliefs for their justification. In this view reasoning or belief revision should consist, first, in subtracting any of one's beliefs that do not have a satisfactory justification and, second, in adding new beliefs that either need no justification or are justified on the basis of other justified beliefs one has.
>
> On the other hand, according to the coherence theory, it is not true that one's ongoing beliefs have or ought to have the sort of justificational structure required by the foundations theory. In this view ongoing beliefs do not usually require any justification. Justification is taken to be required only if one has a special reason to doubt a particular belief. Such a reason may consist in a conflicting belief or in the observation that one's beliefs could be made more "coherent," that is, more organized or simpler or less ad hoc, if the given belief were abandoned (and perhaps if certain other changes were made). According to the coherence theory, belief revision should involve minimal changes in one's beliefs in a way that sufficiently increases overall coherence. (Harman 1986, pp. 29–30)

Levi (1980) defends the coherence theory. He claims (p. 1) that knowledge is not "a matter of pedigree." It is clear that, with the exception of Doyle's TMS, all models of epistemic states presented in this chapter adhere to the coherence theory. This is true also of the probabilistic models to be presented in the following sections. Harman (1986, ch. 4) discusses the pros and cons of the foundations theory versus the coherence theory. Among other things he reviews some psychological experiments [see Ross and Anderson (1982)] that tend to support the coherence theory. It seems to be a matter of fact that people do not keep track of the justifications for their beliefs. The main reason for this is that it would soon lead to a combinatorial explosion, and it is a matter of the *economy of thought* to avoid cluttering one's mind. Harman sees this as the decisive reason in favor of the coherence theory. However, it should be admitted that, in the view of epistemic states as equilibrium points adopted here, combinatorial considerations are of less importance than other rationality criteria.

2.7 Bayesian Models

The best-known models of epistemic states and of how epistemic inputs affect an epistemic state are models that are based on *subjective* or *personalistic probabilities.* A central part of the Bayesian doctrine is that the epistemic state of an individual can be represented by a *probability function* defined over a language or a set of possible worlds. The intended interpretation is that such a probability function provides a measure of the individual's *degrees of belief* in the sentences or propositions. These degrees of belief then provide us with a richer repertoire of epistemic attitudes than is possible with belief sets or traditional possible worlds models.

The criterion of rationality that is used to motivate the assumption that degrees of belief should be represented by a probability function is that degrees of belief should be *coherent.* The meaning of coherence is best explained by a brief description of the so-called Dutch book theorem. In the form of Bayesianism advocated by Ramsey (1931b) and de Finetti (1937), it is assumed that the probability an individual assigns to a sentence can be determined with the aid of his inclination to accept *bets* concerning the truth of the sentence. If he, for example, accepts a bet that A is true at the odds of 1:4, then this is taken to imply that he estimates the probability of A to be at least $1/(1 + 4) = 0.2$. The Dutch book theorem says that if it is not possible to construct a bet where the individual will lose money no matter what happens (a Dutch book), then there is a unique probability measure that describes the individual's degrees of belief in the different sentences of the language.[12] It should be noted that a necessary assumption for this theorem is that, if an individual is not willing to bet on A at odds of $a:b$, then he should be willing to bet on $-A$ at odds of $b:a$. In the betting context the requirement that the individual's beliefs be coherent is simply that no Dutch book can be made against him. The Dutch book theorem then shows that coherence entails that the individual's degrees of belief can be represented by a probability function.

The Bayesian models of epistemic states have been used extensively within many areas of decision theory and game theory. In combination with a utility measure the subjective probabilities have been used in various forms of decision rules, the most familiar being the principle of maximizing expected utility. In decision-theoretic contexts it is often assumed that all information about the world that is relevant for the decision maker is conveyed by his subjective probability function, that is, the Bayesian model of an epistemic state.

I have not yet raised the question of how the probability function on **L** (or W) is to be defined. We want to consider a numerical function defined on **L** that obeys the laws of probability. The standard formulation of these laws is as follows:

(2.8) $0 \leqslant P(A) \leqslant 1$ for all sentences A in **L**.

(2.9) $P(\top) = 1$.

(2.10) For all sentences A and B, if A and B are logically disjoint [that is, if $\vdash -(A \& B)$], then $P(A \vee B) = P(A) + P(B)$.

However, this formulation of the axioms presumes that we know the *logic* of **L**, because otherwise we cannot determine when two sentences are logically disjoint. In order to prepare the ground for chapter 6, I use instead an axiomatization of the laws of probability that shows that we need not presuppose a logic. On the contrary, the axiomatization can be used to *define* a logic for **L**. Historically the first axiomatization of this kind is Popper's (1959, app. *iv) axioms for conditional probability. For my purposes an axiomatization of unconditional probability is sufficient, and I can use the following definition, which is adopted from Stalnaker (1970):

(Def Prob) A (language-based) *probability function* is a function P from sentences in **L** into real numbers that meets the following six conditions for all sentences A, B, and C:

(i) $0 \leqslant P(A) \leqslant 1$.

(ii) $P(A) = P(A \,\&\, A)$.

(iii) $P(A \,\&\, B) = P(B \,\&\, A)$.

(iv) $P(A \,\&\, (B \,\&\, C)) = P((A \,\&\, B) \,\&\, C)$.

(v) $P(A) + P(-A) = 1$.

(vi) $P(A) = P(A \,\&\, B) + P(A \,\&\, -B)$.

In this definition we assume that $-$ and & are the only primitive connectives and that the remaining connectives are defined in the standard way. It is worth noting that Stalnaker introduces these probability functions "as an autonomous semantics for propositional calculus, based on the concept of knowledge rather than truth" (Stalnaker 1970, p. 65). This semantic program has since been developed by Leblanc (1983) and Field (1977) among others.

It is easy to show that, on the standard axiomatization of probability functions, the six conditions in (Def Prob) are satisfied. In order to show that we can use (Def Prob) to derive the standard axiomatization of probability functions, we must first show how it can be used to generate a *logic* for **L**. Let us say that a sentence A is *logically valid* iff $P(A) = 1$ for all probability functions that satisfy conditions (i)–(vi). On the basis of this we can then say that two sentences A and B are *logically disjoint* iff $-(A \,\&\, B)$ is logically valid. From these definitions the following lemmas can be established:[13]

LEMMA 2.1 A sentence is logically valid iff it is a truth-functional tautology.

LEMMA 2.2 If A and B are logically disjoint, then, for any probability function P, $P(A \vee B) = P(A) + P(B)$.

Lemma 2.2 expresses the key axiom in the standard axiomatization of probability functions. Because the other axioms are trivial, we conclude that the standard axiomatization and the one presented here are essentially equivalent.

I next turn to the connections between belief sets and the probability models presented here. Probabilities represent degrees of belief, and belief sets represent beliefs that are accepted as certain. In a probabilistic model it is natural to define the second notion by saying that a sentence A is *accepted as certain relative to* P iff $P(A) = 1$. We can then introduce a relation between probability functions and sets of sentences:

(Def Gen) A probability function P defined over **L** *generates* the set K of sentences iff, for all sentences A in **L**, $P(A) = 1$ iff $A \in K$.

In other words, the set generated by P is the set of sentences that are accepted as certain relative to P. The relation is many to one—different probability functions may generate the same set of sentences. The following result is trivial but useful [for a proof, see Gärdenfors (1978a), p. 398]:

LEMMA 2.3 K is a nonabsurd belief set iff there exists some probability function P that generates K.

This result shows that, in applications where we are interested only in the sentences that are accepted as certain, we need not use the full complexity of probability functions but can confine ourselves to belief sets.

(Def Gen) identifies the accepted sentences with the sentences that have *full belief*, that is, probability 1. Some authors, for example, de Finetti, Carnap, and Levi, drive a wedge between acceptability and maximal probability and thus allow that some sentences that have probability 1 are not accepted [see Levi (1980, ch. 5)]. For example, in a problem of estimating the value of a real-valued parameter x that ranges between 0 and 1, the hypothesis that x has an irrational value normally has probability 1, but, according to these authors, it should not be accepted. Even if a distinction between acceptability and full belief is motivated in some cases, it does not play any role in this book; so, in order to avoid unnecessary complications, I assume (Def Gen) in chapter 5 and to some extent in the applications in part II.

Corresponding to his belief systems, Ellis (1979) also discusses a class of models of epistemic states that use probability values. These models are based on a set of "P-evaluations," which are assignments of subjective probabilities to sentences in **L**. In addition to evaluations of the form $P(A) = x$, Ellis also allows evaluations of the form $Y(A)$, which occur in the model of an individual's epistemic state when she has no determinate degree of belief in A. Thus these systems of P-evaluations are related to standard Bayesian probability models just like Ellis's belief systems are related to belief sets. Forrest (1986) develops Ellis's probabilistic models into a *dynamic* setting. An important difference between the dynamics of probabilistic models to be presented in chapter 5 and Forrest's system is that he does not make the idealizing assumption that belief systems be closed under logical implications. Besides this, his typology of epistemic changes is different from the one presented there. Some further models of "indeterminate" degrees of belief are discussed in the next section.

It is, of course, also possible to have a probabilistic version of the possible worlds models. Such a model consists of a probability function defined over sets of possible worlds taken from some given set *W*. This kind of model is studied by several authors, and an extended version of it is introduced in chapter 8.[14]

A central feature of probabilistic possible worlds models is that, because propositions are defined as sets of possible worlds, we can immediately assign probabilities to propositions as soon as we have a probability function defined over *W*. And if we have an interpretation function *I* available that takes sentences into propositions, we can assign probabilities to sentences by saying that the probability of the sentence is the same as the probability of the set of possible worlds that is its interpretation. It should come as no surprise that such a probability assignment to the sentences in **L** satisfies (Def Prob). Thus it is clear that, in light of the results of section 2.6, a probability model defined over a set *W* of possible worlds is but a slight generalization of a probability model defined over a language **L**.

2.8 Generalized Probabilistic Models

The main rationale for the Bayesian way of modeling epistemic states is the Dutch book theorem or closely related results. As was mentioned earlier, a presupposition for the theorem is that an individual be willing to take either side of a bet so that, if he is not willing to bet on a sentence A at odds of $a:b$, he should be willing to bet on $-A$ at odds of $b:a$. A general criticism of the Bayesian models is that this requirement demands too much of people's willingness to accept bets, and, perhaps more important, it is far from certain that the requirement is rationally motivated. People are often not willing to accept either of the two bets, and they may have good reasons for not doing so. This criticism of the standard Bayesian models of epistemic states is directed against the assumptions needed for the Dutch book theorem, but similar points can be made against other arguments for representing beliefs by a single probability measure.

A probability function is one way of representing beliefs that are partial in the sense that they are neither accepted nor rejected. The probability function provides detailed information about how strongly the beliefs are held, but, as we have seen, such a function may be unrealistically detailed.

In this section I present some other ways of representing partial beliefs that have been suggested within theories of decision making. These models are based on less demanding assumptions about an individual's behavior than in the standard Bayesian model.

The first alternative way of describing partial beliefs is to associate with each sentence in a language (or each proposition in a possible worlds context) a *probability interval.*[15] The intended interpretation of such an interval is that the information available to the individual entails that the "true" probability of the sentence is a point in the interval and that no narrower interval is justified by this information. On this interpretation the width of the probability interval associated with a sentence A can be taken as a measure of the individual's degree of ignorance of the probability of A. The interval associated with a sentence A can be denoted $(P_*(A), P^*(A))$, where $0 \leqslant P_*(A) \leqslant P^*(A) \leqslant 1$. $P_*(A)$ and $P^*(A)$ are called the *lower* and *upper* probabilities of A, respectively.

The rationality criteria governing the assignment of probability intervals take the form of certain restrictions that guarantee that the assignment is *coherent*. If we consider only a single sentence A, then it must be the case that $P_*(A) = 1 - P^*(-A)$ and $P^*(A) = 1 - P_*(-A)$ in order to maintain the interpretation that every value within the interval assigned to A is a "possible" probability of A. Generally the coherence restriction is that, for every value x within the interval $(P_*(A), P^*(A))$, there must be values within the intervals associated with the other sentences in the language such that x together with these values form a standard probability function over the language [see Gärdenfors (1979b), p. 169]. If this criterion is not fulfilled for some assignment of intervals, it is possible to construct a Dutch book against an individual using these intervals as a basis for his betting behavior.

The second generalization of the traditional Bayesian models that I consider is based on utilizing a *class* **P** of probability functions, instead of a single probability function, to represent the beliefs of an individual. The interpretation of the set **P** is that the information available to the individual is not sufficient to single out a unique probability function, but it determines a class of *epistemologically possible* [or, as Levi (1974) calls them, "permissible"] probability functions. A probability function is said to be epistemologically possible if it does not contradict the individual's knowledge in the given epistemic state. Several authors have suggested that a state of belief be represented by such a class of probability functions.[16]

Levi (1974) also assumes that the set **P** is *convex*, which means that, if P and P' are two functions in **P**, then for all a, $0 \leqslant a \leqslant 1$, the "mixture" function $a \cdot P + (1 - a) \cdot P'$ is also in **P**. (The mixture function is defined by $[a \cdot P + (1 - a) \cdot P'](A) = a \cdot P(A) + (1 - a) \cdot P'(A)$ for all A. Mixture functions and their properties are analyzed in section 5.2.) The interpretation of this assumption is that, if P and P' are both epistemologically possible probability functions according to the beliefs of an individual, then any mixture of these functions is also possible.

Note that, if **P** is assumed to be convex, then for any sentence A the set of probabilities assigned to A by the functions in **P** form an interval (possibly open). In this sense we can say that the set **P** generates an interval assignment to the sentences of **L**. Representing beliefs by a convex set of probability functions is, however, a more general method than representing them by probability intervals, because from a convex set of probability functions a unique set of associated intervals can be computed; but if we start from a given assignment of coherent probability intervals, there is in general a large number of convex sets of probability functions that generate the intervals. And when it comes to decision making, this may make a difference. Levi (1974, pp. 416–417) presents an example that shows that there may be two decision situations with the same alternative states and outcomes but with different sets of "permissible" probability functions, which give different decisions when his theory is used, although the probability intervals that are generated are identical.

It is argued in Gärdenfors and Sahlin (1982) that not all aspects of an individual's beliefs that are relevant for decision making can be captured by a set of probability functions. As a second element in the models of states of belief, in addition to the set **P** a measure r of the *epistemic reliability* of the probability functions in **P** is introduced. The motivation for including r is that, even if several probability functions are epistemologically possible, some distributions are more reliable to the individual—they are backed up by more evidence or information than other distributions. The measure of epistemic reliability is intended to represent these degrees of information. In a sense the measure r can be seen as a probabilistic generalization of Spohn's notion of believing more firmly, as presented in section 2.5.

The concept of epistemic reliability is also closely related to the "weight of evidence" that was introduced by Keynes (1921, p. 71). In fact, this way of extending the Bayesian representation of beliefs was forestalled by Peirce:

> To express the proper state of belief, not *one* number but *two* are requisite, the first depending on the inferred probability, the second on the amount of knowledge on which that probability is based. (Peirce 1932, p. 421)

For a discussion of the weight of evidence, see Gärdenfors (1979b, pp. 176–178).

I have now presented some of the main extensions of the standard Bayesian models of belief. There are several variants of each of the types that have been presented here, but my aim has not been to give a complete list. A general approach to both belief set representations and probabilistic models can be found in Domotor (1983). Yet another extension of the Bayesian model is presented in chapter 8 as a tool for analyzing explanations.

2.9 Johnson-Laird's Mental Models

Most of the models of epistemic states presented so far have presumed sentences or propositions as the building blocks of the models. Furthermore, the models have been governed by a propositional logic. In this concluding section I give an outline of Johnson-Laird's (1983) "mental models," which deal with reasoning in first-order logic and which have "things" as building blocks.

One of the main applications of Johnson-Laird's models is *syllogistic reasoning*. He considers syllogisms that involve two premises of the forms 'All X are Y', 'Some X are Y', and 'No X are Y'. The key step in the modeling of premises in a syllogism is to set up a *table* of "actors" taking different "roles" (Johnson-Laird 1983, ch. 5). For example, the premise 'All artists are beekeepers' may be represented as follows:

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

Here three actors are playing the joints roles of artists and beekeepers, and two (optional) actors are taking the role of the beekeepers who are not artists. If a second premise, 'All the beekeepers are chemists', is now added, the table is *expanded* to accommodate this premise in the following way:

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

By inspecting this table, we can readily determine that the conclusion 'All the artists are chemists' follows.

Similarly the premise 'None of the authors are burglars' can be modeled by "fencing off" the actors:

```
author
author
author
-----------------
          burglar
          burglar
          burglar
```

Suppose now that we want to add the premise 'Some of the chefs are burglars'. The most straightforward way of doing this is as follows:

```
(i)  author
     author
     author
     ------------------------
               burglar = chef
               burglar = chef
               (burglar)
                          (chef)
```

If this table is used, it is tempting to conclude that 'None of the authors are chefs', which some experimental subjects do. However, here are two other ways of modeling the premises:

```
(ii)  author
      author
      author   =        chef
      ------------------------
                burglar = chef
                burglar = chef
                (burglar)
```

```
(iii) author  =        chef
      author  =        chef
      author  =        chef
      ____________________
               burglar = chef
               burglar = chef
               (burglar)
```

Because all these tables are consistent with the premises, it follows that the conclusion 'None of the authors are chefs' is invalid (and so is 'Some of the authors are not chefs'). Given these three tables, it may be tempting to claim that there is no valid conclusion connecting the authors with the chefs (as many subjects do). However, in all three 'Some of the chefs are not authors' is valid.

I have here only presented Johnson-Laird's method of modeling syllogistic reasoning by way of examples. The reader is referred to Johnson-Laird (1983) for the theoretical basis of these models. He there uses the tables and the various places where the constructions can go *wrong* in the subjects' minds (for example, not seeing all possible types of tables for a particular pair of premises) to explain several experimental findings concerning people's performance in syllogistic reasoning.

Another application in Johnson-Laird's book is a computer program that constructs models of *spatial descriptions*. For example, if given the input

X is to the right of Y
Z is in front of Y
W is to the left of Z,

the program constructs the following spatial representation:

```
   Y  X
W  Z
```

The interesting cases occur when the program is given indeterminate descriptions, such as

X is to the right of Y
Z is to the left of X.

Even if this information does not fully specify the model, the program nevertheless constructs a model, for example:

Z Y X

However, if the program is subsequently informed that

Z is to the right of Y,

then this information *contradicts* the constructed model. This forces the program to check whether there is any *revision* of the model that is consistent with all the information available. This is accomplished by (recursively) rearranging the indetermined positions of the items in the model and checking whether the rearrangement satisfies the premises. In the example given here, the program returns the model

Y Z X,

as expected.

Also for this application I have chosen to present Johnson-Laird's theory by means of examples. My main reason for presenting these applications is that the models do *not* use sentences or propositions as building blocks. As we have seen, however, it is still possible to make *inferences* from the models. Furthermore, I want to emphasize that *expansions* and *revisions* can be represented in a systematic manner for these models.