1

Introduction

Automated reasoning has been receiving much interest from a number of fields, including philosophy, cognitive science, and computer science. In this chapter, we consider the particular interest of computer science in automated reasoning over the last few decades, and then focus our attention on probabilistic reasoning using Bayesian networks, which is the main subject of this book.

1.1 Automated reasoning

The interest in automated reasoning within computer science dates back to the very early days of artificial intelligence (AI), when much work had been initiated for developing computer programs for solving problems that require a high degree of intelligence. Indeed, an influential proposal for building automated reasoning systems was extended by John McCarthy shortly after the term "artificial intelligence" was coined [McCarthy, 1959]. This proposal, sketched in Figure 1.1, calls for a system with two components: a knowledge base, which encodes what we know about the world, and a reasoner (inference engine), which acts on the knowledge base to answer queries of interest. For example, the knowledge base may encode what we know about the theory of sets in mathematics, and the reasoner may be used to prove various theorems about this domain.

McCarthy's proposal was actually more specific than what is suggested by Figure 1.1, as he called for expressing the knowledge base using statements in a suitable logic, and for using logical deduction in realizing the reasoning engine; see Figure 1.2. McCarthy's proposal can then be viewed as having two distinct and orthogonal elements. The first is the separation between the knowledge base (what we know) and the reasoner (how we think). The knowledge base can be domain-specific, changing from one application to another, while the reasoner is quite general and fixed, allowing one to use it across different application areas. This aspect of the proposal became the basis for a class of reasoning systems known as *knowledge-based* or *model-based systems*, which have dominated the area of automated reasoning since then. The second element of McCarthy's early proposal is the specific commitment to logic as the language for expressing what we know about the world, and his commitment to logical deduction in realizing the reasoning process. This commitment, which was later revised by McCarthy, is orthogonal to the idea of separating the knowledge base from the reasoner. The latter idea remains meaningful and powerful even in the context of other forms of reasoning including probabilistic reasoning, to which this book is dedicated. We will indeed subscribe to this knowledge-based approach for reasoning, except that our knowledge bases will be Bayesian networks and our reasoning engine will be based on the laws of probability theory.

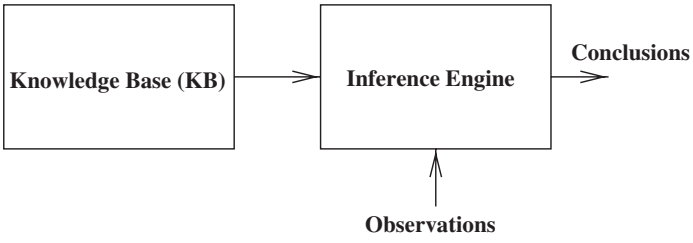


Figure 1.1: A reasoning system in which the knowledge base is separated from the reasoning process. The knowledge base is often called a "model," giving rise to the term "model-based reasoning."

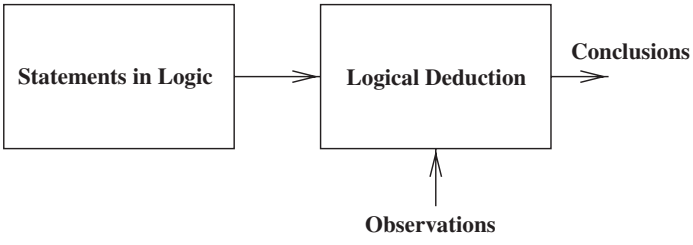


Figure 1.2: A reasoning system based on logic.

1.1.1 The limits of deduction

McCarthy's proposal generated much excitement and received much interest throughout the history of AI, due mostly to its modularity and mathematical elegance. Yet, as the approach was being applied to more application areas, a key difficulty was unveiled, calling for some alternative proposals. In particular, it was observed that although deductive logic is a natural framework for representing and reasoning about facts, it was not capable of dealing with assumptions that tend to be prevalent in commonsense reasoning.

To better explain this difference between facts and assumptions, consider the following statement:

If a bird is normal, it will fly.

Most people will believe that a bird would fly if they see one. However, this belief cannot be logically deduced from this fact, unless we further assume that the bird we just saw is normal. Most people will indeed make this assumption – even if they cannot confirm it – as long as they do not have evidence to the contrary. Hence, the belief in a flying bird is the result of a logical deduction applied to a mixture of facts and assumptions. For example, if it turns out that the bird, say, has a broken wing, the normality assumption will be retracted, leading us to also retract the belief in a flying bird.

This ability to dynamically assert and retract assumptions – depending on what is currently known – is quite typical in commonsense reasoning yet is outside the realm of deductive logic, as we shall see in Chapter 2. In fact, deductive logic is *monotonic* in the sense that once we deduce something from a knowledge base (the bird flies), we can never invalidate the deduction by acquiring more knowledge (the bird has a broken wing). The formal statement of monotonicity is as follows:

If Δ logically implies α, then Δ and Γ will also logically imply α.

Just think of a proof for α that is derived from a set of premises Δ. We can never invalidate this proof by including the additional premises Γ. Hence, no deductive logic is capable of producing the reasoning process described earlier with regard to flying birds.

We should stress here that the flying bird example is one instance of a more general phenomenon that underlies much of what goes on in commonsense reasoning. Consider for example the following statements:

My car is still parked where I left it this morning.
If I turn the key of my car, the engine will turn on.
If I start driving now, I will get home in thirty minutes.

None of these statements is factual, as each is qualified by a set of assumptions. Yet we tend to make these assumptions, use them to derive certain conclusions (e.g., I will arrive home in thirty minutes if I head out of the office now), and then use these conclusions to justify some of our decisions (I will head home now). Moreover, we stand ready to retract any of these assumptions if we observe something to the contrary (e.g., a major accident on the road home).

1.1.2 Assumptions to the rescue

The previous problem, which is known as the *qualification problem* in AI [McCarthy, 1977], was stated formally by McCarthy in the late 1970s, some twenty years after his initial proposal from 1958. The dilemma was simply this: If we write "Birds fly," then deductive logic would be able to infer the expected conclusion when it sees a bird. However, it would fall into an inconsistency if it encounters a bird that cannot fly. On the other hand, if we write "If a bird is normal, it flies," deductive logic will not be able to reach the expected conclusion upon seeing a bird, as it would not know whether the bird is normal or not – contrary to what most humans will do. The failure of deductive logic in treating this problem effectively led to a flurry of activities in AI, all focused on producing new formalisms aimed at counteracting this failure.

McCarthy's observations about the qualification problem were accompanied by another influential proposal, which called for equipping logic with an ability to jump into certain conclusions [McCarthy, 1977]. This proposal had the effect of installing the notion of assumption into the heart of logical formalisms, giving rise to a new generation of logics, *non-monotonic logics*, which are equipped with mechanisms for managing assumptions (i.e., allowing them to be dynamically asserted and retracted depending on what else is known). However, it is critical to note that what is needed here is not simply a mechanism for managing assumptions but also a criterion for deciding on which assumptions to assert and retract, and when. The initial criterion used by many non-monotonic logics was based on the notion of logical consistency, which calls for asserting as many assumptions as possible, as long as they do not lead to a logical inconsistency. This promising idea proved insufficient, however. To illustrate the underlying difficulties here, let us consider the following statements:

A typical Quaker is a pacifist.
A typical Republican is not a pacifist.

If we were told that Nixon is a Quaker, we could then conclude that he is a pacifist (by assuming he is a typical Quaker). On the other hand, if we were told that Nixon is

a Republican, we could conclude that he is not a pacifist (by assuming he is a typical Republican). But what if we were told that Nixon is both a Quaker and a Republican? The two assumptions would then clash with each other, and a decision would have to be made on which assumption to preserve (if either). What this example illustrates is that assumptions can compete against each other. In fact, resolving conflicts among assumptions turned out to be one of the difficult problems that any assumption-based formalism must address to capture commonsense reasoning satisfactorily.

To illustrate this last point, consider a student, Drew, who just finished the final exam for his physics class. Given his performance on this and previous tests, Drew came to the belief that he would receive an A in the class. A few days later, he logs into the university system only to find out that he has received a B instead. This clash between Drew's prior belief and the new information leads him to think as follows:

> Let me first check that I am looking at the grade of my physics class instead of some other class. Hmm! It is indeed physics. Is it possible the professor made a mistake in entering the grade? I don't think so . . . I have taken a few classes with him, and he has proven to be quite careful and thorough. Well, perhaps he did not grade my Question 3, as I wrote the answer on the back of the page in the middle of a big mess. I think I will need to check with him on this . . . I just hope I did not miss Question 4; it was somewhat difficult and I am not too sure about my answer there. Let me check with Jack on this, as he knows the material quite well. Ah! Jack seems to have gotten the same answer I got. I think it is Question 3 after all . . . I'd better see the professor soon to make sure he graded this one.

One striking aspect of this example is the multiplicity of assumptions involved in forming Drew's initial belief in having received an A grade (i.e., Question 3 was graded, Question 4 was solved correctly, the professor did not make a clerical error, and so on). The example also brings out important notions that were used by Drew in resolving conflicts among assumptions. This includes the strength of an assumption, which can be based on previous experiences (e.g., I have taken a few classes with this professor). It also includes the notion of evidence, which may be brought to bear on the validity of these assumptions (i.e., let me check with Jack).

Having reached this stage of our discussion on the subtleties of commonsense reasoning, one could drive it further in one of two directions. We can continue to elaborate on non-monotonic logics and how they may go about resolving conflicts among assumptions. This will also probably lead us into the related subject of *belief revision*, which aims at regulating this conflict-resolution process through a set of rationality postulates [Gärdenfors, 1988]. However, as these subjects are outside the scope of this book, we will turn in a different direction that underlies the formalism we plan to pursue in the upcoming chapters. In a nutshell, this new direction can be viewed as postulating the existence of a more fundamental notion, called a degree of belief, which, according to some treatments, can alleviate the need for assumptions altogether and, according to others, can be used as a basis for deciding which assumptions to make in the first place.

1.2 Degrees of belief

A *degree of belief* is a number that one assigns to a proposition in lieu of having to declare it as a fact (as in deductive logic) or an assumption (as in non-monotonic logic). For example, instead of assuming that a bird is normal unless observed otherwise – which

leads us to tenuously believe that it also flies – we assign a degree of belief to the bird's normality, say, 99%, and then use this to derive a corresponding degree of belief in the bird's flying ability.

A number of different proposals have been extended in the literature for interpreting degrees of belief including, for example, the notion of possibility on which fuzzy logic is based. This book is committed to interpreting degrees of belief as probabilities and, therefore, to manipulating them according to the laws of probability. Such an interpretation is widely accepted today and underlies many of the recent developments in automated reasoning. We will briefly allude to some of the classical arguments supporting this interpretation later but will otherwise defer the vigorous justification to cited references [Pearl, 1988; Jaynes, 2003].

While assumptions address the monotonicity problem by being assertible and retractible, degrees of belief address this problem by being revisable either upward or downward, depending on what else is known. For example, we may initially believe that a bird is normal with probability 99%, only to revise this to, say, 20% after learning that its wing is suffering from some wound. The dynamics that govern degrees of belief will be discussed at length in Chapter 3, which is dedicated to probability calculus, our formal framework for manipulating degrees of belief.

One can argue that assigning a degree of belief is a more committing undertaking than making an assumption. This is due to the fine granularity of degrees of beliefs, which allows them to encode more information than can be encoded by a binary assumption. One can also argue to the contrary that working with degrees of belief is far less committing as they do not imply any particular truth of the underlying propositions, even if tenuous. This is indeed true, and this is one of the key reasons why working with degrees of belief tends to protect against many pitfalls that may trap one when working with assumptions; see Pearl [1988], Section 2.3, for some relevant discussion on this matter.

1.2.1 Deciding after believing

Forming beliefs is the first step in making decisions. In an assumption-based framework, decisions tend to follow naturally from the set of assumptions made. However, when working with degrees of belief, the situation is a bit more complex since decisions will have to be made without assuming any particular state of affairs. Suppose for example that we are trying to capture a bird that is worth $40.00 and can use one of two methods, depending on whether it is a flying bird or not. The assumption-based method will have no difficulty making a decision in this case, as it will simply choose the method based on the assumptions made. However, when using degrees of belief, the situation can be a bit more involved as it generally calls for invoking *decision theory*, whose purpose is to convert degrees of beliefs into definite decisions [Howard and Matheson, 1984; Howard, 1990]. Decision theory needs to bring in some additional information before it can make the conversion, including the cost of various decisions and the rewards or penalties associated with their outcomes. Suppose for example that the first method is guaranteed to capture a bird, whether flying or not, and costs $30.00, while the second method costs $10.00 and is guaranteed to capture a non-flying bird but may capture a flying bird with a 25% probability. One must clearly factor all of this information before one can make the right decision in this case, which is precisely the role of decision theory. This theory is therefore an essential complement to the theory of probabilistic reasoning discussed in this book.

Yet we have decided to omit the discussion of decision theory here to keep the book focused on the modeling and reasoning components (see Pearl [1988], Jensen and Nielsen [2007] for a complementary coverage of decision theory).

1.2.2 What do the probabilities mean?

A final point we wish to address in this section concerns the classical controversy of whether probabilities should be interpreted as objective frequencies or as subjective degrees of belief. Our use of the term "degrees of belief" thus far may suggest a commitment, to the subjective approach, but this is not necessarily the case. In fact, none of the developments in this book really depend on any particular commitment, as both interpretations are governed by the same laws of probability. We will indeed discuss examples in Chapter 5 where all of the used probabilities are degrees of belief reflecting the state of knowledge of a particular individual and not corresponding to anything that can be measured by a physical experiment. We will also discuss examples in which all of the used probabilities correspond to physical quantities that can be not only measured but possibly controlled as well. This includes applications from system analysis and diagnostics, where probabilities correspond to the failure rates of system components, and examples from channel coding, where the probabilities correspond to channel noise.

1.3 Probabilistic reasoning

Probability theory has been around for centuries. However, its utilization in automated reasoning at the scale and rate within AI has never before been attempted. This has created some key computational challenges for probabilistic reasoning systems, which had to be confronted by AI researchers for the first time. Adding to these challenges is the competition that probabilistic methods had initially received from symbolic methods that were dominating the field of AI at the time. It is indeed the responses to these challenges over the last few decades that have led to much of the material discussed in this book. One therefore gains more perspective and insights into the utility and significance of the covered topics once one is exposed to some of these motivating challenges.

1.3.1 Initial reactions

AI researchers proposed the use of numeric degrees of belief well before the monotonicity problem of classical logic was unveiled or its consequences absorbed. Yet such proposals were initially shunned based on cognitive, pragmatic, and computational considerations. On the cognitive side, questions were raised regarding the extent to which humans use such degrees of belief in their own reasoning. This was quite an appealing counterargument at the time, as the field of AI was still at a stage of its development where the resemblance of formalism to human cognition was very highly valued, if not necessary. On the pragmatic side, questions were raised regarding the availability of degrees of beliefs (where do the numbers come from?). This came at a time when the development of knowledge bases was mainly achieved through knowledge elicitation sessions conducted with domain experts who, reportedly, were not comfortable committing to such degrees – the field of statistical machine learning had yet to be influential enough then. The robustness of probabilistic reasoning systems was heavily questioned as well (what happens if I change this .90 to .95?). The issue here was not only whether probabilistic reasoning was robust enough

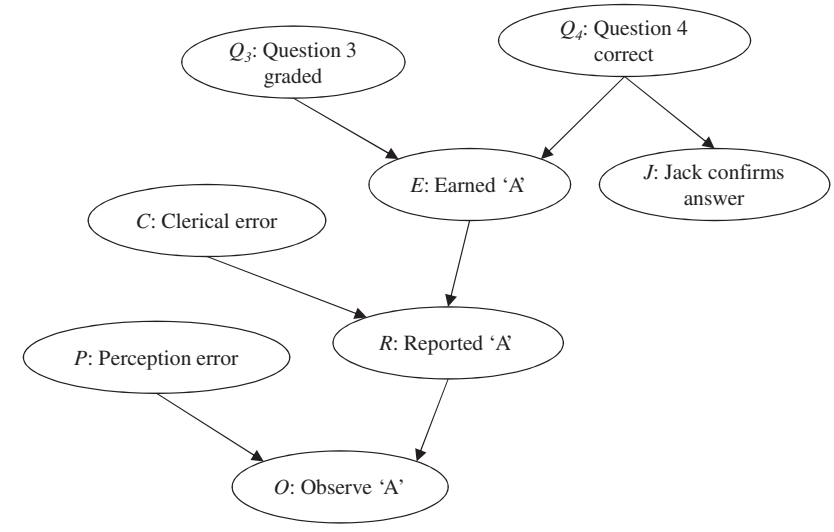
against such perturbations but, in situations where it was shown to be robust, questions were raised about the unnecessary level of detail demanded by specifying probabilities.

On the computational side, a key issue was raised regarding the scale of applications that probabilistic reasoning systems can handle, at a time when applications involving dozens if not hundreds of variables were being sought. Such doubts were grounded in the prevalent perception that joint probability distributions, which are exponentially sized in the number of used variables, will have to be represented explicitly by probabilistic reasoning systems. This would be clearly prohibitive on both representational and computational grounds for most applications of interest. For example, a medical diagnosis application may require hundreds of variables to represent background information about patients, in addition to the list of diseases and symptoms about which one may need to reason.

1.3.2 A second chance

The discovery of the qualification problem, and the associated monotonicity problem of deductive logic, gave numerical methods a second chance in AI, as these problems created a vacancy for a new formalism of commonsense reasoning during the 1980s. One of the key proponents of probabilistic reasoning at the time was Judea Pearl, who seized upon this opportunity to further the cause of probabilistic reasoning systems within AI. Pearl had to confront challenges on two key fronts in this pursuit. On the one hand, he had to argue for the use of numbers within a community that was heavily entrenched in symbolic formalism. On the other hand, he had to develop a representational and computational machinery that could compete with symbolic systems that were in commercial use at the time.

On the first front, Pearl observed that many problems requiring special machinery in logical settings, such as non-monotonicity, simply do not surface in the probabilistic approach. For example, it is perfectly common in probability calculus to see beliefs going up and down in response to new evidence, thus exhibiting a non-monotonic behavior – that is, we often find $\Pr(A) > \Pr(A|B)$ indicating that our belief in A would go down when we observe B. Based on this and similar observations, Pearl engaged in a sequence of papers that provided probabilistic accounts for most of the paradoxes that were entangling symbolic formalisms at the time; see Pearl [1988], Chapter 10, for a good summary. Most of the primitive cognitive and pragmatic arguments (e.g., people do not reason with numbers; where do the numbers come from?) were left unanswered then. However, enough desirable properties of probabilistic reasoning were revealed to overwhelm and silence these criticisms. The culminations of Pearl's efforts at the time were reported in his influential book, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* [Pearl, 1988]. The book contained the first comprehensive documentation of the case for probabilistic reasoning, delivered in the context of contemporary questions raised by AI research. This part of the book was concerned with the foundational aspects of plausible reasoning, setting clear the principles by which it ought to be governed – probability theory, that is. The book also contained the first comprehensive coverage of Bayesian networks, which were Pearl's response to the representational and computational challenges that arise in realizing probabilistic reasoning systems. On the representational side, the Bayesian network was shown to compactly represent exponentially sized probability distributions, addressing one of the classical criticisms against probabilistic reasoning systems. On the computational side, Pearl developed the polytree algorithm [Pearl, 1986b], which was the first general-purpose inference algorithm for networks that contain no



C	$\Pr(C)$
true	.001

Q_3	Q_4	E	$\Pr(E \mid Q_3, Q_4)$
true	true	true	1
true	false	true	0
false	true	true	0
false	false	true	0

Q_4	J	$\Pr(J \mid Q_4)$
true	true	.99
false	true	.20

Figure 1.3: The structure of a Bayesian network, in which each variable can be either true or false. To fully specify the network, one needs to provide a probability distribution for each variable, conditioned on every state of its parents. The figure shows these conditional distributions for three variables in the network.

directed loops.[1] This was followed by the influential jointree algorithm [Lauritzen and Spiegelhalter, 1988], which could handle arbitrary network structures, albeit inefficiently for some structures. These developments provided enough grounds to set the stage for a new wave of automated reasoning systems based on the framework of Bayesian networks (e.g., [Andreassen et al., 1987]).

1.4 Bayesian networks

A *Bayesian network* is a representational device that is meant to organize one's knowledge about a particular situation into a coherent whole. The syntax and semantics of Bayesian networks will be covered in Chapter 4. Here we restrict ourselves to an informal exposition that is sufficient to further outline the subjects covered in this book.

Figure 1.3 depicts an example Bayesian network, which captures the information corresponding to the student scenario discussed earlier in this chapter. This network has two components, one qualitative and another quantitative. The qualitative part corresponds to the directed acyclic graph (DAG) depicted in the figure, which is also known as the

[1] According to Pearl, this algorithm was motivated by the work of Rumelhart [1976] on reading comprehension, which provided compelling evidence that text comprehension must be a distributed process that combines both top-down and bottom-up inferences. This dual mode of inference, so characteristic of Bayesian analysis, did not match the capabilities of the ruling paradigms for uncertainty management in the 1970s. This led Pearl to develop the polytree algorithm [Pearl, 1986b], which appeared first in Pearl [1982] with a restriction to trees, and then in Kim and Pearl [1983] for polytrees.

structure of the Bayesian network. This structure captures two important parts of one's knowledge. First, its variables represent the primitive propositions that we deem relevant to our domain. Second, its edges convey information about the dependencies between these variables. The formal interpretation of these edges will be given in Chapter 4 in terms of probabilistic independence. For now and for most practical applications, it is best to think of these edges as signifying direct causal influences. For example, the edge extending from variable E to variable R signifies a direct causal influence between earning an A grade and reporting the grade. Note that variables Q_3 and Q_4 also have a causal influence on variable R yet this influence is not direct, as it is mediated by variable E. We stress again that Bayesian networks can be given an interpretation that is completely independent of the notion of causation, as in Chapter 4, yet thinking about causation will tend to be a very valuable guide in constructing the intended Bayesian network [Pearl, 2000; Glymour and Cooper, 1999].

To completely specify a Bayesian network, one must also annotate its structure with probabilities that quantify the relationships between variables and their parents (direct causes). We will not delve into this specification procedure here but suffice it to say it is a localized process. For example, the probabilities corresponding to variable E in Figure 1.3 will only reference this variable and its direct causes Q_3 and Q_4. Moreover, the probabilities corresponding to variable C will only reference this variable, as it does not have any causes. This is one of the key representational aspects of a Bayesian network: we are never required to specify a quantitative relationship between two variables unless they are connected by an edge. Probabilities that quantify the relationship between a variable and its indirect causes (or its indirect effects) will be computed automatically by inference algorithms, which we discuss in Section 1.4.2.

As a representational tool, the Bayesian network is quite attractive for three reasons. First, it is a consistent and complete representation as it is guaranteed to define a unique probability distribution over the network variables. Hence by building a Bayesian network, one is specifying a probability for every proposition that can be expressed using these network variables. Second, the Bayesian network is modular in the sense that its consistency and completeness are ensured using localized tests that apply only to variables and their direct causes. Third, the Bayesian network is a compact representation as it allows one to specify an exponentially sized probability distribution using a polynomial number of probabilities (assuming the number of direct causes remains small).

We will next provide an outline of the remaining book chapters, which can be divided into two components corresponding to modeling and reasoning with Bayesian networks.

1.4.1 Modeling with Bayesian networks

One can identify three main methods for constructing Bayesian networks when trying to model a particular situation. These methods are covered in four chapters of the book, which are outlined next.

According to the first method, which is largely subjective, one reflects on their own knowledge or the knowledge of others and then captures it into a Bayesian network (as we have done in Figure 1.3). According to the second method, one automatically synthesizes the Bayesian network from some other type of formal knowledge. For example, in many applications that involve system analysis, such as reliability and diagnosis, one can synthesize a Bayesian network automatically from formal system designs. Chapter 5 will be concerned with these two modeling methods, which are sometimes known as

the *knowledge representation (KR) approach* for constructing Bayesian networks. Our exposure here will be guided by a number of application areas in which we state problems and show how to solve them by first building a Bayesian network and then posing queries with respect to the constructed network. Some of the application areas we discuss include system diagnostics, reliability analysis, channel coding, and genetic linkage analysis.

Constructing Bayesian networks according to the KR approach can benefit greatly from sensitivity analysis, which is covered partly in Chapter 5 and more extensively in Chapter 16. Here we provide techniques for checking the robustness of conclusions drawn from Bayesian networks against perturbations in the local probabilities that annotate them. We also provide techniques for automatically revising these local probabilities to satisfy some global constraints that are imposed by the opinions of experts or derived from the formal specifications of the tasks under consideration.

The third method for constructing Bayesian networks is based on learning them from data, such as medical records or student admissions data. Here either the structure, the probabilities, or both can be learned from the given data set. Since learning is an inductive process, one needs a principle of induction to guide the construction process according to this *machine learning (ML) approach*. We discuss two such principles in this book, leading to what are known as the maximum likelihood and Bayesian approaches to learning. The maximum likelihood approach, which is discussed in Chapter 17, favors Bayesian networks that maximize the probability of observing the given data set. The Bayesian approach, which is discussed in Chapter 18, uses the likelihood principle in addition to some prior information that encodes preferences on Bayesian networks.[2]

Networks constructed by the KR approach tend to have a different nature than those constructed by the ML approach. For example, these former networks tend to be much larger in size and, as such, place harsher computational demands on reasoning algorithms. Moreover, these networks tend to have a significant amount of determinism (i.e., probabilities that are equal to 0 or 1), allowing them to benefit from computational techniques that may be irrelevant to networks constructed by the ML approach.

1.4.2 Reasoning with Bayesian networks

Let us now return to Figure 1.1, which depicts the architecture of a knowledge-based reasoning system. In the previous section, we introduced those chapters that are concerned with constructing Bayesian networks (i.e., the knowledge bases or models). The remaining chapters of this book are concerned with constructing the reasoning engine, whose purpose is to answer queries with respect to these networks. We will first clarify what is meant by reasoning (or inference) and then lay out the topics covered by the reasoning chapters.

We have already mentioned that a Bayesian network assigns a unique probability to each proposition that can be expressed using the network variables. However, the network itself only explicates some of these probabilities. For example, according to Figure 1.3 the probability of a clerical error when entering the grade is .001. Moreover, the probability

[2] It is critical to observe here that the term "Bayesian network" does not necessarily imply a commitment to the Bayesian approach for learning networks. This term was coined by Judea Pearl [Pearl, 1985] to emphasize three aspects: the often subjective nature of the information used in constructing these networks; the reliance on Bayes's conditioning when reasoning with Bayesian networks; and the ability to perform causal as well as evidential reasoning on these networks, which is a distinction underscored by Thomas Bayes [Bayes, 1963].