Chapter 1

UNCERTAINTY IN AI SYSTEMS: AN OVERVIEW

I consider the word probability as meaning the state of mind with respect to an assertion, a coming event, or any other matter on which absolute knowledge does not exist.
— August De Morgan, 1838

1.1 INTRODUCTION

1.1.1 Why Bother with Uncertainty?

Reasoning about any realistic domain always requires that some simplifications be made. The very act of preparing knowledge to support reasoning requires that we leave many facts unknown, unsaid, or crudely summarized. For example, if we choose to encode knowledge and behavior in rules such as "Birds fly" or "Smoke suggests fire," the rules will have many exceptions which we cannot afford to enumerate, and the conditions under which the rules apply (e.g., seeing a bird or smelling smoke) are usually ambiguously defined or difficult to satisfy precisely in real life. Reasoning with exceptions is like navigating a minefield: Most steps are safe, but some can be devastating. If we know their location, we can avoid or defuse each mine, but suppose we start our journey with a map the size of a postcard, with no room to mark down the exact location of every mine or the way they are wired together. An alternative to the extremes of ignoring or enumerating exceptions is to *summarize* them, i.e., provide some warning signs to indicate which areas of the minefield are more dangerous than others. Summarization is

essential if we wish to find a reasonable compromise between safety and speed of movement. This book studies a language in which summaries of exceptions in the minefield of judgment and belief can be represented and processed.

1.1.2 Why Is It a Problem?

One way to summarize exceptions is to assign to each proposition a numerical measure of uncertainty and then combine these measures according to uniform syntactic principles, the way truth values are combined in logic. This approach has been adopted by first-generation expert systems, but it often yields unpredictable and counterintuitive results, examples of which will soon be presented. As a matter of fact, it is remarkable that this combination strategy went as far as it did, since uncertainty measures stand for something totally different than truth values. Whereas truth values in logic characterize the formulas under discussion, uncertainty measures characterize *invisible* facts, i.e., exceptions not covered in the formulas. Accordingly, while the syntax of the formula is a perfect guide for combining the visibles, it is nearly useless when it comes to combining the invisibles. For example, the machinery of Boolean algebra gives us no clue as to how the exceptions to $A \rightarrow C$ interact with those of $B \rightarrow C$ to yield the exceptions to $(A \wedge B) \rightarrow C$. These exceptions may interact in intricate and clandestine ways, robbing us of the modularity and monotonicity that make classical logic computationally attractive.

Although formulas interact in intricate ways, in logic too, the interactions are visible. This enables us to calculate the impact of each new fact *in stages*, by a process of derivation that resembles the propagation of a wave: We compute the impact of the new fact on a set of syntactically related sentences S_1, store the results, then propagate the impact from S_1 to another set of sentences S_2, and so on, without having to return to S_1. Unfortunately, this computational scheme, so basic to logical deduction, cannot be justified under uncertainty unless one makes some restrictive assumptions of *independence*.

Another feature we lose in going from logic to uncertainty is *incrementality*. When we have several items of evidence, we would like to account for the impact of each of them individually: Compute the effect of the first item, then absorb the added impact of the next item, and so on. This, too, can be done only after making restrictive assumptions of independence. Thus, it appears that uncertainty forces us to compute the impact of the entire set of past observations to the entire set of sentences in one global step—this, of course, is an impossible task.

1.1.3 Approaches to Uncertainty

AI researchers tackling these problems can be classified into three formal schools, which I will call *logicist*, *neo-calculist*, and *neo-probabilist*. The logicist school

attempts to deal with uncertainty using nonnumerical techniques, primarily nonmonotonic logic. The neo-calculist school uses numerical representations of uncertainty but regards probability calculus as inadequate for the task and thus invents entirely new calculi, such as the Dempster-Shafer calculus, fuzzy logic, and certainty factors. The neo-probabilists remain within the traditional framework of probability theory, while attempting to buttress the theory with computational facilities needed to perform AI tasks. There is also a school of researchers taking an informal, heuristic approach [Cohen 1985; Clancey 1985; Chandrasekaran and Mittal 1983], in which uncertainties are not given explicit notation but are instead embedded in domain-specific procedures and data structures.

This taxonomy is rather superficial, capturing the syntactic rather than the semantic variations among the various approaches. A more fundamental taxonomy can be drawn along the dimensions of *extensional* vs. *intensional* approaches.† The extensional approach, also known as *production* systems, *rule-based* systems, and *procedure-based* systems, treats uncertainty as a generalized truth value attached to formulas and (following the tradition of classical logic) computes the uncertainty of any formula as a function of the uncertainties of its subformulas. In the intensional approach, also known as *declarative* or *model-based*, uncertainty is attached to "states of affairs" or subsets of "possible worlds." Extensional systems are computationally convenient but semantically sloppy, while intensional systems are semantically clear but computationally clumsy. The trade-off between semantic clarity and computational efficiency has been the main issue of concern in past research and has transcended notational boundaries. For example, it is possible to use probabilities either extensionally (as in PROSPECTOR [Duda, Hart, and Nilsson 1976]) or intensionally (as in MUNIN [Andreassen et al. 1987]). Similarly, one can use the Dempster-Shafer notation either extensionally [Ginsberg 1984] or intensionally [Lowrance, Garvey, and Strat 1986].

1.1.4 Extensional vs. Intensional Approaches

Extensional systems, a typical representative of which is the certainty-factors calculus used in MYCIN [Shortliffe 1976], treat uncertainty as a generalized truth value; that is, the certainty of a formula is defined to be a unique function of the certainties of its subformulas. Thus, the connectives in the formula serve to select the appropriate weight-combining function. For example, the certainty of the conjunction $A \wedge B$ is given by some function (e.g., the minimum or the product) of

† These terms are due to Perez and Jirousek (1985); the terms *syntactic* vs. *semantic* are also adequate.

the certainty measures assigned to A and B individually. By contrast, in intensional systems, a typical representative of which is probability theory, certainty measures are assigned to sets of worlds, and the connectives combine sets of worlds by set-theory operations. For example, the probability $P(A \wedge B)$ is given by the weight assigned to the intersection of two sets of worlds—those in which A is true and those in which B is true—but $P(A \wedge B)$ cannot be determined from the individual probabilities $P(A)$ and $P(B)$.

Rules, too, have different roles in these two systems. The rules in extensional systems provide licenses for certain symbolic activities. For example, a rule $A \xrightarrow{m} B$ may mean "If you see A, then you are given the license to update the certainty of B by a certain amount which is a function of the rule strength m." The rules are interpreted as a summary of past performance of the problem solver, describing the way an agent normally reacts to problem situations or to items of evidence. In intensional systems, the rules denote elastic constraints about the world. For example, in the Dempster-Shafer formalism (see Chapter 9) the rule $A \xrightarrow{m} B$ does not describe how an agent reacts to the finding of A, but asserts that the set of worlds in which A and $\neg B$ hold simultaneously has low likelihood and hence should be excluded with probability m. In the Bayesian formalism the rule $A \xrightarrow{m} B$ is interpreted as a conditional probability expression $P(B \mid A) = m$, stating that among all worlds satisfying A, those that also satisfy B constitute a fraction of size m. Although there exists a vast difference between these two interpretations (as will be shown in Chapters 9 and 10), they both represent summaries of factual or empirical information, rather than summaries of past decisions. We will survey intensional formalisms in Section 1.3, but first, we will briefly discuss their extensional rivals.

1.2 EXTENSIONAL SYSTEMS: MERITS, DEFICIENCIES, AND REMEDIES

1.2.1 Computational Merits

A good way to show the computational merits of extensional systems is to examine the way rules are handled in the certainty-factors formalism [Shortliffe 1976] and contrast it with probability theory's treatment of rules. Figure 1.1 depicts the combination functions that apply to serial and parallel rules, from which one can form a *rule network*. The result is a modular procedure for determining the certainty of a conclusion, given the credibility of each rule and the certainty of the premises (i.e., the roots of the network). To complete the calculus we also need to define combining functions for conjunction and negation. Setting mathematical details aside, the point to notice is that the same combination function applies

uniformly to any two rules in the system, regardless of what other rules might be in the neighborhood.

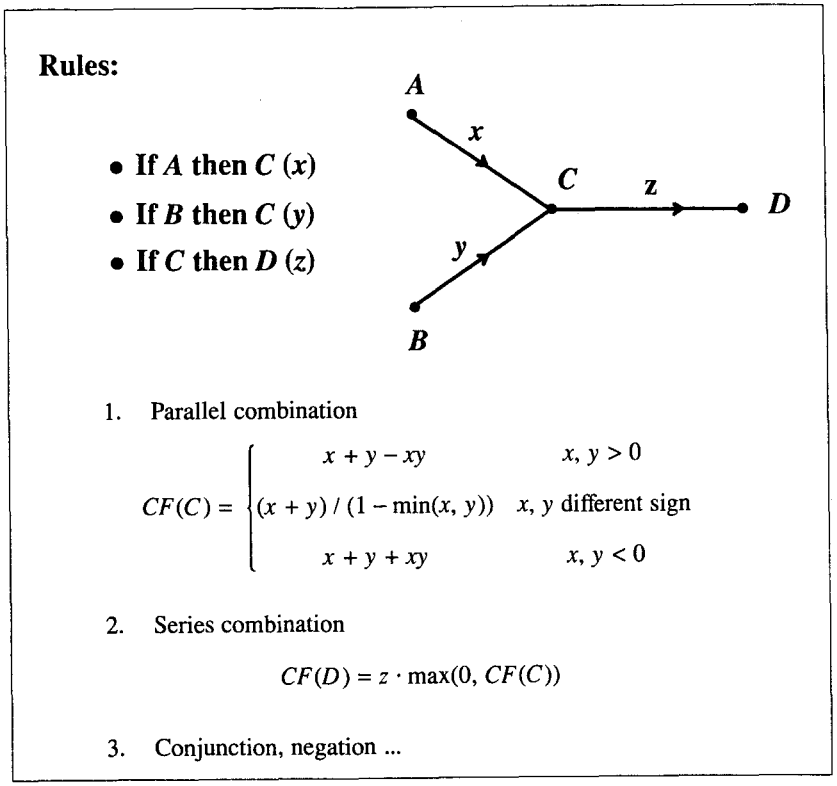

Figure 1.1. *Certainty combination functions used in MYCIN. x, y, and z denote the credibilities of the rules.*

Computationally speaking, this uniformity mirrors the modularity of inference rules in classical logic. For example, the logical rule "If A then B" has the following procedural interpretation: "If you see A anywhere in the knowledge base, then regardless of what other things the knowledge base contains and regardless of how A was derived, you are given the license to assert B and add it to the database." This combination of *locality* ("regardless of other things") and *detachment* ("regardless of how it was derived") constitutes the principle of *modularity*. The numerical parameters that decorate the combination functions in Figure 1.1 do not alter this basic principle. The procedural license provided by the rule $A \xrightarrow{x} B$ reads as follows: "If you see the certainty of A undergoing a change δ_A, then regardless of what other things the knowledge base contains and

regardless of how δ_A was triggered, you are given an unqualified license to modify the current certainty of B by some amount δ_B, which may depend on x, on δ_A, and on the current certainty of B."†

To appreciate the power of this interpretation, let us compare it with that given by an intensional formalism such as probability theory. Interpreting rules as conditional probability statements, $P(B \mid A) = p$, does not give us license to do anything. Even if we are fortunate enough to find A true in the database, we still cannot assert a thing about B or $P(B)$, because the meaning of the statement is "If A is true and A is the *only* thing that you know, then you can attach to B a probability p." As soon as other facts K appear in the database, the license to assert $P(B) = p$ is automatically revoked, and we need to look up $P(B \mid A, K)$ instead. The probability statement leaves us totally impotent, unable to initiate any computation, unless we can verify that everything else in the knowledge base is irrelevant. This is why verification of irrelevancy is so crucial in intensional systems.

In truth, such verifications are crucial in extensional systems too, but the computational convenience of these systems and their striking resemblance to logical derivation tempt people to neglect the importance of verifying irrelevancy. We shall now describe the semantic penalties imposed when relevance considerations are ignored.

1.2.2 Semantic Deficiencies

The price tag attached to extensional systems is that they often yield updating that is incoherent, i.e., subject to surprises and counterintuitive conclusions. These problems surface in several ways, most notably

1. improper handling of bidirectional inferences,
2. difficulties in retracting conclusions, and
3. improper treatment of correlated sources of evidence.

We shall describe these problems in order.

THE ROLE OF BIDIRECTIONAL INFERENCES

The ability to use both predictive and diagnostic information is an important component of plausible reasoning, and improper handling of such information leads to rather strange results. A common pattern of normal discourse is that of

† The observation that the rules refer to changes rather than absolute values was made by Horvitz and Heckerman [1986].

abductive reasoning—if A implies B, then finding that B is true makes A more credible (Polya [1954] called this an *induction* pattern [see Section 2.3.1]). This pattern involves reasoning both ways, from A to B and from B to A. Moreover, it appears that people do not require two separate rules for performing these inferences; the first rule (e.g., "Fire implies smoke") provides the license to invoke the second (e.g., "Smoke makes fire more credible"). Extensional systems, on the other hand, require that the second rule be stated explicitly and, even worse, that the first rule be removed. Otherwise, a cycle would be created where any slight evidence in favor of A would be amplified via B and fed back to A, quickly turning into a stronger confirmation (of A and B), with no apparent factual justification. The prevailing practice in such systems (e.g., MYCIN) is to cut off cycles of that sort, permitting only diagnostic reasoning and no predictive inferences.

Removal of its predictive component prevents the system from exhibiting another important pattern of plausible reasoning, one that we call *explaining away*: If A implies B, C implies B, and B is true, then finding that C is true makes A *less* credible. In other words, finding a second explanation for an item of data makes the first explanation less credible. Such interaction among multiple causes appears in many applications (see Sections 2.2.4, 2.3.1, 4.3.2, and 10.2). For example, finding that the smoke could have been produced by a bad muffler makes fire less credible. Finding that my light bulb emits red light makes it less credible that the red-hued object in my hand is truly red.

To exhibit this sort of reasoning, a system must use bidirected inferences: from evidence to hypothesis (or explanation) and from hypothesis to evidence. While it is sometimes possible to use brute force (e.g., enumerating all exceptions) to restore "explaining away" without the danger of circular reasoning, we shall see that any system that succeeds in doing this must sacrifice the principles of modularity, i.e., locality and detachment. More precisely, every system that updates beliefs modularly at the natural rule level and that treats all rules equally is bound to defy prevailing patterns of plausible reasoning.

THE LIMITS OF MODULARITY

The principle of locality is fully realized in the inference rules of classical logic. The rule "If P then Q" means that if P is found true, we can assert Q with no further analysis, even if the database contains some other knowledge K. In plausible reasoning, however, the luxury of ignoring the rest of the database cannot be maintained. For example, suppose we have a rule R_1 = "If the ground is wet, then assume it rained (with certainty c_1)." Validating the truth of "The ground is wet" does not permit us to increase the certainty of "It rained" because the knowledge base might contain strange items such as K = "The sprinkler was on last night." These strange items, called *defeaters* or *suppressors* (Section 10.3), are sometimes easy to discover (as with K' = "The neighbor's grass is dry," which

directly opposes "It rained"), but sometimes they hide cleverly behind syntactical innocence. The neutral fact K = "Sprinkler was on" neither supports nor opposes the possibility of rain, yet K manages to undercut the rule R_1. This undercutting cannot be implemented in an extensional system; once R_1 is invoked, the increase in the certainty of "It rained" will never be retracted, because no rule would normally connect "Sprinkler was on" to "It rained." Imposing such a connection by proclaiming "Sprinkler was on" as an explicit exception to R_1 defeats the spirit of modularity by forcing the rule-writer to pack together items of information that are only remotely related to each other, and it burdens the rules with an unmanageably large number of exceptions.

Violation of detachment can also be demonstrated in this example. In deductive logic, if K implies P and P implies Q, then finding K true permits us to deduce Q by simple chaining; a derived proposition (P) can trigger a rule ($P \rightarrow Q$) with the same vigor as a directly observed proposition can. Chaining does not apply in plausible reasoning. The system may contain two innocent-looking rules—"If the ground is wet then it rained" and "If the sprinkler was on then the ground is wet"—but if you find that the sprinkler was on, you obviously do not wish to conclude that it rained. On the contrary, finding that the sprinkler was on only takes away support from "It rained."

As another example, consider the relationships shown in Figure 1.2. Normally an alarm sound alerts us to the possibility of a burglary. If somebody calls you at the office and tells you that your alarm went off, you will surely rush home in a hurry, even though there could be other causes for the alarm sound. If you hear a radio announcement that there was an earthquake nearby, and if the last false alarm you recall was triggered by an earthquake, then your certainty of a burglary will diminish. Again, this requires going both ways, from effect to cause (*Radio* $\rightarrow$ *Earthquake*), and from cause to effect (*Earthquake* $\rightarrow$ *Alarm*), and then from effect to cause again (*Alarm* $\rightarrow$ *Burglary*). Notice what pattern of reasoning results from such a chain, though: We have a rule, "If A (*Alarm*) then B (*Burglary*)"; you listen to the radio, A becomes more credible, and the conclusion B becomes less credible. Overall, we have "If $A \rightarrow B$ and A becomes more credible, then B becomes less credible." This behavior is clearly contrary to everything we expect from local belief updating.

In conclusion, we see that the difficulties plaguing classical logic do not stem from its nonnumeric, binary character. Equally troublesome difficulties emerge when truth and certainty are measured on a grey scale, whether by point values, by interval bounds, or by linguistic quantifiers such as "likely" and "credible." There seems to be a basic struggle between procedural modularity and semantic coherence, independent of the notation used.

CORRELATED EVIDENCE

Extensional systems, greedily exploiting the licenses provided by locality and detachment, respond only to the magnitudes of the weights and not to their origins.

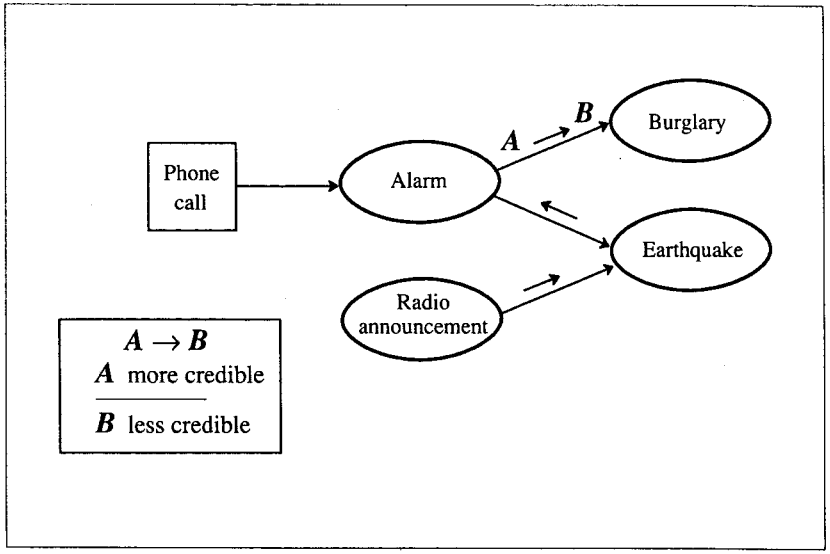

Figure 1.2. *Making the antecedent of a rule more credible can cause the consequent to become less credible.*

As a result they will produce the same conclusions whether the weights originate from identical or independent sources of information. An example from Henrion [1986b] about the Chernobyl disaster helps demonstrate the problems encountered by such a local strategy. Figure 1.3 shows how multiple, independent sources of evidence would normally increase the credibility of a hypothesis (e.g., *Thousands dead*), but the discovery that these sources have a common origin should reduce the credibility. Extensional systems are too local to recognize the common origin of the information, and they would update the credibility of the hypothesis as if it were supported by three independent sources.

1.2.3 Attempted Remedies and their Limitations

The developers of extensional systems have proposed and implemented powerful techniques to remedy some of the semantic deficiencies we have discussed. The remedies, most of which focus on the issue of correlated evidence, take two approaches:

1. *Bounds propagation*: Since most correlations are unknown, certainty measures are combined under two extreme assumptions—that the components have a high positive correlation, and that they have a high

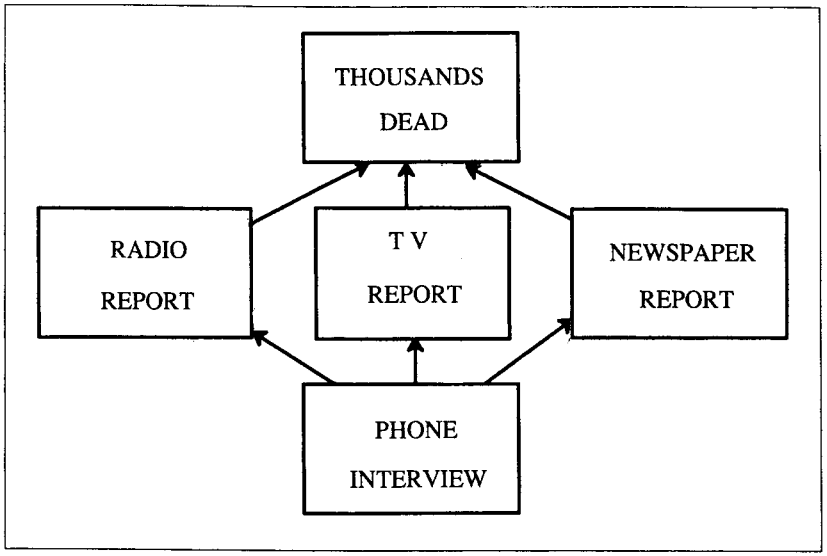

Figure 1.3. *The Chernobyl disaster example (after Henrion) shows why rules cannot combine locally.*

negative correlation. This yields upper and lower bounds on the combined certainty, which are entered as inputs to subsequent computations, producing new bounds on the certainty of the conclusions. This approach has been implemented in INFERNO [Quinlan 1983] and represents a local approximation to Nilsson's probabilistic logic [Nilsson 1986] (see Section 9.3).

2. *User-specified combination functions*: A system named RUM [Bonissone, Gans, and Decker 1987] permits the rule-writer to specify the combination function that should apply to the rule's components. For example, if a, b, and c stand for the weights assigned to propositions A, B, and C in the rule

$$A \wedge B \rightarrow C\,,$$

the user can specify which of the following three combination functions should be used:

$$T_1(a, b) = \max(0, a + b - 1),$$

$$T_2(a, b) = ab,$$

or

$$T_3(a, b) = \min(a, b).$$

These functions (called *T norms*) represent the probabilistic combinations obtained under three extreme cases of correlation between A and B: highly negative, zero, and highly positive.

Cohen, Shafer, and Shenoy [1987] have proposed a more refined scheme, where for any pair of values $P(A)$ and $P(B)$, the user is permitted to specify the value of the resulting probability, $P(C)$.

The difficulties with these correlation-handling techniques are several. First, the bounds produced by systems such as INFERNO are too wide. For example, if we are given $P(A) = p$ and $P(B \mid A) = q$, then the bounds we obtain for $P(B)$ are

$$pq \leq P(B) \leq 1 - p(1 - q)\,,$$

which for small p approach the unit interval [0, 1]. Second, to handle the intricate dependencies that may occur among rules it is not enough to capture pair-wise correlations; higher-order dependencies are often necessary [Bundy 1985]. Finally, even if one succeeds in specifying higher-order dependencies, a much more fundamental limitation exists: Dependencies are dynamic relationships, created and destroyed as new evidence is obtained. For example, dependency between the propositions "It rained last night" and "The sprinkler was on" is created once we find out that the ground is wet. The dependence between a child's shoe size and reading ability is destroyed once we find out the child's age. Thus, correlations and combination functions specified at the knowledge-building phase may quickly become obsolete once the program is put into use.

Heckerman [1986a, 1986b] delineated precisely the range of applicability of extensional systems. He proved that any system that updates certainty weights in a modular and consistent fashion can be given a probabilistic interpretation in which the certainty update of a proposition A is some function of the likelihood ratio

$$\lambda = \frac{P(Evidence \mid A)}{P(Evidence \mid \neg A)}.$$

In MYCIN, for example, the certainty factor CF can be interpreted as

$$CF = \frac{\lambda - 1}{\lambda + 1}\,.$$

Once we have a probabilistic interpretation, it is easy to determine the set of structures within which the update procedure will be semantically valid. It turns out that a system of such rules will produce coherent updates if and only if the rules form a directed tree, i.e., no two rules may stem from the same premise. This limitation explains why strange results were obtained in the burglary example of Figure 1.2. There, the alarm event points to two possible explanations, *Burglary* and *Earthquake*, which amounts to two evidential rules stemming from the premise, *Alarm*.

Hájek [1985] and Hájek and Valdes [1987] have developed an algebraic theory that characterizes an even wider range of the extensional systems and combining functions, including those based on Dempster-Shafer intervals. The unifying properties common to all such systems is that they form an ordered Abelian group. Again, the knowledge base must form a tree so that no evidence is counted twice via alternative paths of reasoning.

1.3 INTENSIONAL SYSTEMS AND NETWORK REPRESENTATIONS

We have seen that handling uncertainties is a rather tricky enterprise. It requires a fine balance between our desire to use the computational permissiveness of extensional systems and our ability to refrain from committing semantic sins. It is like crossing a minefield on a wild horse. You can choose a horse with good instincts, attach certainty weights to it and hope it will keep you out of trouble, but the danger is real, and highly skilled knowledge engineers are needed to prevent the fast ride from becoming a disaster. The other extreme is to work your way by foot with a semantically safe intensional system, such as probability theory, but then you can hardly move, since every step seems to require that you examine the entire field afresh. We shall now examine means for making this movement brisker.

In intensional systems, the syntax consists of declarative statements about states of affairs and hence mirrors world knowledge rather nicely. For example, conditional probability statements such as "Most birds fly" are both empirically testable and conceptually meaningful. Additionally, intensional systems have no problem handling bidirected inferences and correlated evidence; these emerge as built-in features of one globally coherent model (see Chapters 2 and 4). However, since the syntax does not point to any useful procedures, we need to construct special mechanisms that convert the declarative input into routines that answer queries. Such a mechanism is offered by techniques based on *belief networks*, which will be a central topic of discussion in this book.

1.3.1 Why Networks?

Our goal is to make intensional systems operational by making relevance relationships explicit, thus curing the impotence of declarative statements such as $P(B \mid A) = p$. As mentioned earlier, the reason one cannot act on the basis of such declarations is that one must first make sure that other items in the knowledge base are irrelevant to B and hence can be ignored. The trick, therefore, is to encode knowledge in such a way that the ignorable is recognizable, or better yet, that the unignorable is quickly identified and is readily accessible. Belief networks encode relevancies as neighboring nodes in a graph, thus ensuring that by consulting the neighborhood one gains a license to act; what you don't see locally doesn't matter. In effect, what network representations offer is a dynamically updated list of all currently valid licenses to ignore, and licenses to ignore constitute permissions to act.

Network representations are not foreign to AI systems. Most reasoning systems encode relevancies using intricate systems of pointers, i.e., networks of indices that group facts into structures, such as frames, scripts, causal chains, and inheritance hierarchies. These structures, though shunned by pure logicians, have proved to be indispensable in practice, because they place the information required to perform an inference task close to the propositions involved in the task. Indeed, many patterns of human reasoning can be explained only by people's tendency to follow the pathways laid out by such networks.

The special feature of the networks discussed in this book is that they have clear semantics. In other words, they are not auxiliary devices contrived to make reasoning more efficient but are an integral part of the semantics of the knowledge base, and most of their features can even be derived from the knowledge base.

Belief networks play a central role in two uncertainty formalisms: probability theory, where they are called *Bayesian networks, causal nets,* or *influence diagrams,* and the Dempster-Shafer theory (see Chapter 9), where they are referred to as *galleries* [Lowrance, Garvey, and Strat 1986], *qualitative Markov networks* [Shafer, Shenoy, and Mellouli 1988], or *constraint networks* [Montanari 1974]. Probabilistic networks will be given a formal treatment in Chapter 3 and will serve as a unifying theme throughout this book. In the next subsection we briefly discuss the theory of graphoids, which provides formal semantics for graphical representations in terms of information relevance.

1.3.2 Graphoids and the Formalization of Relevance and Causality

A central requirement for managing intensional systems is to articulate the conditions under which one item of information is considered relevant to another, given what we already know, and to encode knowledge in structures that display these conditions vividly as the knowledge undergoes changes. Different formalisms give rise to different definitions of relevance. For example, in probability theory, relevance is identified with dependence; in database theory, with induced constraints—two variables are said to be relevant to each other if we can restrict the range of values permitted for one by constraining the other.

The essence of relevance can be identified with a structure common to all of these formalisms. It consists of four axioms which convey the simple idea that when we learn an irrelevant fact, the relevance relationships of all other

propositions remain unaltered; any information that was irrelevant remains irrelevant, and that which was relevant remains relevant. Structures that conform to these axioms are called *graphoids* [Pearl and Paz 1985] and will be treated more fully in Chapter 3. Interestingly, both undirected graphs and directed acyclic graphs conform to the graphoids axioms (hence the name) if we associate the sentence "Variable X is irrelevant to variable Y once we know Z" with the graphical condition "Every path from X to Y is intercepted by the set of nodes corresponding to Z." (A special definition of intercept is required for directed graphs [see Section 3.3.1]).

With this perspective in mind, graphs, networks, and diagrams can be viewed as inference engines devised for efficiently representing and manipulating relevance relationships. The topology of the network is assembled from a list of local relevance statements (e.g., direct dependencies). This input list implies (using the graphoid axioms) a host of additional statements, and the graph ensures that a substantial portion of the latter can be verified by simple graphical procedures such as path tracing and path blocking. Such procedures enable one to determine, at any state of knowledge Z, what information is relevant to the task at hand and what can be ignored. Permission to ignore, as we saw in Section 1.1, is the fuel that gives intensional systems the power to act.

The theory of graphoids shows that a belief network can constitute a sound and complete inference mechanism relative to probabilistic dependencies, i.e., it identifies, in polynomial time, every conditional independence relationship that logically follows from those used in the construction of the network (see Section 3.3). Similar results hold for other types of relevance relationships, e.g., partial correlations and constraint-based dependencies. The essential requirement for soundness and completeness is that the network be constructed *causally,* i.e., that we identify the most relevant predecessors of each variable recursively, in some total order, say temporal. (Once the network is constructed, the original order can be forgotten; only the partial order displayed in the network matters.)

It is this soundness and completeness that gives causality such a central role in this book, and perhaps in knowledge organization in general. However, the precise relationship between causality as a representation of irrelevancies and causality as a commitment to a particular inference strategy (e.g., chronological ignorance [Shoham 1986]) has yet to be fully investigated.

1.4 THE CASE FOR PROBABILITIES

The aim of artificial intelligence is to provide a computational model of intelligent behavior, most importantly, commonsense reasoning. The aim of probability theory is to provide a coherent account of how belief should change in light of partial or uncertain information. Since commonsense reasoning always applies to incomplete information, one might naturally expect the two disciplines to share

language, goals, and techniques. However, ever since McCarthy and Hayes [1969] proclaimed probabilities to be "epistemologically inadequate," AI researchers have shunned probability adamantly. Their attitude has been expressed through commonly heard statements like "The use of probability requires a massive amount of data," "The use of probability requires the enumeration of all possibilities," and "People are bad probability estimators." "We do not have those numbers," it is often claimed, and even if we do, "We find their use inconvenient."

Aside from the obvious corrections to these claims, this book will try to communicate the idea that "probability is not really about numbers; it is about the structure of reasoning," as Glenn Shafer recently wrote.† We will emphasize, for example, that when a physician asserts, "The chances that a patient with disease D will develop symptom S is p," the thrust of the assertion is not the precise magnitude of p so much as the specific reason for the physician's belief, the context or assumptions under which the belief should be firmly held, and the sources of information that would cause this belief to change. We will also stress that probability theory is unique in its ability to process context-sensitive beliefs, and what makes the processing computationally feasible is that the information needed for specifying context dependencies can be represented by graphs and manipulated by local propagation.

1.4.1 Why Should Beliefs Combine Like Frequencies?

On the surface, there is really no compelling reason that beliefs, being mental dispositions about unrepeatable and often unobservable events, should combine by the laws of proportions that govern repeatable trials such as the outcomes of gambling devices. The primary appeal of probability theory is its ability to express useful *qualitative* relationships among beliefs and to process these relationships in a way that yields intuitively plausible conclusions, at least in cases where intuitive judgments are compelling. A summary of such qualitative relationships will be given in the next subsection. What we wish to stress here is that the fortunate match between human intuition and the laws of proportions is not a coincidence. It came about because beliefs are formed not in a vacuum but rather as a distillation of sensory experiences. For reasons of storage economy and generality we forget the actual experiences and retain their mental impressions in the forms of averages, weights, or (more vividly) abstract qualitative relationships that help us determine future actions. The organization of knowledge and beliefs must strike a delicate balance between the computational resources these relationships consume and the frequency of their use. With these considerations in mind, it is

† Personal communication.

hard to envision how a calculus of beliefs can evolve that is substantially different from the calculus of proportions and frequencies, namely probability.

1.4.2 The Primitive Relationships of Probability Language

Although probabilities are expressed in numbers, the merit of probability calculus rests in providing a means for articulating and manipulating qualitative relationships that are found useful in normal discourse. The following four relationships are viewed as the basic primitives of the language:

1. Likelihood ("Tim is *more likely* to fly than to walk").
2. Conditioning ("*If* Tim is sick, he can't fly").
3. Relevance ("Whether Tim flies *depends on whether* he is sick").
4. Causation ("Being sick *caused* Tim's inability to fly").

LIKELIHOOD

The qualitative relationship of the form "A is more likely than B" has traditionally been perceived as the prime purpose of using probabilities. The practical importance of determining whether one event is more likely than another is best represented by the fact that probability calculus was pioneered and developed by such ardent gamblers as Cardano (1501-1576) and De Moivre (1667-1754). However, the importance of likelihood relationships goes beyond gambling situations or even management decisions. Decisions depending on relative likelihood of events are important in every reasoning task because likelihood translates immediately to *processing time*—the time it takes to verify the truth of a proposition, to consider the consequence of a rule, or to acquire more information. A reasoning system unguided by likelihood considerations (my ex-lawyer is a perfect example of one) would waste precious resources in chasing the unlikely while neglecting the likely.

Philosophers and decision theorists have labored to obtain an axiomatic basis for probability theory based solely on this primitive relationship of "more likely," namely, to identify conditions under which an ordering of events has a numerical representation P that satisfies the properties of probability functions [Krantz et al. 1971; Fine 1973; Fishburn 1986]. More recently, the task of devising a nonnumeric logic for manipulating sentences that contain the qualifier likely has received considerable attention [Halpern and Rabin 1987; Fagin and Halpern 1988] and has turned out to be a tougher challenge than expected.

CONDITIONING

Probability theory adopts the autoepistemic phrase "...given that what I know is C" as a primitive of the language. Syntactically, this is denoted by placing C behind the conditioning bar in a statement such as $P(A \mid C) = p$. This statement combines the notions of knowledge and belief by attributing to A a degree of belief p, given the knowledge C. C is also called the *context* of the belief in A, and the notation $P(A \mid C)$ is called *Bayes conditionalization*. Thomas Bayes (1702–1761) made his main contribution to the science of probability by associating the English phrase "...given that I know C" with the now-famous ratio formula

$$P(A \mid C) = \frac{P(A, C)}{P(C)} \tag{1.1}$$

[Bayes 1763], which has become a definition of conditional probabilities (see Eq. (2.8)).

It is by virtue of Bayes conditionalization that probability theory facilitates nonmonotonic reasoning, i.e., reasoning involving retraction of previous conclusions (see Section 1.5). For example, it is perfectly acceptable to assert simultaneously $P(Fly(a) \mid Bird(a)) = HIGH$ and $P(Fly(a) \mid Bird(a), Sick(a)) = LOW$. In other words, if all we know about individual a is that a is a bird, we jump to the conclusion that a most likely flies. However, upon learning that a is also sick, we retract our old conclusion and assert that a most likely cannot fly.

To facilitate such retraction it is necessary both that the original belief be stated with less than absolute certainty and that the context upon which we condition beliefs be consulted constantly to see whether belief revision is warranted. The dynamic of belief revision under changing contexts is not totally arbitrary but must obey some basic laws of plausibility which, fortunately, are embedded in the syntactical rules of probability calculus. A typical example of such a plausibility law is the rule of the *hypothetical middle:*

> If two diametrically opposed assumptions impart two different degrees of belief onto a proposition Q, then the unconditional degree of belief merited by Q should be somewhere between the two.

For example, our belief that Tim flies given that Tim is a bird must be between our belief that Tim flies given that he is a sick bird and our belief that Tim flies given that he is a healthy bird. Such a qualitative, commonsense restriction is built into the syntax of probability calculus via the equality

$$P(B \mid C) = \alpha P(B \mid C, A) + (1 - \alpha) P(B \mid C, \neg A), \tag{1.2}$$

where $\alpha = P(A \mid C)$ is some number between 0 and 1. Other typical patterns of plausible reasoning are those of abduction and "explaining away," mentioned in Section 1.2.2 and further elaborated in Section 2.3.1.

RELEVANCE

Relevance is a relationship indicating a potential change of belief due to a specified change in knowledge (see Section 1.3.2). Two propositions A and B are said to be relevant to each other in context C if adding B to C would change the likelihood of A. Clearly, relevance can be defined in terms of likelihood and conditioning, but it is a notion more basic than likelihood. For example, a person might be hesitant to assess the likelihood of two events but feel confident about judging whether or not the events are relevant to each other. People provide such judgments swiftly and consistently because—we speculate—relevance relationships are stored explicitly as pointers in one's knowledge base.

Relevance is also a primitive of the language of probability because the language permits us to specify relevance relationships directly and qualitatively before making any numerical assessment. Later on, when numerical assessments of likelihood are required, they can be added in a consistent fashion, without disturbing the original relevance structure (see Chapter 3).

CAUSATION

Causation is a ubiquitous notion in man's conception of his environment, yet it has traditionally been considered a psychological construct, outside the province of probability or even the physical sciences [Russell 1913]. In Section 3.3 we present a new account of causation, according to which it can be given a nontemporal probabilistic interpretation based solely on the notion of relevance. The temporal component of causation [Suppes 1970; Shoham 1988] is viewed merely as a convenient indexing standard chosen to facilitate communication and predictions.

Causation is listed as one of the four basic primitives of the language of probability because it is an indispensable tool for structuring and specifying probabilistic knowledge (see Sections 3.3 and 10.4) and because the semantics of causal relationships are preserved by the syntax of probabilistic manipulations; no auxiliary devices are needed to force conclusions to conform with people's conception of causation. The following is a brief summary of our notion of causation, to be further developed in Sections 3.3, 8.2, and 10.3.

Causation is a language with which one can talk efficiently about certain structures of relevance relationships, with the objective of separating the relevant from the superfluous. For example, to say that a wet pavement was a direct cause of my slipping and breaking a leg is a concise way of identifying which events should no longer be considered relevant to my accident, once the wetness of the pavement is confirmed. The facts that it rained that day, that the rain was welcomed by farmers, and that my friend also slipped and broke his leg should no longer be considered relevant to the accident once we establish the truth of *Wet pavement* and identify it as the direct cause of the accident.

The asymmetry conveyed by causal directionality is viewed as a notational device for encoding still more intricate patterns of relevance relationships, such as

nontransitive and *induced* dependencies. For example, by designating *Rain* and *Sprinkler* as potential causes of the wet pavement we permit the two causes to be independent of each other and still both be relevant to *Wet pavement* (hence forming a nontransitive relationship). Moreover, by this designation we also identify the consequences *Wet pavement* and *Accident* as potential sources of new dependencies between the two causes; once a consequence is observed, its causes can no longer remain independent, because confirming one cause lowers the likelihood of the other. This connection between nontransitive and induced dependencies is, again, a built-in feature of the syntax of probability theory—the syntax ensures that nontransitive dependencies always induce the appropriate dependencies between causes (see Exercise 3.10).

To summarize, causal directionality conveys the following pattern of dependency: Two events do not become relevant to each other merely by virtue of predicting a common consequence, but they do become relevant when the consequence is actually observed. The opposite is true for two consequences of a common cause; typically the two become independent upon learning the cause. (Chapter 8 deals with using this asymmetry to identify causal directionality in nontemporal empirical data.)

1.4.3 Probability as a Faithful Guardian of Common Sense

In the preceding subsections we presented qualitative patterns of commonsense reasoning that are naturally embedded within the syntax of probability calculus. Among these intuitive patterns are nonmonotonicity (context sensitivity), abduction, "explaining away," causation, and hypothetical middle. It is possible to assemble some of these desirable patterns of inference and pose them as axioms that render probability calculus "inevitable," i.e., to show that any calculus respecting these desired patterns behaves as if it were driven by a probability engine. This route was a favorite preoccupation of many philosophers, most notably Ramsey [1931], de Finetti [1937], Cox [1946], Good [1950], and Savage [1954]. Cox assembled seven semi-qualitative arguments for the conditional relation $(A \mid B)$ (to read, "The plausibility of A conditioned on the evidence B") and showed that they lead to Bayes' ratio formula (Eq. (1.1)) and thus to probability calculus. This axiomatic approach placed probability on firm qualitative ground, but it has also been the subject of lively debates and refutations (e.g., Savage [1962], Lindley [1982], and Shafer [1986a]). When posed as a stand-alone axiomatic system, any chosen subset of reasoning patterns is vulnerable to criticism because we can always imagine a situation where one of the axioms ceases to be necessary, thus discrediting the entire system. The interested reader is referred to the classical literature on the foundations of probability [Fine 1973; Krantz et al. 1971].

The approach taken in this book is somewhat different. We take for granted that probability calculus is unique in the way it handles context-dependent information and that no competing calculus exists that closely covers so many qualitative aspects of plausible reasoning. So the calculus is worthy of exploitation, emulation, or at the very least, serious exploration. We therefore take probability calculus as an initial model of human reasoning from which more refined models may originate, if needed. By exploring the limits of using probability calculus in machine implementations of plausible inference, we hope to identify conditions under which extensions, refinements, and simplifications are warranted.

Obviously, there are applications where strict adherence to the dictates of probability theory would be computationally infeasible, and there compromises will have to be made. Still, we find it more comfortable to compromise an ideal theory that is well understood than to search for a new surrogate theory, with only gut feeling for guidance.

The merits of a theory-based approach are threefold:

1. The theory can be consulted to ensure that compromises are made only when necessary and that their damage is kept to a minimum.
2. When system performance does not match expectations, knowing which compromises were made helps identify the adjustments needed.
3. Compromised theories facilitate scientific communication; one need specify only the compromises made, treating the rest of the theory as common knowledge.

HOW BAD ARE THOSE NUMBERS?

People are notoriously bad numerical estimators. They find it hard to assess absolute probabilities as well as distances, weights, and times. A person would much rather assert qualitatively that one object is heavier than another than assess the absolute weight of a given object. Still, the lack of an accurate scale does not preclude the use of the laws of physics when it comes to deciding which bag is lighter, the one containing 2000 dimes or the one containing 1000 quarters. It is quite conceivable that a person has never before seen bags containing thousands of coins, yet the limited experience gathered from handling small quantities of coins, teaching us that two dimes are lighter than one quarter, can be amplified by the laws of physics and extended to situations never seen before. We might assign a single dime a rough weight estimate of 10 grams, consult our experience and assign a quarter an estimate of 30 grams, then multiply the two estimates by the respective numbers of coins and compare the results. The absolute estimates in this example can be completely off, but as long as their ratio reflects genuine experience, the conclusions will still be useful. (Deriving these conclusions

symbolically, using axioms to describe how weights combine, often requires much more work.) In other words, if we strongly believe in the rules by which exact quantities combine, we can use the same combination rules on the rough estimates at hand.

This heuristic strategy gives reasonably good results for several reasons. First, by using reliable combination rules, we make the utmost use of the available knowledge and keep the damage due to imprecision from extending beyond well-defined boundaries. Second, when we commit ourselves to a particular set of numbers, no matter how erroneous, the consistency of the model prevents us from reaching inconsistent conclusions. For example, we will never reach a conclusion that the 2000-dime bag is lighter than the 1000-quarter bag and a simultaneous conclusion that 3000 dimes are heavier than 1500 quarters. Finally, and most importantly for dealing with uncertainty in AI systems, adhering to a coherent model of reality helps us debug our inferences when they do not match expectations. In our coin example, if it turns out, contrary to calculations, that the 2000 dimes are not lighter than the 1000 quarters, we know immediately that we have either wrongly estimated the relative weights of a dime and a quarter or miscounted the coins in the bags; we need not tamper with the rules of inference or with their calculus of combination. In general, we know precisely how the model should be refined or improved.

ON THE USEFULNESS OF NUMBERS

If people prefer to reason qualitatively, why should machines reason with numbers? Probabilities are summaries of knowledge that is left behind when information is transferred to a higher level of abstraction. The summaries can be encoded logically or numerically; logic enjoys the advantages of parsimony and simplicity, while numbers are more informative and sometimes are necessary.

The minefield metaphor used in Section 1.1 will help illustrate the usefulness of numerical summarization. Imagine that before we start our journey across the minefield, we are given access to a complete record of the field, specifying in full detail the exact location of each mine as recorded six months earlier by the team that laid these mines. However, since we cannot carry with us the entire record, we must somehow summarize that information on a miniature map, the size of a postcard. There are many ways we might summarize the data on the postcard, but one of the most effective methods is to color the map to reflect the density of mines in any given area: the darker the color, the higher the density. Viewing dark colors as high numbers, this is the essence of numerical summarization of uncertainty. Why is this scheme effective?

Imagine that you start your journey by pursuing what appears to be a rather safe path to your destination. After two days you reach a roadblock; the path chosen is not usable and an alternative path must be found. Here is where the color code begins to show its usefulness. While traversing the original path you

passed many side roads branching out from the one you chose. At the time, these junctions were abandoned because your path appeared more promising, but now that your first choice turned into a disappointment, you must look back at those branching points and decide which one to pursue next. Had you summarized your decisions using a bi-valued predicate, say "possible" or "not possible," you would now be at a loss. Among those marked "possible," you would not know which one is actually the least dangerous and the quickest, especially in light of the new roadblocks you have discovered. The colored map provides exactly this information.

To make the analogy closer to mental reasoning tasks, let us further imagine that we can communicate with headquarters and ask them to wire us a more detailed map of any region under consideration. The question is which map we should request. In the absence of priority ranking among the viable alternatives, precious time will be wasted transmitting and examining maps that, in view of the new road conditions discovered, will again lead to dead ends. The function of colored maps, and of numeric labels in general, is to prioritize the flow of information and focus on items more likely to yield beneficial results.

The translation to reasoning tasks is obvious. Raw experiential data is not amenable to reasoning activities such as prediction and planning; these require that data be abstracted into a representation with a coarser grain. Probabilities are summaries of details lost in this abstraction, similar in role to the colors on our maps. The importance of maintaining such summaries in AI systems can be appreciated in the context of planning systems, where a major obstacle has been the impracticability of enumerating all preconditions that might trigger, inhibit, or enable a given event. (This problem is known as the *qualification* problem [McCarthy 1980], a refinement of the infamous *frame* problem [McCarthy and Hayes 1969; Brown 1987]). Probabilistic formalisms enable us to *summarize* the presumed existence of exceptional conditions without explicating the details of their interactions unless the need arises. Probability does not offer a complete solution to the frame problem because it does not provide rules for recomputing the summaries when unanticipated refinements are warranted. It does, however, provide a way to express summaries of unexplicated information, procedures for manipulating these summaries, and criteria for deciding when additional chunks of knowledge warrant explication.

To show what is still needed, let us examine how an ideal system might reason about the burglar alarm situation of Figure 1.2. Upon receiving the phone call from your neighbor, only the burglary hypothesis is triggered; your decision whether to drive home or stay at work is made solely on the basis of the parameter $P(\textit{False alarm})$, which summarizes all other (unexplicated) causes for an alarm sound. After a moment's reflection, the possibility of an April Fools' Day joke may enter your mind, in which case a two-stage inference chain is assembled, governed by two probabilistic parameters, $P(\textit{False alarm})$ and $P(\textit{Prank call})$. Later, when the possibility of an earthquake enters consideration, the parameter

$P(\textit{False alarm})$ undergoes a partial explication; a fragment of knowledge is brought over from the remote frame of earthquake experiences and is appended to the link $\textit{Burglary} \rightarrow \textit{Alarm}$ as an alternative cause or explanation. The catchall hypothesis *All other causes* shrinks (to exclude earthquakes), and its parameters are readjusted. The radio announcement strengthens your suspicion in the earthquake hypothesis and permits you to properly readjust your decisions without elaborating the mechanics of the pressure transducer used in the alarm system. The remote possibility of having forgotten to push the reset button will be invoked only if it is absolutely needed for explaining some observed or derived phenomenon, e.g., finding your home burglarized and your alarm system silent.

Systems using probabilistic formalisms have so far drawn inferences from static knowledge bases, where the set of variables, their relationships, and all probabilistic parameters are provided by external agents, at predetermined levels of granularity. This is far from the reasoning pattern just portrayed by our burglary example, where relationships are explicated, refined, and quantified mechanically when the need arises. Clearly, what is lacking is the ability to transfer information back and forth between knowledge strata at different levels of abstraction, the ability to identify how information in one strata bears on information in another, and a means of properly adjusting the parameters of each item transferred.† Research toward the development of such facilities should bring together logic's aptitude for handling the visible and probability's ability to summarize the invisible.

1.5 QUALITATIVE REASONING WITH PROBABILITIES

In the preceding section we described some of the merits of using numerical representations in reasoning tasks. There are applications, however, where categorical abstractions may suffice and knowledge can be summarized by hard logical facts, merely distinguishing the possible from the impossible. For example, when the number of possibilities is small, instead of calculating which option is preferred we might settle for an indication of which option is still a candidate for exploration. In such cases we enter the province of logical analysis, and the problem becomes one of representing exceptions and reflecting nonmonotonic reasoning. The connection between probability theory and nonmonotonic logic will be expounded more fully in Chapter 10. Here we merely outline how probability theory, even stripped of all its numbers, can be useful as a paradigm facilitating purely qualitative reasoning.

† Variable precision logic [Michalski and Winston 1986] is an attempt to formulate this dynamics.

1.5.1 Softened Logic vs. Hardened Probabilities

The ills of classical logic have often been attributed to its rigid, binary character. Indeed, when one tries to explain why logic would not predict the obvious fact that penguin are birds but do not fly, the first thing that one tends to blame is logic's rigid stance toward exceptions to the rule "Birds fly." It is therefore natural to assume that once we soften the constraints of Boolean logic and allow truth values to be measured on a grey scale, these problems will disappear. There have been several such attempts. Rich [1983] proposed a likelihood-based interpretation of default rules, managed by certainty-factors calculus. Ginsberg [1984] and Baldwin [1987] have pursued similar aspirations using the Dempster-Shafer notion of belief functions (see Chapter 9). While these attempts can produce valuable results (revealing, for instance, how sensitive a conclusion is to the uncertainty of its premises), the fundamental problem of monotonicity remains unresolved. For example, regardless of the certainty calculus used, these analyses always yield an increase in the belief that penguins can fly if one adds the superfluous information that penguins are birds and birds normally fly. Identical problems surface in the use of incidence calculus and softened versions of truth-maintenance systems [Falkenhainer 1986; D'Ambrosio 1987].

Evidently, it is not enough to add a soft probabilistic veneer to a system that is built on hard monotonic logic. The problem with monotonic logic lies not in the hardness of its truth values, but rather in its inability to process context-dependent information. Logic does not have a device equivalent to the conditional probability statement "$P(B \mid A)$ is high," whose main function is to define the context A under which the proposition B can be believed and to make sure that the only context changes permitted are those that do not change the belief in B (e.g., going from A = *Birds* to A' = *Feathered birds*).

Lacking an appropriate logical device for conditionalization, the natural tendency is to interpret the English sentence "If A then B" as a softened version of the material implication constraint $A \supset B$. A useful consequence of such softening is the freedom from outright contradictions. For example, while the classical interpretation of the three rules "Penguins do not fly," "Penguins are birds," and "Birds fly" yields a blatant contradiction, attaching uncertainties to these rules renders them manageable. They are still managed in the wrong way, however, because the material-implication interpretation of if–then rules is so fundamentally wrong that its maladies cannot be rectified simply by allowing exceptions in the form of shaded truth values. The source of the problem lies in the property of transitivity, $(a \rightarrow b, b \rightarrow c) \Longrightarrow a \rightarrow c$, which is inherent to the material-implication interpretation. On some occasions rule transitivity must be totally suppressed, not merely weakened, or else strange results will surface. One such occasion occurs in property inheritance, where subclass specificity should override superclass properties. Another occurs in causal reasoning, where predictions should not trigger explanations (e.g., "Sprinkler was on" predicts "Ground is wet,"

"Ground is wet" suggests "It rained," yet "Sprinkler was on" should not suggest "It rained"). In such cases, softening the rules weakens the flow of inference through the rule chain but does not bring it to a dead halt, as it should.

Apparently what is needed is a new interpretation of if–then statements, one that does not destroy the context sensitivity of probabilistic conditionalization. McCarthy [1986] remarks that *circumscription*† indeed provides such an interpretation. In his words:

> Since circumscription doesn't provide numerical probabilities, its probabilistic interpretation involves probabilities that are either infinitesimal, within an infinitesimal of one, or intermediate—without any discrimination among the intermediate values. The circumscriptions give conditional probabilities. Thus we may treat the probability that a bird can't fly as an infinitesimal. However, if the rare event occurs that the bird is a penguin, then the conditional probability that it can fly is infinitesimal, but we may hear of some rare condition that would allow it to fly after all.

Rather than contriving new logics and hoping that they match the capabilities of probability theory, we can start with probability theory, and if we can't get the numbers or we find their use inconvenient, we can extract the infinitesimal approximation as an idealized abstraction of the theory, while preserving its context-dependent properties. In this way, a nonmonotonic logic should crystallize that is guaranteed to capture the context-dependent features of natural defaults.

1.5.2 Probabilities and the Logic of "Almost True"

This program was in fact initiated over twenty years ago by the philosopher Ernest Adams, who developed a logic of conditionals based on probabilistic semantics [Adams 1966]. The sentence "If A then B" is interpreted to mean that the conditional probability of B given A is very close to 1 but is short of actually being 1. An adaptation of Adams's logic to default schemata of the form $Bird(x) \rightarrow Fly(x)$, where x is a variable, is described in Section 10.2. The resulting logic is nonmonotonic relative to learning new facts, in accordance with McCarthy's desiderata. For example, learning that Tweety is a bird will yield the conclusion that Tweety can fly. Subsequently learning that Tweety is also a penguin will yield the opposite conclusion: Tweety can't fly. Further, learning that Tweety is black and white will not alter this belief, because black and white is a typical color combination for penguins. However, and this is where Adams's logic falls short of expectations, learning that Tweety is clever would force us to

† Circumscription is a system developed by McCarthy for nonmonotonic reasoning. With circumscription, the conclusions are sanctioned relative to the minimal models of the theory.

retract all previously held beliefs about Tweety's flying and answer, "I don't know." The logic is so conservative that it never jumps to conclusions that some new rule schemata might invalidate (just in case clever penguins *can* fly). In other words, the logic does not capture the usual convention that unless we are told otherwise, properties are presumed to be irrelevant to each other.

Attempts to enrich Adams's logic with relevance-based features are reported in Geffner and Pearl [1987b] and briefly described in Section 10.2.5. The idea is to follow a default strategy similar to that of belief networks (Section 3.1): Dependencies exist only if they are mentioned explicitly or if they follow logically from other explicit dependencies. However, whereas the stratified method of constructing belief networks ensures that all relevant dependencies were already encoded in the network, this can no longer be assumed in the case of partially specified models of isolated default rules. A new logic is needed to capture the conventions by which we proclaim properties to be irrelevant to each other.

There is another dimension along which probabilistic analysis can assist current research into nonmonotonic logics—the logics provide no criterion for testing whether a database comprising default rules is internally consistent. The prevailing attitude is that once we tolerate exceptions we might as well tolerate anything [Brachman 1985]. There is a sharp qualitative difference, however, between exceptions and outright contradictions. For example, the statement "Red penguins can fly" can be accepted as a description of a world in which redness defines an abnormal type of penguin, but the statements "Typically, birds fly" and "Typically, birds do not fly" stand in outright contradiction to each other, and because there is no world in which the two statements can hold simultaneously, they will inevitably lead to strange, inconsistent conclusions. While such obvious contradictions can easily be removed from the database [Touretzky 1986], more subtle ones might escape detection, e.g., "Birds fly," "Birds are feathered animals," "Feathered animals are birds," and "Feathered animals do not fly." Adams's logic provides a criterion for detecting such inconsistencies, in the form of three axioms that should never be violated. These axioms, and their implied graphical test for consistency, will be discussed in Sections 10.1 and 10.2.

1.6 BIBLIOGRAPHICAL AND HISTORICAL REMARKS

Broad surveys of uncertainty formalisms proposed for AI can be found in Prade [1983], Thompson [1985], Stephanou and Sage [1987], and the works collected in Kanal and Lemmer [1986] and Smets et al. [1988]. The February 1987 issue of *Statistical Science*, devoted to the calculus of uncertainty in artificial intelligence and expert systems, includes a lively debate between advocates of the Bayesian methods and advocates of the Dempster-Shafer approach. The February 1988 issue of *Computational Intelligence* offers a similar debate between advocates of the probabilistic and logicist schools in AI.

Systems—primarily expert systems—that provide practical solutions to various problems of reasoning with uncertainty include MYCIN [Shortliffe 1976], INTERNIST [Miller, Poole, and Myers 1982; Pople 1982], PROSPECTOR [Duda, Hart, and Nilsson 1976], MEDAS [Ben-Bassat et al. 1980], INFERNO [Quinlan 1983], RUM [Bonissone, Gans, and Decker 1987], MUM [Cohen et al. 1987], MDX [Chandrasekaran and Mittal 1983], and MUNIN [Andreassen et al. 1987]. Of these, only MEDAS and MUNIN would be classified as intensional systems; the rest are extensional (i.e., rule-based) systems. An in-depth study of rule-based systems, including the uncertainty management technique used in MYCIN, can be found in Buchanan and Shortliffe [1984] and the survey articles by Davis, Buchanan, and Shortliffe [1977] and Buchanan and Duda [1983]. Critical discussions of the use of probabilistic reasoning in medical decisions are given in Szolovits and Pauker [1978] and Pauker and Kassirer [1987].

Cox's [1946] argument for the use of probability theory has also been expounded by Reichenbach [1949] and restated in Horvitz, Heckerman, and Langlotz [1986] and Cheeseman [1988] for an AI audience. Heckerman [1986b] has generalized Cox's argument to measures of confirmation, i.e., the impact evidence has on the belief in a hypothesis. A stronger argument, based entirely on qualitative axioms, has been developed by Aleliunas [1988], who included the hypothetical-middle pattern (Section 1.4.2) as one of his axioms.

Arguments based on pragmatic considerations go back to Ramsey [1931] and de Finetti [1937]. These are often called "Dutch book" arguments, because they show that a gambler deviating from the rules of probability calculus will, in the long run, lose against an opponent who adheres to those rules. Lindley [1982] introduced a pragmatic argument based on the notion of a scoring rule, i.e., a payoff function that depends both on one's degree of belief in an event and on whether the event actually occurred (see Exercise 6.9). He showed that under rather general conditions, an agent can maximize his expected payoffs only by adopting the axioms of probability theory. Rebuttals to this argument are given in the discussion following Lindley's article.

Our treatment of MYCIN's certainty calculus (Figure 1.1) follows that of Heckerman [1986a]. A coherent treatment of bidirectional inferences in trees was given in Pearl [1982] and will be described in Section 4.2. The distinction between rebutting and undercutting defeaters (Section 1.2.2) was first made in Pollock [1974], and the example of an object observed in red light is his. A probabilistic model for such defeaters was proposed by Kim and Pearl [1983] and implemented in CONVINCE [Kim 1983; Kim and Pearl 1987] (see Section 4.3). A logic-based model was proposed in Pearl [1988b] and will be described in Section 10.3.

Bibliographical references for graphoids and nonmonotonic logic are in Chapters 3 and 10, respectively.

References to recent literature on various approaches to uncertainty in AI can be found in the following volumes:

Kanal L.N.; and Rosenfeld A. (series eds.). 1986-1991. Uncertainty in Artificial Intelligence 1-6 [1] Elsevier Science Publishers B.V. (North-Holland).

Shafer, G., and Pearl, J. (Eds.). 1990. *Readings in Uncertain Reasoning*, Morgan Kaufmann, Palo Alto, CA.

Neapolitan , R.E. 1990. *Probabilistic Reasoning in Expert Systems: Theory and Algorithms*, Wiley, New York.

Shachter, R., (ed.), Special Issue on Influence Diagrams, *Networks: an International Journal*, Vol. 20, No. 5, August 1990.

Oliver, R.M., and Smith, J.Q. (Eds.). 1990. *Influence Diagrams, Belief Nets and Decision Analysis*, Sussex, England: John Wiley & Sons, Ltd.

The following articles describe general uncertainty-management systems:

Andersen, S. K., et al. 1989. "HUGIN — A Shell for Building Bayesian Belief Universes for Expert Systems," *Proceedings, IJCAI-89*, 1080-1085.

Poole, D. "Representing Diagnostic Knowledge for Probabilistic Horn Abduction," *Proceedings IJCAI-91*, Sydney, Australia, August, 1991, 1129-1137.

Srinivas, S. and Breese, J., 1989. *IDEAL: Influence Diagram Evaluation and Analysis in Lispu*, Rockwell International Science Center, Palo Alto, CA.

Systems designed for specific applications include:

Heckerman, D.E., Horvitz, E.J., and Nathwany, B.N. 1990. "Toward normative expert systems: The Pathfinder project." Technical Report KSL-90-08, Medical Computer Science Group, Section on Medical Informatics, Stanford University, Stanford, CA. (diagnosis of pathological findings)

Peng, Y., and Reggia, J.A. 1990. *Abductive Inference Models for Diagnostic Problem-Solving*, Springer-Verlag, New York. (medical diagnosis)

Levitt, T.S., Agosta, J.M., and Binford, T.O. 1990. "Model-Based Influence Diagrams for Machine Vision," *UAI* 5, 371-388.

Charniak, E., and Goldman, R. 1991. "A Probabilistic Model of Plan Recognition, " *Proceedings, AAAI-91*, Anaheim, CA, 160-165. (story understanding)

Agogino, A.M., Srinivas, S. and Schneider, K. 1988. "Multiple sensor expert system for diagnostic reasoning, monitoring and control of mechanical systems, *Mechanical Systems and Signal Processing*, 2(2), 165-85.

Abramson, B. 1991. "ARCO1: An application of belief networks to the oil market," *Proceedings of the 1991 Conference on Uncertainty in AI*, Los Angeles, CA., Morgan Kaufmann, 1-8. (economic forcasting)

[1] In subsequent references, these volumes will be denoted *UAI*-1 through *UAI*-6.

Chapter 2

BAYESIAN INFERENCE

The purpose I mean is, to show what reason we have for believing that there are in the constitution of things fixed laws according to which events happen...

— Richard Price, 1763

(Introduction to Bayes' essay)

2.1 BASIC CONCEPTS

2.1.1 Probabilistic Formulation and Bayesian Inversion

Bayesian methods provide a formalism for reasoning about partial beliefs under conditions of uncertainty. In this formalism, propositions are given numerical parameters signifying the degree of belief accorded them under some body of knowledge, and the parameters are combined and manipulated according to the rules of probability theory. For example, if A stands for the statement "Ted Kennedy will seek the nomination for president in 1992," then $P(A \mid K)$ stands for a person's subjective belief in A given a body of knowledge K, which might include that person's assumptions about American politics, specific proclamations made by Kennedy, and an assessment of Kennedy's past and personality. In defining belief