Chapter 1

Introduction

Uncertainty is a fundamental—and unavoidable—feature of daily life. In order to deal with uncertainty intelligently, we need to be able to represent it and reason about it. How to do that is what this book is about.

Reasoning about uncertainty can be subtle. If it weren't, this book would be much shorter. The following puzzles and problems hint at some of the subtleties.

- Consider the *second-ace puzzle*: Suppose that we have a deck with four cards: the ace and deuce of hearts, and the ace and deuce of spades. After a fair shuffle of the deck, two cards are dealt to Alice. It is easy to see that, at this point, there is a probability of $1/6$ that Alice has both aces, probability $5/6$ that Alice has at least one ace, probability $1/2$ that Alice has the ace of spades, and probability $1/2$ that Alice has the ace of hearts: Out of the six possible deals of two cards out of four, Alice has both aces in one of them, at least one ace in five of them, the ace of hearts in three of them, and the ace of spades in three of them. (For readers unfamiliar with probability, there is an introduction in Chapter 3.)

  Alice then says "I have an ace". Conditioning on this information (by discarding the possibility that Alice was dealt no aces), Bob computes the probability that Alice holds both aces to be $1/5$. This seems reasonable: The probability of Alice having two aces goes up if we find out she has an ace. Next, Alice says "I have the ace of spades". Conditioning on this new information, Bob now computes the probability that Alice holds both aces to be $1/3$. Of the three deals in which Alice holds the ace of spades, she holds both aces in one of them. As a result of learning not only that Alice holds at least one ace, but that the ace is actually the ace of spades, the conditional probability that Alice holds both aces goes up from 1/5 to 1/3. Similarly, if Alice had said "I have the ace of hearts", the conditional probability that Alice holds both aces would be 1/3.

But is this reasonable? When Bob learns that Alice has an ace, he knows that she must have either the ace of hearts or the ace of spades. Why should finding out which particular ace it is raise the conditional probability of Alice having two aces?

- The second-ace puzzle is very similar to the *Monty Hall puzzle*: Suppose you're on a game show and given a choice of three doors. Behind one is a car; behind the others are goats. You pick door Number 1. Before opening door Number 1, Monty Hall, the host (who knows what is behind each door) opens door Number 3, which has a goat. He then asks you if you still want to take what's behind door Number 1, or to take what's behind door Number 2 instead. Should you switch?

- Puzzles like the second-ace puzzle and the Monty Hall puzzle are the stuff of puzzle books. The following *two-coin problem* doesn't seem to have the subtlety of these puzzles. Suppose Alice has two coins. One of them is fair, and so has equal likelihood of landing heads and tails. The other is biased, and is twice as likely to land heads as to land tails. Alice chooses one of her coins (assume she can tell them apart by their weight and feel) and is about to toss it. Bob knows that one coin is fair and the other is twice as likely to land heads as tails. He does not know which coin Alice has chosen, nor is he given a probability that the fair coin is chosen. What is the probability, according to Bob, that the outcome of the coin toss will be heads? What is the probability according to Alice? (Both of these probabilities are for the situation *before* the coin is tossed.)

- Suppose instead that both Bob and Alice know that Alice is using the fair coin. In this *one-coin problem*, Alice tosses the coin and looks at the outcome. What is the probability of heads (after the coin toss) according to Bob? One argument would say that the probability is still 1/2. After all, Bob hasn't learned anything about the outcome of the coin toss, so why should he change his valuation of the probability? On the other hand, runs the counterargument, once the coin has been tossed, can we really talk about the probability of heads? It has either landed heads or tails, so at best, Bob can say that the probability is either 0 or 1, but he doesn't know which.

- On a more serious note, consider a doctor who is examining a patient Eric. The doctor can see that Eric has jaundice, no temperature, and red hair. According to his medical textbook, 90% of patients with jaundice have hepatitis and 80% of patients with hepatitis have a temperature. This is all the information he has that is relevant to the problem. Should he proceed under the assumption that Eric has hepatitis?

  There is clearly some ambiguity in the presentation of this problem (far more than, say, in the presentation of the second-ace puzzle). For example, we have not been told what other options the doctor has. Even ignoring this ambiguity, there are many issues that arise here. An obvious one is how the doctor's statistical information should affect his beliefs regarding what to do. There are many others though. For example, we must decide what it means that the doctor has no other relevant information. Typically the doctor has a great deal of information, and part of the problem lies in deciding what is and is not relevant. Another issue is perhaps more pragmatic. How can we represent the doctor's information? For example, how can we represent the fact that certain things are irrelevant? The doctor may feel that the fact that Eric has red hair is irrelevant to the question of whether he has hepatitis. How do we represent that?

  In many cases, we do not have quantitative information, only qualitative information. For example, rather than knowing that 90% of patients with jaundice have hepatitis and 80% of patients with hepatitis have a temperature, he may only know that patients with jaundice typically have hepatitis, and patients with hepatitis typically have a temperature. How does this affect things?

- The world is not static. We typically acquire new information all the time. How should the new information affect our beliefs and how should it be incorporated in whatever representation we choose? The standard way of incorporating new information in probability theory is by *conditioning*. This is what Bob used in the second-ace puzzle above to incorporate the information he got from Alice, such as the fact that she holds an ace, or that she holds the ace of hearts. This puzzle already suggests that there are subtleties involved with conditioning. Things get even more complicated if we do not represent our uncertainty using probability, or if the information that we get cannot be conditioned on.

So how do we reason about uncertainty? A standard approach is to design a logic, replete with axioms and inference rules, argue that the axioms capture some fundamental features of importance, and then build an inference engine based on this system. While axioms and inference are discussed from time to time, they tend to occupy the back seat for most of this book. I spend much more time discussing and comparing different approaches to representing uncertainty. Only after this discussion do I consider formal logics for reasoning about uncertainty. Still, since there have been several attacks recently on the role of logic, some justification for the use of logic seems called for.

There is a great deal of complexity in the real world. Typically, we want to reason about certain phenomena of interest, ignoring many details that are taken to be irrelevant. To reason about these phenomena in a rigorous way, it is often useful to try to capture them using a formal mathematical model. That is the role of a formal logic. The term "formal logic" as I use it here means a *syntax* or *language*—that is, a collection of well-formed formulas—together with a *semantics*—a method for deciding whether a given formula in the language is true or false. But not just any syntax and semantics will do. The semantics should bear a clear and natural relationship to the real-world phenomena it is trying to model, and the syntax should be well-suited to its purpose. It should be easy to render the statements we want to express as formulas in our language. If we cannot capture a lot of statements we want to make in an easy and natural way, the logic is not doing its job. Of course, "ease", "clarity", and "naturalness" are in the eye of the beholder. To complicate the matter, expressive power usually comes at a price. A more expressive logic is typically more complex than a less expressive one. This makes the task of designing a useful logic, or choosing among several pre-existing candidates, far more of an art than a science, and one that requires a deep understanding of the phenomena that we are reasoning about.

So what is the advantage of using logic at all? Is it really worth the overhead? I am biased, of course, since much of my professional career has involved studying logic. Nevertheless, I believe a good formal logic can certainly help guide our reasoning. More importantly, it can help clarify representation issues. In particular, a well-chosen syntax may force us to clarify ambiguities that often exist in natural language, and a good semantics should be able to show relationships between situations far more clearly than we could just through symbol manipulation. In particular, as we shall see, having the "right" syntax and semantics can certainly help in understanding the puzzles presented above. Once we have an appropriate representation, we can also investigate the problem of automating the process of reasoning. While this is an extremely important topic, it is one that I do not discuss in this book—the emphasis is instead on the underlying representation.

I make heavy use throughout the book of the *possible-world* approach to representing uncertainty. The intuitive idea is that there are many possible ways the world could be. The agent does not know which one is the true description of the real world, so she must consider them all. Of course, the agent may have additional structure on the set of worlds she considers possible. For example, she may consider some worlds more likely than others. This likelihood may be represented in a quantitative way, for example, by using probability, or, more qualitatively, by a preference ordering on worlds. There are a number of alternatives here; using possible worlds provides a general framework in which to explore them.

The rest of this book is organized as follows. After a brief discussion of propositional logic, the possible worlds framework is introduced in Chapter 2, together with some propositional *modal* logics that are appropriate for reasoning about knowledge and belief in the possible-worlds framework in a qualitative way.

Chapter 3 deals with more quantitative approaches to representing uncertainty, still in the possible-worlds framework. Probability is considered, of course, but so are other methods of representing uncertainty, including *Dempster-Shafer belief functions*, *possibility measures*, *ranking functions*, and *possibility measures*.

Once we have a method of representing uncertainty, we need to find ways of updating it in the light of new information. This is the subject of Chapter 4. The effect of receiving new information on our beliefs depends on the relationship between various propositions of interest. In particular, it is important to know whether they are related or independent. Independence is considered in more detail in Chapter **??**. I also consider *Bayesian networks*, which provide a way of representing the dependencies between propositions. Bayesian networks can be viewed as complementary to the possible-worlds approach; the two are carefully related. Finally, I consider the notion of *expectation* in Chapter **??**.

Chapter 5 marks a return to logic. There are many possible logics for reasoning about uncertainty. The appropriate choice depends in part on the underlying method for representing uncertainty. I consider logics for each of the methods of representing uncertainty discussed in the preceding chapters.

Chapter 6 deals with *defaults* and *counterfactuals*. Default reasoning involves reasoning about statement like "birds typically fly" and "patients with hepatitis typically have jaundice". Such reasoning may be *nonmonotonic*: Strengthening hypotheses may cause us to change our conclusions. For example, although birds typically fly, penguins typically do not fly. Thus, if we learn about a bird that it is a penguin, we may want to retract our initial conclusion that it flies. Counterfactual reasoning involves reasoning about statements that may be counter to what actually occurred. Statements like "If I hadn't slept in [although I did] I wouldn't have been late for my wedding" are counterfactuals. Somewhat surprisingly, defaults and counterfactuals can be modeled using very similar techniques; moreover, these techniques use the methods of representing uncertainty introduced in Chapter 3.

In Chapter 7 considers combined logics, that allow for reasoning about several methods of representing uncertainty at the same time, so that, for example, we can reason about knowledge and probability. It turns out that there are some surprising interactions between knowledge and probability, which are related to subtleties involving the distinction between probability and nondeterminism. Dealing with these subtleties gives some insight into the two-coin and one-coin problems mentioned earlier.

In Chapter 8 deals with more dynamic aspects of belief and probability. To talk about these dynamic aspects, it is helpful to add time explicitly to the framework. We can then do a formal analysis of the second-ace puzzle. It turns out that in order to represent the puzzle formally, we must describe the *protocols* being followed by Alice and Bob. The protocol, in turn, determines the set of *runs*, or possible sequences of events that might happen. The key question here is what Alice's protocol says to do after she has answered "Yes" to Bob's question as to whether she has an ace. Roughly speaking, if her protocol is "if I have the ace of spades, then I will say that, otherwise I will say nothing", then $1/3$ is indeed appropriate as Bob's estimate of the probability that Alice has both aces. This is the conditional probability of Alice having both aces given that she has the ace of spades; the original information is subsumed by the new statement. On the other hand, if her protocol is "I will tell Bob which ace I have; if I have both, I will choose at random between the ace of hearts and the ace of spades", then in fact, Bob's conditional probability should not go up to $1/3$, but should stay at $1/5$. In this case, the second statement contains no new information. The different protocols determine different possible runs, and so result in different probability spaces. More generally, we show how a logic based on protocols allows us to reason correctly about situations of this kind.

Having time in the framework also makes it easier to consider the problem *belief revision* in a more qualitative setting. How should an agent revise her beliefs in the light of new information, especially when the information contradicts her old beliefs? This is the subject of Chapter **??**.

Propositional logic is known to be quite weak. Modalities such as knowledge, belief, and probability give us a great deal of added expressive power. Moving to first-order logic also gives us a great deal of additional expressive power, but in a much different dimension. It allows to reason about individuals and their properties. Of course, we can combine modal logic and first-order logic. In Chapter **??**, after a brief introduction to first-order logic, I consider first-order modal logic. When considering first-order logics of probabilities, we need to make a distinction between two kinds of "probabilities" that are often confounded: statistical information (such as "90% of birds fly") and degrees of belief (such as "My degree of belief that Tweety—a particular bird—flies is .9."). I discuss an approach that allows us to carefully make such distinctions. As we shall see, the same issues arise in more qualitative approaches to reasoning about likelihood, and the same general approach works for them as well.

Once we make the distinction between these two different types of probability, we are immediately led to asking what the connection should be between them. To understand this problem, suppose that we have the statistical information that 90% of birds fly and we know that Tweety is a bird. What should our degree of belief be that Tweety flies? If this is all we know about Tweety, then it seems reasonable to take the degree of belief to be .9. But what if we know that Tweety is a yellow bird? Should the fact that it is yellow influence our degree of belief that Tweety flies? What if also know that Tweety is a penguin, and only 5% of penguins fly? Then it seems more reasonable to take our degree of belief to be .05 rather than .9. But how can we justify this? More generally, how do we go about computing degrees of belief. In Chapter **??**, I describe a general approach to this problem. The basic idea is quite simple. Given a knowledge base KB, we construct a set of possible worlds: the different worlds consistent with the knowledge base. We then put a uniform probability distribution on this set of possible worlds. Our degree of belief in a fact φ is then the fraction of the worlds consistent with the KB in which φ is true. We examine this approach and some of its variants, and show that it has some rather attractive (and some not so attractive) properties. I also discuss one application of this approach, to default reasoning.

I have tried to write this book in as modular a fashion as possible. Figure **??** describes the dependencies between chapters. An arrow from one chapter to another indicates that it is necessary to read (at least part of) the first to understand (at least part of) the second. Where the dependency involves only one or two sections, I have labeled the arrows. For example, ***

I cover much of this material in a one-semester course at Cornell University. In a typical semester, I cover ...

Formal proofs of many of the statements in the text are left to the exercises, as well as a more detailed examination of some tangential topics. I would strongly encourage the reader to read over all the exercises and attempt as many as possible. This is the best way to master the material!

Each chapter ends with a section of notes, that provide references to material and, occasionally, more details on some material not covered in the chapter. Although the bibliography is extensive, reasoning about uncertainty is a huge area, and I am sure that I have (inadvertently!) left out relevant references. I apologize in advance for any such omissions.

Notes

Many books have been written recently regarding alternative approaches to reasoning about uncertainty. Shafer [1976] provides a good introduction to the Dempster-Shafer approach; [Klir and Folger 1988] is a good introduction to fuzzy logic; [Shafer and Pearl 1990] contains a good overview of a number of approaches.

There are numerous discussions of the subtleties in probabilistic reasoning. A particularly good one is [Bar-Hillel and Falk 1982], where a discussion of the second-ace puzzle (and the related three-prisoners puzzle, discussed in Section 4.3) can be found. Further discussion of the second-ace puzzle can be found in [Freund 1965; Shafer 1985]. The Monty Hall puzzle is also an old problem [Mosteller 1965]; it has generated a great deal of discussion since it was discussed by vos Savant in *Parade Magazine* [1990a, 1990b, 1991]. Further discussion can be found in [Morgan, Chaganty, Dahiya, and Doviak 1991].

The discussion of the role of logic in reasoning about uncertainty is largely taken from [Halpern 1990]. This article is a response to an article by Cheeseman [1985] that suggests that logic is unnecessary in reasoning about uncertainty.

See the notes at the end of later chapters for more references on the specific subjects discussed in these chapters.

Chapter 2

Propositional Modal Logic

As I said in Chapter 1, logic can be a useful tool in grappling with the complexities of reasoning about uncertainty. Perhaps the simplest logic considered in the literature, and the one that most students encounter initially, is *propositional logic* (sometimes called *sentential logic*). It is intended to capture features of arguments such as the following:

> Borogroves are mimsy whenever it is brillig. It is now brillig and this thing is a borogrove. Hence this thing is mimsy.

While propositional logic is useful for reasoning about conjunctions, negations, and implications, it is not so useful when it comes to dealing with notions like knowledge or likelihood. For example, notions like "Alice *knows* it is mimsy" or "it is more likely to be mimsy than not" cannot be expressed in propositional logic. Such statements are crucial for reasoning about uncertainty. Knowledge is an example of what philosophers have called a *propositional attitude*. Such propositional attitudes can be expressed using *modal logic*.

Modal logic gives us a powerful tool for modeling uncertainty. This chapter is an introduction to propositional modal logic. To make the presentation self-contained, I start with the basics—propositional logic—and then move to modal logic.

2.1 Propositional Logic

The standard approach to reasoning using logic typically starts with *syntax*, a formal language for capturing what it is we want to say. It is not obvious we always need a language; sometimes it may be easier to work directly with a model. I return to this issue in Section 2.2.4. However, for now, I review the standard approach.

The formal syntax for propositional logic is quite straightforward: we start with a *vocabulary*—a nonempty set Φ of *primitive propositions* or *basic facts*, which I typically label by letters such as p and q (perhaps with a prime or subscript). These primitive propositions can be thought of as representing statements such as "it is brillig" or "this thing is mimsy". More complicated formulas are formed by closing off under conjunction and negation, so that if φ and ψ are formulas, then so are $\neg\varphi$ and $\varphi \wedge \psi$ (read "not φ" and "φ and ψ", respectively). Thus, if p stands for "it is brillig" and q stands for "this thing is mimsy", then $\neg p$ says "it is not brillig", while $p \wedge q$ says "it is brillig and this thing is mimsy". The set $\mathcal{L}^{Prop}(\Phi)$ of formulas consists precisely of all the formulas that can be formed in this way.

There are some other standard connectives that can easily be defined in terms of conjunction and negation:

- $\varphi \vee \psi$ (read "φ or ψ") is an abbreviation for $\neg(\neg\varphi \wedge \neg\psi)$
- $\varphi \Rightarrow \psi$ (read "φ implies ψ" or "if φ then ψ") is an abbreviation for $\neg\varphi \vee \psi$
- $\varphi \Leftrightarrow \psi$ (read "φ if and only if ψ") is an abbreviation for $(\varphi \Rightarrow \psi) \wedge (\psi \Rightarrow \varphi)$
- *true* is an abbreviation for $p \vee \neg p$ (where p is a fixed primitive propositions in Φ)
- *false* is an abbreviation for $\neg$*true*.

If p and q stand for "it is brillig" and "this thing is mimsy", as before, and r stands "this thing is a borogrove", then the argument above can be easily captured in propositional logic. "Borogroves are mimsy whenever it is brillig" can be restated as "if it is brillig then [if this thing is a borogrove, then it is mimsy]". (Check that these two English statements really are saying the same thing!) Thus, this becomes $p \Rightarrow (r \Rightarrow q)$. "It is now brillig and this thing is a borogrove" becomes $p \wedge r$. We want to conclude q: "this thing is mimsy". This seems to be a legitimate conclusion. How can we formalize its legitimacy?

So far, we have defined a formal language, along with an intended reading of formulas in the language. The intended reading is supposed to correspond to intuitions that we have regarding words like "and", "or", and "not". We capture these intuitions by providing a *semantics* for the formulas in the language, that is, a method for deciding whether a given formula is true or false.

The key component of the semantics for propositional logic is a *truth assignment* v, a function that maps the primitive propositions in Φ to a *truth value*, that is, an element of the set $\{\mathbf{true}, \mathbf{false}\}$. The form of the truth assignment guarantees that each primitive proposition has exactly one truth value. It is either true or false; it cannot be both true and false, or neither true nor false. This commitment, which gives us *classical* or *two-valued* logic, has been subject to criticism; some alternative approaches are mentioned in the notes at the end of the chapter.

A truth assignment determines which primitive propositions are true and which are false. There are standard rules for then determining whether an arbitrary formula φ is *true under truth assignment* v, or *satisfied by truth assignment* v, written $v \models \varphi$. Formally, the $\models$ relation is defined by induction on the structure of φ. That is, we start with the simplest formulas, namely, the primitive propositions, and extend the definition to more complicated formulas of the form $\neg\varphi$ or $\varphi \wedge \psi$ under the assumption that we have already computed whether each of the constituents is true under v. For a primitive proposition p, we have

$$v \models p \text{ iff } v(p) = \mathbf{true}.$$

Thus, a primitive proposition is true under truth assignment v if and only if the truth assignment assigns it truth value **true**.

Intuitively, $\neg\varphi$ is true if and only if φ is false. This intuition is captured as follows:

$$v \models \neg\varphi \text{ iff } v \not\models \varphi.$$

It is also easy to formalize the intuition that $\varphi \wedge \psi$ is true if and only if both φ and ψ are true:

$$v \models \varphi \wedge \psi \text{ iff } v \models \varphi \text{ and } v \models \psi.$$

Now we need to check that our abbreviations capture their intended intuitions. For example, we would expect that $\varphi \vee \psi$ is true exactly if one of φ or ψ is true. It is not immediately obvious that our definition of $\varphi \vee \psi$ as an abbreviation for $\neg(\neg\varphi \wedge \neg\psi)$ enforces this intuition. As the following lemma shows, it does, not only for $\varphi \vee \psi$, but for all the other abbreviations that were defined earlier.

Lemma 2.1.1 *For every truth assignment v, we have:*

(a) $v \models \varphi \vee \psi$ iff $v \models \varphi$ or $v \models \psi$.

(b) If $v \models \varphi \Rightarrow \psi$, then if $v \models \varphi$ then $v \models \psi$.

(c) $v \models \varphi \Leftrightarrow \psi$ iff either $v \models \varphi$ and $v \models \psi$ or $v \models \neg\varphi$ and $v \models \neg\psi$.

(d) $v \models true$.

(e) $v \not\models false$.

Proof I prove part (a), leaving the remainder of the parts as an exercise for the reader (Exercise 2.1). Suppose $v \models \varphi \vee \psi$. This is the case iff $v \models \neg(\neg\varphi \wedge \neg\psi)$. This, in turn, is the case iff $v \not\models \neg\varphi \wedge \neg\psi$. The definition of $\models$ guarantees that v does not satisfy a conjunction iff it does not satisfy one of the conjuncts: that is, iff $v \not\models \neg\varphi$ or $v \not\models \neg\psi$. But this last situation holds iff $v \models \varphi$ or $v \models \psi$. This completes the proof of part (a). ∎

Lemma 2.1.1 says, among other things, that the formula *true* is always true and *false* is always false. This is certainly the intent! Similarly, it says that $\varphi \Leftrightarrow \psi$ is true exactly if φ and ψ have the same truth value: either they must both be true, or they must both be false. It also says that if $\varphi \Rightarrow \psi$ is true, then if φ is true, then ψ is true. Put another way, it says that the truth of φ implies the truth of ψ. Again, this seems consistent with our interpretation of implication. However, notice that our view of $\varphi \Rightarrow \psi$ as an abbreviation for $\neg\varphi \vee \psi$ guarantees that $\varphi \Rightarrow \psi$ will automatically be true if φ is false. This may seem counterintuitive. There has been a great deal of discussion regarding the reasonableness of this definition of $\Rightarrow$; alternative logics have been proposed that attempt to retain the intuition that $\varphi \Rightarrow \psi$ is true if the truth of φ implies the truth of ψ without automatically making $\varphi \Rightarrow \psi$ true if φ is false. I use the standard definition in this book because it has proved so useful; references for alternative approaches are provided in the notes at the end of the chapter. One important thing to remember (perhaps the most important thing, as far as the proofs in this book are concerned) is that when trying to show that $v \models \varphi \Rightarrow \psi$ for some valuation v, then we can assume without loss of generality that $v \models \varphi$, and then try to show $v \models \psi$; for if $v \models \neg\varphi$, then $v \models \varphi \Rightarrow \psi$ is vacuously true.

As we have seen, a formula such as *true* is true under every truth assignment. A formula that is true under every truth assignment is said to be a *tautology*, or *valid*. Other valid formulas include $(p \wedge q) \Leftrightarrow (q \wedge p)$, and $p \Leftrightarrow \neg\neg p$. The first one says that the truth value of a conjunction is independent of the order the conjuncts are taken; the second says that two negations cancel each other out. A formula that is true under some truth assignment is said to be *satisfiable*. It is easy to see that φ is valid if and only if $\neg\varphi$ is not satisfiable (Exercise 2.2).

One last definition for now: A set Σ of formulas *entails* a formula φ if every truth assignment that makes the formulas in Σ true also makes φ true. Note that a formula φ is valid iff $\emptyset$ (the empty set of formulas) entails φ. I can finally say in what sense the original argument is legitimate: $\{p \Rightarrow (r \Rightarrow q), p \wedge r\}$ entails q (Exercise 2.3).

2.2 A Modal Logic of Knowledge

I now move beyond propositional logic to a logic that allows reasoning about uncertainty within the logic. Perhaps the simplest kind of reasoning about uncertainty involves reasoning about whether certain situations are possible or impossible. I start with a logic of knowledge that allows just this kind of reasoning.

2.2.1 Syntax and Semantics

The syntax of propositional modal logic is just a slight extension of that for propositional logic. As an example, consider a propositional modal logic for reasoning about the knowledge of n agents. We start with a nonempty set Φ of primitive propositions, and close off under negation and conjunction. In addition, we have modal operators $K_1, \ldots, K_n$, one for each agent. We close off under application of the modal operators, so that if φ is a formula, so is $K_i\varphi$. Depending on context, we may want to view K_i as standing for either knowledge or belief. Although $K_i\varphi 4$ is typically read "agent i knows φ", in some contexts it may be more appropriate to read it as "agent i believes φ". Let $\mathcal{L}_n^K(\Phi)$ be the language consisting of all formulas that can be built up this way; for notational convenience, we often suppress the Φ. The subscript n denotes that there are n agents; I typically omit the subscript if $n = 1$. I also make use of abbreviations such as $\vee$, $\Rightarrow$, and $\Leftrightarrow$, just as in propositional logic.

We can express quite complicated statements in a straightforward way in this language. For example, the formula

$$K_1K_2p \wedge \neg K_2K_1K_2p$$

says that agent 1 knows that agent 2 knows p, but agent 2 does not know that agent 1 knows that agent 2 knows p. More colloquially, if I am agent 1 and you are agent 2, this can be read as "I know that you know p, but you don't know that I know that you know it."

Notice that possibility is the dual of knowledge. Agent i considers φ possible exactly if he does not know $\neg\varphi$. This situation can be described by the formula $\neg K_i\neg\varphi$. A statement such as "Alice does not know whether it is sunny in San Francisco" means that Alice considers possible both that it is sunny in San Francisco and that it is not sunny in San Francisco. This can be expressed by formula $\neg K_A\neg p \wedge \neg K_A\neg(\neg p)$, if p stands for "it is sunny in San Francisco".

How do we decide whether a formula in modal logic is true? This is where the notion of *possible worlds* comes in. The intuition is that, besides the true state of affairs, there are a number of other states of affairs, or "worlds", that an agent considers possible. The set of worlds that an agent considers possible can be viewed as a qualitative measure of her uncertainty. The more worlds she considers possible, the more uncertain she is as to the true state of affairs, and the less she knows. An agent *knows* a fact φ if φ is true in all the worlds she considers possible. For example, Alice may be walking on the streets in San Francisco on a sunny day, but have no information at all about the weather in London. Thus, in all the worlds that she considers possible, it is sunny in San Francisco. (I am implicitly assuming here that Alice does not consider it possible that she is hallucinating and in fact it is raining heavily in San Francisco.) On the other hand, since Alice has no information about the weather in London, there is a world she considers possible in which it is sunny in London, and another in which it is raining in London. Thus, Alice knows that it is sunny in San Francisco, but does not know whether it is sunny in London.

In a situation such as a poker game, these "possible worlds" have a concrete interpretation: they are simply all the possible ways the cards could have been distributed among the players. (Actually, this is a very simple model of the possible worlds in a poker game, that leaves out a number of potentially very relevant details. For example, it does not take into account whether a player is likely to bluff and how good a player may be. Nevertheless, it suffices for my purposes here.) Players may acquire additional information in the course of the play of the game. This additional information allows them to eliminate some of the worlds they consider possible. At some point Alice might know that Bob holds the ace of spades, since in all worlds (distributions of cards among players) that she currently considers possible, Bob holds the ace of spades. Note here that possibility is viewed as the dual of knowledge. Intuitively, the more worlds an agent considers possible, the greater her uncertainty, and the less she knows.

These intuitions regarding possible worlds can be formalized by means of *(Kripke) structures*. Suppose for simplicity we start with just one agent (and write $K\varphi$ rather than $K_1\varphi$). A *simple (epistemic) structure* M (over Φ) is a triple (W, w_0, π), where W can be thought of as the set of states of affairs or worlds the agent considers possible, w_0 is the actual world, and π is an *interpretation*, a function that associates with each world in $W \cup \{w_0\}$ a truth assignment to the primitive propositions. That is, $\pi(w)(p) \in \{\mathbf{true}, \mathbf{false}\}$ for each primitive proposition $p \in \Phi$ and world $w \in W \cup \{w_0\}$. Notice that I am not identifying a world with a truth assignment. There may be two worlds associated with the same truth assignment; that is, we may have $\pi(w) = \pi(w')$ for $w \neq w'$. This amounts to saying that there may be more to a world than what can be described by the primitive propositions. For the simple logic I am about to present, it would actually be safe to identify worlds with truth assignments (and thus "combine" two worlds that were associated with the same truth assignment), but with multiple agents, or even in somewhat more sophisticated logics involving just one agent, this cannot be done.

The set W of worlds in a simple structure can be empty. This amounts to the agent not considering any worlds as possible. A simple structure is *consistent* if W is nonempty (which implies, as we shall see, that the agent's beliefs are consistent), and *reliable* if $w_0 \in W$. A reliable structure is one where the agent considers the actual world to be possible (so his "knowledge" is reliable).

In propositional logic, a formula is true or false given a valuation. In the possible-worlds approach, of which this is an example, the truth of a formula depends on the world. A primitive proposition such as p may be true in one world and false in another. Thus, we define truth relative to a world in a structure, writing $(M, w) \models \varphi$, which is read "φ is true in world w of structure M". We define $\models$ by induction on the structure of formulas:

$(M, w) \models p$ (for a primitive proposition $p \in \Phi$) iff $\pi(w)(p) = \mathbf{true}$

$(M, w) \models \varphi \wedge \varphi'$ iff $(M, w) \models \varphi$ and $(M, w) \models \varphi'$

$(M, w) \models \neg\varphi$ iff $(M, w) \not\models \varphi$

$(M, w) \models K\varphi$ iff $(M, w') \models \varphi$ for all $w' \in W$.

The first three clauses are just what we would expect from propositional logic; the last captures the intuition that the agent knows φ if φ is true in all the worlds the agent considers possible.

Notice that in a simple structure, the truth of a formula of the form $K\varphi$ depends only on the set W of possible worlds, and not on the actual world. That is, if $(M, w) \models K\varphi$ for some world w (in $W \cup \{w_0\}$), then $(M, w') \models K\varphi$ for all worlds w'. The same is easily seen to be true for any *Boolean combination* of such formulas, that is, the result of combining such formulas using $\wedge$ and $\neg$ (Exercise 2.4).

As an example of a simple structure, suppose $W = \{w_1, w_2\}$ and we have two primitive propositions, p and q. Think of p as standing for "it is sunny in San Francisco" and q as standing for "it is raining in London". Suppose π is such that p is true at w_1 and w_2 and false at w_0 (that is, $\pi(w_0)(p) = \mathbf{true}$ and $\pi(w_1)(p) = \pi(w_2)(p) = \mathbf{false}$) and q is true at w_0 and w_1 and false at w_2. Then we have $(M, w_0) \models \neg p \wedge Kp$, that is, the agent believes that it is sunny in San Francisco but is mistaken (since the agent is mistaken here, "believes" seems like a more appropriate reading than "knows"). We also have $(M, w_0) \models q \wedge \neg Kq \wedge \neg K\neg q$—it is in fact raining in London, but the agent does not know whether or not it is raining. Put another way, the agent considers it possible that it is raining in London and also considers it possible that it is not raining in London; that is precisely the effect of having both w_1 and w_2 in W. Note that this structure is not safe, since $w_0 \notin W$. As we shall see, it is exactly because it is not safe that the agent can have mistaken beliefs.

While simple structures are just that—simple—they lack some expressive power. In simple structures, it is implicitly assumed that the set of worlds the agent considers possible in world w is the same as the set of worlds the agent considers possible in world w', even if $w \neq w'$. In general, this is clearly inappropriate. The set of worlds the agent considers possible when it is raining is clearly different from the set of worlds the agent considers possible when it is sunny.

The notion that the set of worlds considered possible is independent of the actual world may seem less objectionable if we think of W as the set of worlds that the agent considers possible given some fixed information (or set of observations and perceptions). Then our implicit assumption amounts to saying that the set of worlds the agent considers possible is determined by her "internal state"—roughly speaking, what she has seen and heard thus far, together with her genetic makeup and so on. The impact of the external world on the set of worlds she considers possible is summarized by her internal state. I return to this viewpoint in the Section 8.1.

Nevertheless, there are times when we want the set of worlds that the agent considers possible to depend on the actual world. This can be done in a straightforward way. Define a *Kripke structure* M for one agent to be a tuple $(W, \mathcal{K}, \pi)$, where $\mathcal{K}$ is a *binary relation on* W, that is, a subset of $W \times W$. Intuitively, $(w, w') \in \mathcal{K}$ if the agent considers w' a possible world in world w. In a Kripke structure there is no distinguished "actual world" (although we could add one if desired). Define $\mathcal{K}(w) = \{w' : (w, w') \in \mathcal{K}\}$; then $\mathcal{K}(w)$ describes the set of worlds that the agent considers possible in world w. Then define

$$(M, w) \models K\varphi \text{ iff } (M, w') \models \varphi \text{ for all } w' \in \mathcal{K}(w)$$

Notice that simple structures can be viewed as Kripke structures where $\mathcal{K}(w) = \mathcal{K}(w')$ for $w, w' \in W$.

Once we model the agent's possibilities by means of a binary relation, we can consider the effect of putting various constraints on that relation. For example, if the agent always considers the actual world possible, then we would have $(w, w) \in \mathcal{K}$ for all $w \in W$, i.e., $\mathcal{K}$ is *reflexive*. Similarly, we might want to assume that if (u, v) and (v, w) are both in $\mathcal{K}$ (so that v is considered possible in world u, and w is considered possible in v) then $(u, w) \in \mathcal{K}$ (so that w is considered possible in u). This just says that $\mathcal{K}$ is *transitive*. There are many other constraints that could be placed on $\mathcal{K}$. I mention three here.

- $\mathcal{K}$ is *Euclidean* if $(u,v),(u,w) \in \mathcal{K}$ implies $(v,w) \in \mathcal{K}$, for all $u,v,w \in W$.
- $\mathcal{K}$ is *symmetric* if $(u,v) \in \mathcal{K}$ implies that $(v,u) \in \mathcal{K}$ for all $u,v,\in W$.
- $\mathcal{K}$ is *serial* if for all $w \in W$, there is some $w' \in W$ such that $(w,w') \in \mathcal{K}$.
- $\mathcal{K}$ is an *equivalence relation* if it is reflexive, symmetric, and transitive. It is easy to see that this is equivalent to being reflexive, Euclidean, and transitive (Exercise 2.5).

There are many applications for which it is convenient to think of the $\mathcal{K}$ relation as being an equivalence relation, or at least Euclidean and transitive. In particular, we can capture the intuition behind the notion of possibility in simple structures—namely, that the set of worlds that the agent considers possible depends only on the agent's internal state—by taking the $\mathcal{K}$ relation to be transitive and Euclidean. This guarantees that the set of worlds that the agent considers possible is the same in all worlds that she considers possible. In fact, a Kripke structure for one agent where the $\mathcal{K}$ relation is Euclidean and transitive is essentially a union of simple structures. If we further assume that $\mathcal{K}$ is serial, then we get a union of consistent structures, and if we further assume that $\mathcal{K}$ is reflexive (so that $\mathcal{K}$ is in fact an equivalence relation), we get a union of reliable structures. This is all made precise in the following lemma.

Lemma 2.2.1 *Suppose $M = (W, \mathcal{K}, \pi)$ is a Kripke structure such that $\mathcal{K}$ is Euclidean and transitive and $w \in W$.*

(a) For all $w' \in \mathcal{K}(w)$, we have $\mathcal{K}(w') = \mathcal{K}(w)$.

(b) Let $M(w)$ be the simple structure $(\mathcal{K}(w), w, \pi')$, where π' is π restricted to $\mathcal{K}(w)$. (This means that $\pi(w')(p) = \pi'(w')(p)$ for all $w' \in \mathcal{K}(w)$ and all primitive propositions $p \in \Phi$.) Then for all $w' \in \mathcal{K}(w) \cup \{w\}$ and all formulas φ, we have $(M,w') \models \varphi$ if and only if $(M(w),w') \models \varphi$.

(c) If $\mathcal{K}$ is serial, then $M(w)$ is consistent; if $\mathcal{K}$ is reflexive, then $M(w)$ is reliable.

Proof See Exercise 2.6. ▮

Of course, this framework can be generalized easily to multiple agents. We simply have one possibility relation for each agent. In particular, a

Kripke structure M for n agents is a tuple $(W, \mathcal{K}_1, \ldots, \mathcal{K}_n, \pi)$, where each $\mathcal{K}_i$ is a binary relation on W. Then we have

$$(M, w) \models K_i\varphi \text{ iff } (M, w') \models \varphi \text{ for all } w' \in \mathcal{K}_i(w). \tag{2.1}$$

One of the advantages of a Kripke structure is that it can be viewed as a labeled graph, that is, a set of labeled nodes connected by directed, labeled edges. The nodes are the states of W; the label of state $w \in W$ describes which primitive propositions are true and false at w; and there is an edge from w to w' labeled i exactly if $(w, w') \in \mathcal{K}_i$. The graphical viewpoint makes it easier to see the connection between worlds. This is illustrated by the following example.

Suppose that a deck consists of three cards labeled A, B, and C. Agents 1 and 2 each get one of these cards; the third card is left face down. A possible world is characterized by describing the cards held by each agent. For example, in the world (A, B), agent 1 holds card A and agent 2 holds card B (while card C is face down). There are clearly six possible worlds: (A, B), (A, C), (B, A), (B, C), (C, A), and (C, B). Moreover, it is clear that in a world such as (A, B), agent 1 thinks two worlds are possible: (A, B) itself and (A, C). Agent 1 knows that he has card A, but considers it possible that agent 2 could hold either card B or card C. Similarly, in world (A, B), agent 2 also considers two worlds: (A, B) and (C, B). In general, in a world (x, y), agent 1 considers (x, y) and (x, z) possible, while agent 2 considers (x, y) and (z, y) possible, where z is different from both x and y.

From this description, we can easily construct the $\mathcal{K}_1$ and $\mathcal{K}_2$ relations. It is easy to check that they are equivalence relations. This is because an agent's knowledge is determined by the information he has, namely, the card he is holding. (Considerations similar to these lead to the use of equivalence relations in many examples involving knowledge.) The structure is described in the Figure 2.1 below, where, since the relations are equivalence relations, I omit the self loops and the arrows on edges for simplicity (if there is an edge from state w to state w', there is bound to be an edge from w' to w as well by symmetry).

The example points out the need for having worlds that an agent does not consider possible included in the structure. For example, in the world (A, B), agent 1 knows perfectly well that the world (B, C) cannot be the case (after all, agent 1 knows perfectly well that his own card is an A). Nevertheless, because agent 1 considers it possible that agent 2 considers it possible that (B, C) is the case, so we must include (B, C) in the structure. This is captured in the structure by the fact that there is no edge from (A, B) to (B, C) labeled 1, but there is an edge labeled 1 to (A, C), from

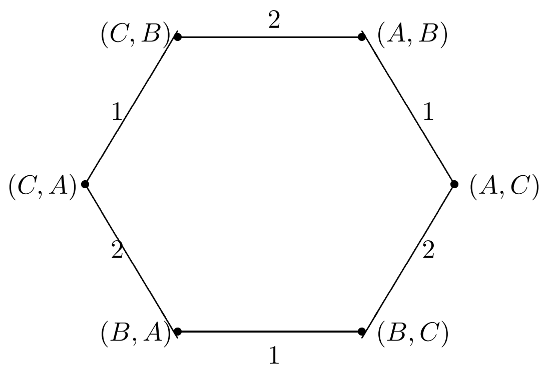

Figure 2.1: The Kripke structure describing a simple card game

which there is an edge labeled 2 to (B, C). Thus, in (A, B) agent 1 considers it possible that (A, C) is the case, and in (A, C), agent 2 considers it possible that (B, C) is the case.

I still have not discussed the language to be used in this example. Since we are interested in reasoning about the cards held by agents 1 and 2, it seems reasonable to have primitive propositions of the form $1A$, $2A$, $2B$, and so on, which are to be interpreted as "agent 1 holds card A", "agent 2 holds card A", "agent 2 holds card B", and so on. Given this reading, we can define π in the obvious way; let M_c be the Kripke structure describing this card game. Then, for example, we have $(M_c, (A, B)) \models 1A \wedge 2B$. I leave it to the reader to check that we also have $(M_c, (A, B)) \models K_1(2B \vee 2C)$, which expresses the fact that if agent 1 holds an A, then he knows that agent 2 holds either B or C. Similarly, we have $(M_c, (A, B)) \models K_1 \neg K_2(1A)$: agent 1 knows that agent 2 does not know that he holds an A.

This example shows that the possible-worlds semantics for knowledge formalized in (2.1) does capture some of the intuitions we naturally associate with the word "knowledge". However, this is far from a complete justification for reading $K_i\varphi$ as "agent i knows φ". Further justification for this semantics can be found in Chapter 8, where I show that this interpretation of knowledge—including the assumption that the $\mathcal{K}_i$'s are equivalence relations—is useful and reasonably appropriate for reasoning about multi-agent systems. In the next section, I provide an another justification for this semantics by characterizing its properties *axiomatically*.

2.2.2 Properties of Knowledge

One way to assess the reasonableness of the semantics of knowledge as truth in all worlds the agent considers possible is to try to characterize its properties. I start with simple structures. A formula φ is *valid in simple structure* $M = (W, w_0, \pi)$, denoted $M \models \varphi$, if $(M, w) \models \varphi$ for all w in $W \cup \{w_0\}$; φ is *satisfiable in a simple structure* M if $(M, w) \models \varphi$ for some world w in M. A formula φ is *valid in simple structures* if φ is valid in every simple structure; φ is *satisfiable in simple structures* if φ is satisfiable in some simple structure.

One important property of the definition of knowledge is that each agent knows all the logical consequences of his knowledge. If an agent knows φ and knows that φ implies ψ, then both φ and $\varphi \Rightarrow \psi$ are true at all worlds he considers possible. Thus ψ must be true at all worlds that the agent considers possible, so he must also know ψ. It follows that

$$\models (K\varphi \wedge K(\varphi \Rightarrow \psi)) \Rightarrow K\psi$$

is valid in simple structures. This property is called the *Distribution Axiom* since it allows us to distribute the K operator over implication. It suggests that the definition of K assumes that agents are quite powerful reasoners.

Further evidence of this comes from the fact that agents know all the formulas that are valid in a given structure. If φ is true at all the possible worlds of structure M, then φ must be true at all the worlds that an agent considers possible in any given world in M, so it must be the case that $K\varphi$ is true at all possible worlds of M. More formally, we have the following *Rule of Knowledge Generalization*:

$$\text{For all simple structures } M \text{, if } M \models \varphi \text{ then } M \models K\varphi.$$

Note that from this we can deduce that if φ is valid, then so is $K\varphi$. This rule is very different from the formula $\varphi \Rightarrow K\varphi$, which says that if φ is true then the agent knows it. An agent does not necessarily know all things that are true. (For example, it may be the case that it is sunny in San Francisco, and Alice does not know this.) However, agents do know all valid formulas. Intuitively, these are the formulas that are *necessarily* true, as opposed to the formulas that just happen to be true at a given world. It requires a rather powerful reasoner to know all necessarily true facts.

In simple structures, agents can also do introspection regarding their knowledge. They know what they know and what they do not know. That is, the following two formulas are valid in simple structures:

$$K_i\varphi \Rightarrow K_iK_i\varphi,$$
$$\neg K_i\varphi \Rightarrow K_i\neg K_i\varphi.$$

Imagine that an agent has the collection of all facts that he knows written in a database. Then the first of these properties, called the *Positive Introspection Axiom*, says that the agent can look at this database and see what facts are written there, so that if he knows φ, then he knows that he knows it (and thus the fact that he knows φ is also written in his database). The second property, called the *Negative Introspection Axiom*, says that he can also look over his database to see what he doesn't know. Thus, if he doesn't know φ, so that φ is not written in his database, he knows that φ is not written there, so that he knows that he doesn't know φ.

It is possible that for K*false* to hold in a simple structure, but only if the set of possible worlds is empty. In the simple structure $M = (\emptyset, w_0, \pi)$, we have $(M, w_0) \models K\varphi$ for all φ (and, in particular, for $\varphi =$ *false*). However, if the set of possible worlds is nonempty, then this somewhat anomalous phenomenon cannot occur. It is easy to see that if M is a consistent structure, then $M \models \neg K$*false*.

In reliable structures, an even stronger property holds: what an agent knows to be true is in fact true; more precisely, $K\varphi \Rightarrow \varphi$ is valid in reliable structures. This property, occasionally called the *Knowledge Axiom* or the *veridicality* property (since "veridical" means "truthful"), has been taken by philosophers to be the major one distinguishing knowledge from *belief*. Although you may have false beliefs, you cannot know something that is false. In nonreliable structures, the knowledge axiom may fail. For example, consider the simple structure $M = (\{w\}, w', \pi)$, where (according to π) p is true at w and false at w'. Then $(M, w') \models \neg p \wedge Kp$.

The preceding discussion is summarized in the following theorem.

Theorem 2.2.2 *For all simple structures M,*

(a) $M \models (K\varphi \wedge K(\varphi \Rightarrow \psi)) \Rightarrow K\psi$,

(b) *if* $M \models \varphi$ *then* $M \models K\varphi$,

(c) $M \models K\varphi \Rightarrow KK\varphi$,

(d) $M \models \neg K\varphi \Rightarrow K\neg K\varphi$.

If M is consistent, then

(e) $M \models \neg K$*false*.

If M is reliable, then

(f) $M \models K\varphi \Rightarrow \varphi$.

Proof Suppose $M = (W, w_0, \pi)$.

(a) If $(M,w) \models K\varphi \wedge K(\varphi \Rightarrow \psi)$, then for all worlds $w' \in W$, we have both that $(M,w') \models \varphi$ and $(M,w') \models \varphi \Rightarrow \psi$. By the definition of $\models$, we have that $(M,w') \models \psi$ for all such $w' \in W$, and therefore $(M,w) \models K\psi$.

(b) If $M \models \varphi$ then $(M,w') \models \varphi$ for all worlds w' in $W \cup \{w_0\}$. It immediately follows that $(M,w) \models K\varphi$ for all $w \in W \cup \{w_0\}$.

(c) Suppose $(M,w) \models K\varphi$. Then $(M,w') \models \varphi$ for all $w' \in W$. It follows from the definition of $\models$ that $(M,w') \models K\varphi$ for all $w' \in W$. Thus, $(M,w) \models K\varphi$.

(d) Suppose $(M,w) \models \neg K\varphi$. Then $(M,w'') \models \neg\varphi$ for some $w'' \in W$. It follows from the definition of $\models$ that $(M,w') \models \neg K\varphi$ for all $w' \in W$. Thus, $(M,w) \models K\neg K\varphi$.

(e) Since M is consistent, there is some world $w' \in W$. Clearly, $(M,w') \models \neg \mathit{false}$. Thus, for all $w \in W \cup \{w_0\}$, we have $(M,w) \models \neg K \mathit{false}$.

(f) If $(M,w) \models K\varphi$, then for all $w' \in W$, we have $(M,w') \models \varphi$. Since M is reliable, we must have $w \in W$. Thus, $(M,w) \models \varphi$. ∎

As we shall see in Section 2.2.3, the properties described in Theorem 2.2.2 essentially characterize knowledge in simple structures.

Moving now to Kripke structures, similar properties hold. Indeed, once we have the flexibility of considering a binary relation, we can see exactly what properties of the relation give us each of the properties considered in Theorem 2.2.2. We can define validity and satisfiability in Kripke structures analogously to simple structures. Thus, for example, if M is a Kripke structure, then φ is *valid in M*, denoted $M \models \varphi$, if $(M,w) \models \varphi$ for all worlds w in M. In general, once we have a notion of a structure (like a simple structure or a Kripke structure), we can talk about a formula being valid or satisfiable in a structure, or with respect to a class of structures. If $\mathcal{N}$ is a class of structures (for example, the class of all Kripke structures, or the class of all Kripke structures where the $\mathcal{K}_i$ relations are equivalence relations, or the class of all reliable structures), I use $\mathcal{N} \models \varphi$ to denote that φ is valid in every structure in $\mathcal{N}$.

Theorem 2.2.3 *Suppose $M = (W, \mathcal{K}_1, \ldots, \mathcal{K}_n, \pi)$ is a Kripke structure. Then for all agents i, we have:*

(a) $M \models (K_i\varphi \wedge K_i(\varphi \Rightarrow \psi)) \Rightarrow K_i\psi$,

(b) if $M \models \varphi$ then $M \models K_i\varphi$,

(c) *if* $\mathcal{K}_i$ *is transitive, then* $M \models K_i\varphi \Rightarrow K_iK_i\varphi$,

(d) *if* $\mathcal{K}_i$ *is Euclidean, then* $M \models \neg K_i\varphi \Rightarrow K_i\neg K_i\varphi$,

(e) *if* $\mathcal{K}_i$ *is serial, then* $M \models \neg K_i$*false*,

(f) *if* $\mathcal{K}_i$ *is reflexive, then* $M \models K_i\varphi \Rightarrow \varphi$.

Proof See Exercise 2.7. ∎

Theorem 2.2.3 makes it clear how the various properties we are considering are related to properties of the possibility relation. For example, if $\mathcal{K}_i$ is transitive, then we get positive introspection; if $\mathcal{K}_i$ is Euclidean, then we get negative introspection; and if $\mathcal{K}_i$ is reflexive, then we get veridicality. However, two properties that seem forced on us by the possible worlds approach are the Distribution Axiom and the Rule of Knowledge Generalization. These properties hold no matter what we assume of the $\mathcal{K}_i$ relations. To the extent that we think of knowledge as something acquired by agents through some reasoning process, these properties suggest that we must think in terms of idealized agents who can do perfect reasoning. Clearly, this assumption is not always reasonable. Of course, there are other intuitions behind the notion of knowledge besides just the idea that knowledge is something gained via a reasoning process. In Chapter 8, knowledge is ascribed to agents in a multi-agent system based on their current information, as encoded in their state. This notion of knowledge explicitly does not take computation into account. While it is useful in many contexts, not surprisingly, it is no longer adequate when we need to think in terms of processes computing their knowledge.

2.2.3 Axiomatizing Knowledge

Are there other important properties of the possible-worlds definition of knowledge that I have not yet mentioned? In a precise sense, the answer is no. All the properties of knowledge (at least, in the propositional case) follow from those we have already seen. To formalize this, we need to consider the notion of *provability* and that of a *sound and complete axiomatization.*

An *axiom system AX* consists of a collection of *axioms* and *rules of inference*. An axiom is just a formula, while a rule of inference has the form "from $\varphi_1, \dots, \varphi_k$ infer ψ, where $\varphi_1, \dots, \varphi_k, \psi$ are formulas. An inference rule can be viewed as a method for inferring new formulas from old ones. A *proof* in AX consists of a sequence of steps, each of which is either an instance of an axiom in AX, or follows from previous steps by an application of an inference rule. More precisely, if "from $\varphi_1, \dots, \varphi_k$ infer ψ" is an inference rule, and the formulas $\varphi_1, \dots, \varphi_k$ have appeared earlier in the proof, then ψ follows by an application of this inference rule. A proof is a *proof of the formula* φ if the last step of the proof is φ. A formula φ is *provable in* AX, denoted $AX \vdash \varphi$, if there is a proof of φ in AX.

Suppose that we have a class $\mathcal{N}$ of structures and a language $\mathcal{L}$, such that a notion of validity is defined for the formulas in $\mathcal{L}$ with respect to the structures in $\mathcal{N}$. For example, if $\mathcal{L} = \mathcal{L}^{Prop}$, then $\mathcal{N}$ could consist of all truth assignments; if $\mathcal{L} = \mathcal{L}_n^K$, then $\mathcal{N}$ could be the class of Kripke structures. An axiom system AX is *sound* for $\mathcal{L}$ with respect to $\mathcal{N}$ if every formula of $\mathcal{L}$ that is provable in AX is valid in every structure in $\mathcal{N}$. AX is *complete* for $\mathcal{L}$ with respect to $\mathcal{N}$ if every formula in $\mathcal{L}$ that is valid in every structure in $\mathcal{N}$ is provable in AX. AX can be viewed as characterizing the class $\mathcal{N}$ if it provides a sound and complete axiomatization of that class; notationally, this amounts to saying that for all formulas φ, we have $AX \vdash \varphi$ iff $\mathcal{N} \models \varphi$. Soundness and completeness provide a connection between the *syntactic* notion of provability and the *semantic* notion of validity.

There are well-known sound and complete axiomatizations for propositional logic (see the notes at the end of the chapter). Here I want to focus on axioms for knowledge, so I just take for granted that we have access to all the tautologies of propositional logic. Consider the following collection of axioms and inference rules.

Prop. All substitution instances of tautologies of propositional calculus

K1. $(K_i\varphi \wedge K_i(\varphi \Rightarrow \psi)) \Rightarrow K_i\psi$ (Distribution Axiom).

K2. $K_i\varphi \Rightarrow \varphi$ (Knowledge Axiom).

K3. $\neg K_i \mathit{false}$ (Consistency Axiom).

K4. $K_i\varphi \Rightarrow K_iK_i\varphi$ (Positive Introspection Axiom).

K5. $\neg K_i\varphi \Rightarrow K_i\neg K_i\varphi$ (Negative Introspection Axiom).

MP. From φ and $\varphi \Rightarrow \psi$ infer ψ (Modus ponens).

Gen. From φ infer $K_i\varphi$ (Knowledge Generalization).

Technically, Prop and K1–K5 are axiom schemes, rather than single axioms. K1, for example, holds for all formulas φ and ψ and all agents $i = 1, \ldots, n$. Prop allows gives us all propositional tautologies "for free". A formula such as $K_1q \vee \neg K_1q$ is an instance of axiom Prop (since it is a substitution instance of the propositional tautology $p \vee \neg p$, obtained by substituting K_1q for p).

Historically, axiom K1 has been called **K**, K2 has been called **T**, K3 has been called **D**, K4 has been called **4**, and K5 has been called **5**. We get different modal logics by considering various subsets of these axioms. In the case of one agent, the system with axioms and rules Prop, K1, MP, and Gen has been called K. One approach to naming these logics is to name them after the significant axioms used. For example, the axiom system KD45 is the result of combining the axioms **K**, **D**, **4**, and **5** with Prop, MP, and Gen, while KT4 is the result of combining the axioms **K**, **T**, and **4** with Prop, MP, and Gen. Some of the axiom systems are commonly called by other names as well. The K is quite often omitted, so that KT becomes T, KD becomes D, and so on; KT4 has traditionally been called S4 and KT45 has been called S5. I stick with the traditional names here for those logics that have them, since they are in common usage, except that I use the subscript n to emphasize the fact that these are systems with n agents rather than only one agent. I occasionally omit the subscript if $n = 1$, in line with more traditional notation. In this book I focus mainly on $\mathrm{S5}_n$ and $\mathrm{KD45}_n$.

Philosophers have spent years arguing which of these axioms, if any, best captures the knowledge of an agent. I do not believe that there is one "true" notion of knowledge; rather, the appropriate notion depends on the application. For many applications, the axioms of S5 seem most appropriate (see Chapter 8), although philosophers have argued quite vociferously against them, particularly axiom K5. Rather than justify these axioms further, I focus here on the relationship between these axioms and the properties of the $\mathcal{K}_i$ relation.

Theorem 2.2.3 suggests that there is a connection between K4 and transitivity of $\mathcal{K}_i$, K5 and the Euclidean property, K3 and seriality, and K2 and reflexivity. This connection is a rather close one. To formalize it, let $\mathcal{M}_n^r$ (resp., $\mathcal{M}_n^{rt}$; $\mathcal{M}_n^{rst}$; $\mathcal{M}_n^{elt}$; $\mathcal{M}_n^{et}$) be the class of all structures for n agents where the possibility relations are reflexive (resp., reflexive and transitive; reflexive, symmetric, and transitive; Euclidean, serial, and transitive; Euclidean and transitive).

Theorem 2.2.4 *For formulas in the language $\mathcal{L}^K$:*

(a) K_n *is a sound and complete axiomatization with respect to* $\mathcal{M}_n$*,*

(b) T_n *is a sound and complete axiomatization with respect to* $\mathcal{M}_n^r$*,*

(c) $\mathrm{S4}_n$ *is a sound and complete axiomatization with respect to* $\mathcal{M}_n^{rt}$*,*

(d) $\mathrm{K45}_n$ *is a sound and complete axiomatization with respect to* $\mathcal{M}_n^{et}$*.*

(e) $\mathrm{KD45}_n$ *is a sound and complete axiomatization with respect to* $\mathcal{M}_n^{elt}$*,*

(f) $\mathrm{S5}_n$ *is a sound and complete axiomatization with respect to* $\mathcal{M}_n^{rst}$*.*

Proof Soundness follows immediately from Theorem 2.2.3. The proof of completeness is beyond the scope of this book. ∎

This theorem tells us that, for example, the axiom system $\mathrm{S5}_n$ completely characterizes propositional modal reasoning in $\mathcal{M}_n^{rst}$. There are no "extra" properties (beyond those that can be proved from $\mathrm{S5}_n$) that are valid in structures in $\mathcal{M}_n^{rst}$. There is an analogous theorem for simple structures.

Theorem 2.2.5 *For formulas in the language $\mathcal{L}^K$:*

(a) K45 *is a sound and complete axiomatization with respect to the class of simple structures,*

(b) KD45 *is a sound and complete axiomatization with respect to the class of consistent simple structures,*

(c) S5 *is a sound and complete axiomatization with respect to the class of reliable structures.*

2.2.4 A Digression: Events vs. Propositions

I have presented logic here using the standard logician's approach (which is also common in the philosophy and AI communities): we start with a language and assign formulas in the language truth values in a semantic structure. In other communities (such as the statistics and economics communities), it is more standard to dispense with language, and work directly with semantic structures.

For example, economists use *Aumann structures*, which can be viewed as Kripke structures without an interpretation π. That is, an Aumann structure M for n agents is a tuple $(W, \mathcal{K}_1, \ldots, \mathcal{K}_n)$. (Economists typically restrict to Aumann structures where the $\mathcal{K}_i$ relation is an equivalence relation, so that it partitions the set of worlds.) An Aumann structure is a special case of what is called a *frame* in the modal logic literature; a frame focuses on the accessibility relations, ignoring the interpretation. This means we cannot give formulas a truth value. Economists still want to talk about knowledge, but without formulas. To do this, they use a knowledge operator that works directly on sets of possible worlds.

To understand how this is done, it is best to start with the propositional operators, $\wedge$ and $\neg$. To each of these we can also attach an operation on sets of worlds. The operator corresponding to $\wedge$ is intersection, and the operator corresponding to $\neg$ is complementation. To understand the connection, imagine we have a Kripke structure $M = (W, \mathcal{K}_1, \ldots, \mathcal{K}_n, \pi)$ and two formulas φ and ψ. Let A and B be the sets of worlds in M where the formulas φ and ψ, respectively, are true. Then the set of worlds where $\varphi \wedge \psi$ is true is $A \cap B$, and the set of worlds where $\neg\varphi$ is true is $\overline{A}$ (where $\overline{A}$ denotes the complement of A). Thus, the operations of intersection and complementation are the semantic analogues of $\wedge$ and $\neg$.

What is the semantic analogue of K_i? Let $[\![\varphi]\!]_M = \{w : (M, w) \models \varphi\}$. Thus, $[\![\varphi]\!]_M$, called the *intension* of φ, is the set of worlds where φ is true. What is the connection between $[\![\varphi]\!]_M$ and $[\![K_i\varphi]\!]_M$? It is captured by the operator $\mathtt{K}_i$ on sets, which is defined so that $\mathtt{K}_i(U) = \{w : \mathcal{K}_i(w) \subseteq U\}$. The following theorem makes this precise.

Proposition 2.2.6 *For all formulas φ and ψ, we have*

(a) $[\![\varphi \wedge \psi]\!]_M = [\![\varphi]\!]_M \cap [\![\psi]\!]_M$,

(b) $[\![\neg\varphi]\!]_M = \overline{[\![\varphi]\!]_M}$,

(c) $[\![K_i\varphi]\!]_M = \mathtt{K}_i([\![\varphi]\!]_M)$.

Proof See Exercise 2.8. ∎

The most interesting clause in Proposition 2.2.6 is (c), which makes clear in what sense $\mathtt{K}_i$ is the semantic analogue of K_i. Not surprisingly, $\mathtt{K}_i$ satisfies many properties analogous to K_i.

Proposition 2.2.7 *For all Aumann structures $M = (W, \mathcal{K}_1, \ldots, \mathcal{K}_n)$, the following properties hold for all $U, V \subseteq W$ and all agents i:*

(a) $\mathtt{K}_i(U \cap V) = \mathtt{K}_i(U) \cap \mathtt{K}_i(V)$,

(b) if $\mathcal{K}_i$ is reflexive, then $\mathtt{K}_i(U) \subseteq U$,

(c) if $\mathcal{K}_i$ is transitive, then $\mathtt{K}_i(U) = \mathtt{K}_i(\mathtt{K}_i(U))$,

(d) if $\mathcal{K}_i$ is Euclidean, then $\overline{\mathtt{K}_i(U)} \subseteq \mathtt{K}_i(\overline{\mathtt{K}_i(U)})$.

Proof See Exercise 2.9. ∎

Part (a) is the semantic analogue of $K_i(\varphi \wedge \psi) \Leftrightarrow (K_i\varphi \wedge K_i\psi)$, which is easily seen to be valid in all Kripke structures (Exercise 2.10). Parts (b), (c), and (d) are the the semantic analogues of axioms K2, K4, and K5, respectively.

If we have a fixed structure in mind, there is certainly an advantage in working with Aumann structures rather than Kripke structures. We can dispense with the interpretation π and with defining $\models$, working directly with sets (events) rather than formulas.

So why do we bother with the overhead of syntax? Having a language has a number of advantages; I discuss three of them here.

The first is that there are times when it is useful to distinguish logically equivalent formulas. For example, in any given structure, it is not hard to check that the events corresponding to the formulas $K_i true$ and $K_i((p \Rightarrow q) \vee (q \Rightarrow p))$ are logically equivalent, since $(p \Rightarrow q) \vee (q \Rightarrow p)$ is a tautology. However, a computationally bounded agent may not recognize that $(p \Rightarrow q) \vee (q \Rightarrow p)$ is a tautology, and thus may not know it. Although all the semantics for modal logic that I consider in this book have the property that they do not distinguish logically equivalent formulas, it is possible to give semantics where such formulas are distinguished. (See the notes at the end of this chapter for some pointers to the literature.) This clearly would not be possible if we used a set-based approach.

The second advantage is that the structure of the syntax gives us a way to reason and carry out proofs. For example, many technical results proceed by induction on the structure of formulas. Similarly, formal axiomatic reasoning typically takes advantage of the syntactic structure of formulas.

The third advantage of using formulas is that it allows us to formulate notions in a structure-independent way. For example, economists are interested in notions of rationality, and would like to express rationality in terms of knowledge and belief. The description of what it means to be rational will thus be a formula, involving knowledge (and perhaps other operators; see the references in the notes for some discussion). Similarly, formulas also allow us to compare two or more structures that, intuitively, are "about" the same basic phenomena. For a simple example, consider the following two Kripke structures. In M_1, both agents know the true situation. There are two worlds in M_1: in one, p is true, both agents know it, know that they know it, and so on; in the other, p is false, both agents know it, know that they know it, and so on. In M_2, agent 1 knows the true situation, although agent 2 does not. There are also two worlds in M_2: in one, p is true and agent 1 knows it; in the other, p is false and agent 1 knows that it is false; agent 2 cannot distinguish these two worlds. Thus, $K_1p \vee K_1\neg p$ holds in every state of both M_1 and M_2, while $K_2p \vee K_2\neg p$ holds in both states of M_1, but not in both states of M_2. Using formulas allows us to compare M_1 and M_2. We cannot make an analogous comparison using events. The problem is that formulas such as p and K_1p are represented by totally different sets in M_1 and M_2. That is, the set of worlds where p or K_ip are true in M_1 and M_2 bear no relationship to each other. Nevertheless, we would like to be able to say that these sets correspond in some way. We cannot do that in any obvious way without invoking a language.

2.3 A Modal Logic of Relative Likelihood

Knowledge gives us a rather coarse-grained way of modeling uncertainty. It allows us to model the fact that one world may be considered possible while another is not, but it does not allow us to capture the fact that one world is considered more likely than another. In this section, I consider a qualitative approach to modeling uncertainty that lets us reason about relative likelihood. We again have possible worlds, but now we also order them according to likelihood.

2.3.1 From Likelihood on Worlds to Likelihood on Sets of Worlds

Let $\succeq$ be a reflexive and transitive relation on a set W of worlds. Technically, $\succeq$ is a *partial preorder*. It is *partial* because two worlds might be incomparable as far as $\succeq$ goes; that is, we may have neither $w \succeq w'$ nor $w' \succeq w$ for some worlds w and w'. It is a partial *preorder* rather than a partial *order* because it is not necessarily *antisymmetric*; we may have $w \preceq w'$ and $w' \preceq w$, even though w and w' are different worlds. I typically write $w \succeq w'$ rather than $(w, w') \in \succeq$. (It may seem strange to write $(w, w') \in \succeq$, but recall that $\succeq$ is just a binary relation.) I also write $w \succ w'$ if $w \succeq w'$ and it is not the case that $w' \succeq w$. The relation $\succ$ is the *strict* partial order *determined by* $\succeq$; it is strict because is irreflexive and transitive. An irreflexive and transitive relation must be antisymmetric (Exercise 2.11), so $\succ$ is an order rather than just a preorder.

Think of $\succeq$ as providing a likelihood ordering on the worlds in W. If $w \succeq w'$, then w is at least as likely as w'. Given this interpretation, the fact that $\succeq$ is assumed to be a partial preorder is easy to justify. Transitivity just says that if u is at least as likely as v, and v is at least as likely as w, then u is at least as likely as w; reflexivity just as says that world w is at least as likely as itself. The fact that $\succeq$ is partial allows for an agent that is not able to compare two worlds in likelihood. It is sometimes (but not always!) reasonable to further assume that $\succeq$ is *total*, that is, all worlds are comparable. This assumption makes some of the technical arguments in this section easier.

To take advantage of the additional structure provided by $\succeq$, it is useful to expand the language somewhat. Since we have added likelihood to the worlds, it seems that we should add likelihood to the language, to allow us to say "$\varphi \gg \psi$" which is read as "φ is more likely than ψ". But what exactly should this mean? Although having $\succeq$ in our semantic model allows us to say that one world is more likely than another, it does not immediately tell us how to say that a set of worlds is more likely than another set. But this is just what we need to make sense of "φ is more likely than ψ", since it corresponds to saying "the set of worlds where φ is true is more likely than the set where ψ is true".

How should we extend the likelihood ordering $\succeq$ on worlds to a likelihood ordering $\rhd$ on sets? We clearly want the ordering on sets to preserve the ordering on worlds, that is, we would expect $\{w\} \rhd \{w'\}$ to hold iff $w \succeq w'$. It also seems reasonable to expect that if every element in U is greater than every element in V (that is, if $u \succeq v$ for all $u \in U$ and $v \in V$), then $U \rhd V$. But these two properties do not uniquely determine $\rhd$. Indeed, a number of different approaches to defining an order on sets have been studied in the literature (see the notes at the end of the chapter); no one can lay claim to being the "right" one. I explore one general approach here, which has the advantage of being reasonably natural and of having some nice technical properties. (Other ways of defining likelihood directly on sets are studied in Chapter 3.) Roughly speaking, we take U to be more likely than V if for every world in V, there is a more likely world in U.

If $U, V \subseteq W$, I write $U \succeq^s V$ if for every world $v \in V$, there is a world $u \in U$ such that $u \succeq v$. (The superscript s in $\succeq^s$ is meant to emphasize that $\succeq^s$ is an ordering on *sets* of worlds.) It is easy to check that $\succeq^s$ as defined on finite sets is a partial preorder, that is, it is reflexive and transitive (Exercise 2.16) (and not necessarily anti-symmetric, so not necessarily a partial order). As we might expect, we have $u \succeq v$ iff $\{u\} \succeq^s \{v\}$, so the $\succeq^s$ relation on sets of worlds can be viewed as a generalization of the $\succeq$ relation on worlds.

In a similar spirit, define a relation $\succ^s$ on finite sets by taking $U \succ^s V$ to hold if U is nonempty, and for every world $v \in V$, there is a world $u \in U$ such that $u \succ v$. As the notation suggests, $\succ^s$ is a strict partial order on 2^W. However, note that the requirement that U must be nonempty in the definition of $U \succ^s V$ is necessary to ensure this; otherwise we would have $\emptyset \succ^s \emptyset$, so $\succ^s$ would not be irreflexive. Another strict partial order $\succ'$ on sets can be defined by taking $U \succ' V$ as an abbreviation for $U \succeq^s V$ and not($V \succeq^s U$). It is easy to see that $u \succ v$ if and only if $\{u\} \succ^s \{v\}$ if and only if $\{u\} \succ' \{v\}$. Thus, $\succ^s$ and $\succ'$ agree on singleton sets and extend the $\succ$ relation on worlds. It is not too hard to show that $\succ'$ and $\succ^s$ are in fact identical if the underlying relation $\succeq$ on worlds is total (Exercise 2.12). However, as the following example shows, $\succ^s$ and $\succ'$ are not identical in general.

Example 2.3.1 Suppose $W = \{w_1, w_2\}$, and $\succeq$ is such that w_1 and w_2 are incomparable. Then it is easy to see that $\{w_1, w_2\} \succ' \{w_1\}$. However, it is not the case that $\{w_1, w_2\} \succ^s \{w_1\}$, since there is no element of $\{w_1, w_2\}$ that is strictly more likely than w_1. ∎

Notice that I was careful to define $\succ$ only on finite sets. Finite sets are guaranteed to have *maximal* elements with respect to $\succ$ (and hence also with respect to $\succ'$ and $\succeq$). Indeed, given $u \in U \subseteq W$, we can always find $v \in U$ such that $v \succeq u$ and it is not the case that $v' \succ v$ for all $v' \in U$. The following example illustrates the importance of this.

Example 2.3.2 Let $V_\infty = \{w_0, w_1, w_2, \ldots\}$, and suppose that $\succeq$ is such that

$$w_0 \prec w_1 \prec w_2 \prec \ldots$$

Then it is easy to see that if we were to apply the definition of $\succ^s$ to infinite sets, then we would have $W_\infty \succ^s W_\infty$, and $\succ$ would not be irreflexive. ∎

For the remainder of this section, I assume that the set W of possible worlds is finite. This simplifies the exposition. Exercise 2.13 discusses how all of the definitions given can be modified to deal with the case that W is infinite.

2.3.2 A Logic for Reasoning About Relative Likelihood

I now consider a logic for reasoning about relative likelihood. I again start with a nonempty set Φ of primitive propositions, and close off under negation and conjunction; now I also close off under the application of the binary operators $\gg_1, \ldots, \gg_n$. Thus, if φ and ψ are formulas, then so is $\varphi \gg \psi$, read "agent i considers φ to be more likely than ψ". Thus, a statement like "agent i thinks this thing is more likely mimsy than not" would be expressed as $q \gg \neg q$ (taking q to represent "this thing is mimsy", as before). Let $\mathcal{L}_n^{\gg}(\Phi)$ be the resulting language. (Again, I typically suppress Φ and omit the subscript if $n = 1$.)

For the semantics, just as in the case of knowledge, let's first consider the case where there is only one agent (so I write $\varphi \gg \psi$ rather than $\varphi \gg_1 \psi$). Define a *simple preferential structure* M (over Φ) to be a tuple $M = (W, \succeq, \pi)$, where W is a finite nonempty set of possible worlds, $\succeq$ is a partial preorder on W, and π is an interpretation.

As usual, the semantics of formulas in $\mathcal{L}^{\gg}$ in simple preferential structures is defined by induction on structure. For primitive propositions, conjunction, and negation, the definition is just as it was for simple epistemic structures, so I do not bother repeating it. The clause for $\gg$ is straightforward:

$$(M, w) \models \varphi \gg \psi \text{ iff } [\![\varphi]\!]_M \succ^s [\![\psi]\!]_M.$$

Just as in the case of simple epistemic structures, the truth of a formula of the form $\varphi \gg \psi$ does not depend on the actual world, since the preferential ordering $\succeq$ is assumed to be independent of the world. It is easy to check that if $(M,w) \models \varphi \gg \psi$ for some $w \in W$, then $(M,w') \models \varphi \gg \psi$ for all $w' \in W$.

From the definitions, it follows that $(M,w) \models \neg(\neg\varphi \gg \mathit{false})$ exactly if $[\![\neg\varphi]\!]_M = \emptyset$; this, in turn, is true exactly if $[\![\varphi]\!]_M = W$. Thus, $\neg(\neg\varphi \gg \mathit{false})$ is true exactly if φ is true at all the worlds in W. This means that we can view $\neg(\neg\varphi \gg \mathit{false})$ as a way of expressing $K\varphi$ in $\mathcal{L}^{\gg}$. Indeed, from here on, I take $K\varphi$ to be an abbreviation for $\neg(\neg\varphi \gg \mathit{false})$ when working in the language $\mathcal{L}^{\gg}$.

Just as in the case of knowledge, we can move from simple preferential structures to more general preferential structures, where there may be more than one agent and the preferential order may depend on the world. A *preferential structure* for n agents is a tuple $(W, O_1, \ldots, O_n, \pi)$, where, for each world $w \in W$, $O_i(w)$ is a pair $(W_{w,i}, \succeq_{w,i})$, where $W_{w,i} \subseteq W$ and $\succeq_{w,i}$ is a partial preorder on $W_{w,i}$. $W_{w,i}$. Intuitively, $W_{w,i}$ is the set of worlds that agent i considers possible at worlds w, and $\succeq_{w,i}$ is agent i's likelihood ordering at world w. Again, the semantics of formulas in $\mathcal{L}_n^{\gg}$ in preferential structures is defined by induction on the structure of formulas. $O_i(w)$ is used in the clause for $\gg_i$:

$$(M,w) \models \varphi \gg_i \psi \text{ iff } W_{w,i} \cap [\![\varphi]\!]_M \succ^s W_{w,i} \cap [\![\psi]\!]_M, \text{ where } O_i(w) = (W_{w,i}, \succeq).$$

Let $\mathcal{M}_n^{pref}$ denote the class of all preferential structures for n agents and let $\mathcal{M}^{pref}$ denote the class of simple preferential structures.

2.3.3 Properties of Relative Likelihood

The fact that $\succ^s$ is a strict partial order on sets of worlds that is derived from a partial preorder on worlds gives it special properties. These properties, in turn, are reflected in the axioms for relative likelihood. In this section, I examine these properties and their impact on the axiomatization.

As we have observed, $\succ^s$ is transitive. Another obvious property that it is *orderliness*: If $U \succ^s V$, then taking a superset of U or a subset of V preserves $\succ^s$. Formally, a relation $\rhd$ on 2^W is *orderly* if $U \rhd V$, $U' \supseteq U$, and $V' \subseteq V$ implies $U' \rhd V'$. Clearly $\succ^s$ is orderly.

Another property that $\succ^s$ has is the *union property*. A relation $\rhd$ *satisfies the union property* if $V_1 \rhd V_2$ and $V_1 \rhd V_3$ implies $V_1 \rhd (V_2 \cup V_3)$. Clearly the union property generalizes to arbitrary finite unions. In particular, this means that if $u \succ v_j$ for $j = 1, \ldots, N$, then $\{u\} \succ \{v_1, \ldots, v_N\}$, no matter how large N is. As we shall see in Chapter 3, this union property is one of the main properties distinguishing probability from more qualitative notions of likelihood. With probability, sufficiently many "small" probabilities eventually can dominate a "large" probability. This suggests that $u \succ v$ should perhaps be interpreted as "u is *much* more likely than v".

The key property that characterizes $\succ^s$ is somewhat more subtle. A relation $\rhd$ on 2^W is *qualitative* if $(V_1 \cup V_2) \rhd V_3$ and $(V_1 \cup V_3) \rhd V_2$ implies $V_1 \rhd (V_2 \cup V_3)$. If we think of $\rhd$ as meaning "much more likely", then this property says that if $V_1 \cup V_2$ is much more likely than V_3 and $V_1 \cup V_3$ is much more likely than V_2, then most of the likelihood has to be concentrated in V_1. Thus, V_1 must be much more likely than $V_2 \cup V_3$.

My goal now is to show how the properties we have defined help us characterize the properties of $\succ^s$.

The first result says that, in the presence of orderliness, the qualitative property already implies transitivity and the union property.

Lemma 2.3.3 *If $\rhd$ is an orderly qualitative relation on 2^W, then $\rhd$ is transitive and satisfies the union property.*

Proof See Exercise 2.14. ∎

The converse to Lemma 2.3.3 does not hold. Indeed, an orderly strict partial order on 2^W may satisfy the union property and still not be qualitative (see Exercise 2.15).

The following proposition summarizes the properties of $\succ^s$.

Proposition 2.3.4 *$\succ^s$ is an orderly qualitative strict partial order on 2^W.*

Proof See Exercise 2.16. ∎

Neither $\succ'$ nor $\succeq^s$ is qualitative in general. In Example 2.3.1, we have $\{w_1, w_2\} \succ' \{w_1\}$ and $\{w_1, w_2\} \succ' \{w_2\}$, but we do not have $\{w_1, w_2\} \succ' \{w_1, w_2\}$. This example also shows that $\succeq^s$ is not qualitative, since if it were, we could conclude from $\{w_1, w_2\} \succeq^s \{w_2\}$ (taking $V_1 = \{w_1\}$ and $V_2 = V_3 = \{w_2\}$ in the definition of qualitative) that $\{w_1\} \succeq^s \{w_2\}$, a contradiction. It turns out that the qualitative property allows us to make interesting connections between these qualitative likelihood orderings and some notions that we shall see in later chapters, like *plausibility measures* (see Chapter 3) and *nonmonotonic reasoning* (see Chapter 6). That is why I used $\succ^s$ as the basis for the definition of relative likelihood here, and not $\succ'$ or $\succeq^s$.

With these concepts in hand, I can describe a sound and complete axiomatization for this logic of relative likelihood. Let $\mathrm{AX}_{\gg}$ consist of the following axioms and inference rules.

Prop. All substitution instances of tautologies of propositional calculus.

RL1. $\neg(\varphi \gg_i \varphi)$.

RL2. $(((\varphi_1 \vee \varphi_2) \gg_i \varphi_3) \wedge ((\varphi_1 \vee \varphi_3) \gg_i \varphi_2)) \Rightarrow (\varphi_1 \gg_i (\varphi_2 \vee \varphi_3))$.

RL3. $(K_i(\varphi \Rightarrow \varphi') \wedge K_i(\psi' \Rightarrow \psi) \wedge (\varphi \gg_i \psi)) \Rightarrow \varphi' \gg_i \psi'$.

MP. From φ and $\varphi \Rightarrow \psi$ infer ψ (Modus ponens).

Gen. From φ infer $K_i\varphi$ (Generalization).

Note that RL1, RL2, and RL3 just express the fact that $\succ^s$ is irreflexive, qualitative, and orderly, respectively.

Theorem 2.3.5 *$AX_{\gg}$ is a sound and complete axiomatization of the language $\mathcal{L}_n^{\gg}$ with respect to preferential structures.*

Proof The soundness of Prop is immediate. It is clear that the fact that $\succ^s$ is irreflexive and qualitative, as shown in Proposition 2.3.4, implies that RL1 and RL2 are sound. To see that RL3 corresponds to orderliness, note that if $M \models K_i(\varphi \Rightarrow \varphi') \wedge K_i(\psi' \Rightarrow \psi)$ and $\varphi \gg_i \psi$, then $[\![\varphi]\!]_M \subseteq [\![\varphi']\!]_M$, $[\![\psi']\!]_M \subseteq [\![\psi]\!]_M$, and $[\![\varphi]\!]_M \succ^s [\![\psi]\!]_M$. Since $\succ^s$ is orderly, it follows that $[\![\varphi']\!]_M \succ^s [\![\psi']\!]_M$, so $M \models \varphi' \gg_i \psi'$. Thus, RL3 is sound. It is also clear that MP and Gen preserve validity.

The completeness proof is quite difficult; its proof is beyond the scope of this book. See Exercise 2.24 for a related proof. ∎

In simple preferential structures, we get an additional axiom that reflects the fact that the ordering is independent of the world. Consider the following axiom (where I use $\gg$ and K rather than $\gg_i$ and K_i, since in simple preferential structures there is only one agent).

RL4(a). $(\varphi \gg \psi) \Rightarrow K(\varphi \gg \psi)$.

RL4(b). $\neg(\varphi \gg \psi) \Rightarrow K\neg(\varphi \gg \psi)$.

Theorem 2.3.6 *$AX_{\gg} \cup \{RL4\}$ is a sound and complete axiomatization for the language $\mathcal{L}^{\gg}$ with respect to simple preferential structures.*

Proof The soundness of RL4 follows immediately from the observation that if M is a simple preferential structure and $(M, w) \models \varphi \gg \psi$, then $(M, w') \models \varphi \gg \psi$ for all worlds w' in M. Again, completeness is beyond the scope of this book. ∎

Finally, let's consider what happens if the $\succeq$ relation on possible worlds is total. What is the impact of this assumption on the properties of $\succ^s$? It turns out that it is characterized by a notion called *modularity.* A relation $\rhd$ on an arbitrary set W' (not necessarily of the form 2^W) is *modular* if $w_1 \rhd w_2$ implies that, for all w_3, either $w_3 \rhd w_2$ or $w_1 \rhd w_3$. Modularity is the "footprint" of a total preorder on the strict order derived from it. This is made precise in the following lemma.

Lemma 2.3.7 *If $\succeq$ is a total preorder, then the strict partial order $\succ$ determined by $\succeq$ is modular. Moreover, if $\rhd$ is a modular, strict partial order on W, then there is a total preorder $\succeq$ on W such that $\rhd$ is the strict partial order determined by $\succeq$.*

Proof See Exercise 2.17. ∎

Modularity is preserved when we lift the order from W to 2^W.

Lemma 2.3.8 *If $\succ$ is a modular relation on W, then $\succ^s$ is a modular relation on 2^W.*

Proof See Exercise 2.18. ∎

Modularity is easily axiomatized as follows:

RL5. $\varphi_1 \gg_i \varphi_2 \Rightarrow ((\varphi_1 \gg_i \varphi_3) \vee (\varphi_3 \gg_i \varphi_2))$.

A preferential structure is *total* if it has the form $(W, O_1, \ldots, O_n, \pi)$, where $O_i(w)$ is a total preorder on W for each $w \in W$ and each agent $i \in \{1, \ldots, n\}$. Let $\mathcal{M}_n^{tot}$ be the subset of $\mathcal{M}_n^{pref}$ consisting of all total structures and let $\mathrm{AX}_{\gg}^{M}$ consist of $\mathrm{AX}_{\gg}$ together with RL5.

Theorem 2.3.9 *$AX_{\gg}^{M}$ is a sound and complete axiomatization of the language $\mathcal{L}_n^{\gg}$ with respect to $\mathcal{M}_n^{tot}$.*

Proof Soundness is straightforward and left to the reader (Exercise 2.19). An outline of the proof of completeness can be found in Exercise 2.24. ∎

Recall that we showed earlier that the converse to Lemma 2.3.3 does not hold in general for strict partial orders. A strict partial order may satisfy the union property and still not be qualitative. However, in the presence of modularity, the union property does imply qualitativeness.

Lemma 2.3.10 *If $\rhd$ is a modular, transitive relation that satisfies the union property, then $\rhd$ is qualitative.*

Proof See Exercise 2.25. ∎

Not surprisingly, we can replace RL2 in $\mathrm{AX}_{\gg}^{M}$ by axioms saying that $\gg_i$ is transitive and satisfies the union property, namely:

RL6. $((\varphi_1 \gg_i \varphi_2) \wedge (\varphi_2 \gg_i \varphi_3)) \Rightarrow (\varphi_1 \gg_i \varphi_3)$.

RL7. $((\varphi_1 \gg_i \varphi_2) \vee (\varphi_1 \gg_i \varphi_3)) \Rightarrow (\varphi_1 \gg_i \varphi_2 \vee \varphi_3)$.

Let AX_n^{tot} consist of RL1, RL3, RL5, RL6, RL7, MP, and Gen; that is, we replace RL2 in $\mathrm{AX}_{\gg}^{M}$ by RL6 and RL7.

Theorem 2.3.11 *AX_n^{tot} and $AX_{\gg}^{M}$ are equivalent. That is, φ is provable in AX_n^{tot} iff it is provable in $AX_{\gg}^{M}$.*

Proof See Exercise 2.26. ∎

These results are intended to show that we can go a long way towards doing comparative reasoning about likelihood without invoking numbers. In the next chapter, we delve into more quantitative notions of likelihood.

Exercises

2.1 Prove the remaining part of Lemma 2.1.1.

2.2 Prove that a formula φ is valid if and only if $\neg\varphi$ is not satisfiable.

2.3 Show that $\{p \Rightarrow (r \Rightarrow q), p \wedge r\}$ entails q.

2.4 Show that if ψ is a Boolean combination of formulas of the form $K\varphi$ and $M = (W, w_0, \pi)$ is a simple structure, then $(M, w) \models \psi$ for some world $w \in W \cup \{w_0\}$, then $(M, w') \models K\varphi$ for all worlds $w' \in W \cup \{w_0\}$.

2.5 Show that a binary relation is reflexive, symmetric, and transitive if and only if it is reflexive, Euclidean, and transitive.

2.6 Prove Lemma 2.2.1.

2.7 Prove Theorem 2.2.3.

2.8 Prove Proposition 2.2.6.

2.9 Prove Proposition 2.2.7.

2.10 Show that $K_i(\varphi \wedge \psi) \Leftrightarrow (K_i\varphi \wedge K_i\psi)$ is valid in all Kripke structures.

2.11 Show that an irreflexive and transitive relation must be antisymmetric.

2.12 Show that if $U \succ^s V$ then $U \succ' V$. Note that Example 2.3.1 shows that the converse does not hold in general. However, show that the converse does hold if $\succeq$ is total. (Thus, for total preorders, $\succ^s$ and $\succ'$ agree.)

* **2.13** This exercise provides a semantics for $\succ^s$ that has appropriate properties even in infinite domains. As we have seen, to have $U \succ^s V$, it is not enough that for every element v in V there is some element u in U that is more likely than v. This is what allows $W_\infty \succ W_\infty$ in Example 2.3.2. Notice that in the finite case, it is easy to see that if $U \succ^s V$, then for every element $v \in V$, there must be some $u \in U$ such that, not only do we have $u \succ v$, but u *dominates* V in that for no $v' \in V$ do we have $v' \succ u$. (Notice that if $\succeq$ is a total preorder, this is equivalent to saying that $w \succeq v$ for all $v \in V$.) It is precisely this domination condition that does not hold in Example 2.3.2. Now write $U \succ^s V$ if U is nonempty and, for all $v \in V$, there exists $u \in U$ such that $u \succ v$ and u dominates V. Show that this definition agrees with the earlier definition if U and V are finite. Moreover, show that with this definition, $\succ^s$ is irreflexive and transitive, even in infinite domains.

2.14 Prove Lemma 2.3.3.

2.15 Construct a set W and an orderly strict partial order on 2^W that satisfies the union property but is not qualitative.

2.16 Prove Proposition 2.3.4.

2.17 Prove Lemma 2.3.7.

2.18 Prove Lemma 2.3.8.

2.19 Show that all the axioms in $\mathrm{AX}^M_{\gg}$ are valid in $\mathcal{M}^{tot}_n$.

2.20 Show that if W is finite and $\rhd$ is a modular order on 2^W, then there is a total order $\succeq$ on W such that $\rhd = \succ^s$. As Exercise 2.21 shows, the analogue of this result does not hold without the assumption that $\rhd$ is modular, even if it is an orderly, qualitative, strict partial order.

2.21 Suppose that $W = \{a, b, c\}$. Define a relation $\rhd$ on 2^W such that $\{b, c\} \rhd \{a\}$, $W \rhd \{a\}$, $X \rhd \emptyset$ for all nonempty $X \subseteq W$, and these are the only pairs of sets in the $\rhd$ relation. Show that $\rhd$ is an orderly, qualitative, strict partial order, but there is no total preorder $\succeq$ on W such that $\rhd = \succ^s$.

2.22 This exercise and the one following, like the previous pair, demonstrate an important difference between total and partial preorders. This exercise shows that in the case of total preorders, we can safely identify worlds with truth assignments; this is not the case with partial preorders.

Suppose that $M = (W, \succeq, \pi)$ is a preferential structure, where W is finite and $\succeq$ is a total preorder on W. Define a subset W' of W that consists of the maximal worlds (with respect to $\succeq$) satisfying each truth assignment. More precisely, for each truth assignment v, let $W_v = \{w \in W : \pi(w) = v\}$. If $W_v \neq \emptyset$, choose $w_v \in W_v$ that is maximal in that, for all $w' \in W_v$, it is not the case that $w' \succ w_v$. Let $W' = \{w_v : W_v \neq \emptyset\}$. Let $\succeq'$ and π' be the restrictions of $\succeq$ and π to W', and let $M' = (W', \succeq', \pi')$. Notice that in W', distinct worlds are associated with different truth assignments. Show that $M \models \varphi \gg \psi$ iff $M' \models \varphi \gg \psi$. In this sense, we can assume that if we restrict to total preorders, then there is at most one world per truth assignment in a structure, and thus we can identify worlds with truth assignments.

2.23 Suppose $\Phi = \{p, q\}$. Let φ be the formula $(p \gg \neg p \wedge q) \wedge \neg(p \wedge q \gg \neg p \wedge q) \wedge \neg(p \wedge \neg q \gg \neg p \wedge q)$. Let $M = (W, \succeq, \pi)$ be such that $W = \{w_1, w_2, w_3, w_4\}$, $w_1 \succ w_3$, $w_2 \succ w_4$, $p \wedge q$ is true at w_1, $p \wedge \neg q$ is true at w_2, and $\neg p \wedge q$ is true at both w_3 and w_4. Show that $M \models \varphi$. However, show that φ is not satisfiable in any structure where there is at most one world satisfying $\neg p \wedge q$.

* **2.24** This exercise sketches the completeness proof for Theorem 2.3.6.

(a) A formula is φ is *consistent with* $\mathrm{AX}^M_{\gg}$ if $\neg\varphi$ is not provable in $\mathrm{AX}^M_{\gg}$. Recall that we want to show that every formula valid in all $\mathcal{M}^{tot}_n$ is provable in $\mathrm{AX}^M_{\gg}$. Show that it suffices to prove that every consistent formula is satisfiable in some total structure.

(b) Suppose that φ is consistent. We want to show that φ is satisfiable. Let $\varphi_1, \ldots, \varphi_m$ be the subformulas of φ, with $\varphi = \varphi_1$. Define an *atom over* φ to be a conjunction of the form $\psi_1 \wedge \ldots \wedge \psi_m$, where ψ_i is either φ_i or $\neg\varphi_i$, and $\psi_1 = \varphi$. Using propositional reasoning (Prop and MP), show that φ is provably equivalent to the disjunction of the consistent atoms over φ.

(c) Show that it follows from part (b) that some atom over φ, say σ, must be consistent.

(d) The goal is now to construct a total structure satisfying the atom σ of part (c), since this structure must also satisfy φ. Let $p_1, \ldots, p_n$ be the primitive propositions that appear in φ. Let Σ consist of all the $N = 2^n$ truth assignments to these primitive propositions. Let $W = \Sigma$. Define a total preorder $\rhd$ on 2^W as follows. Given a truth assignment α, let φ_α consist of the conjunction $q_1 \wedge \ldots \wedge q_n$, where q_i is p_i if $\alpha(p_i) = \textbf{true}$ and $\neg p_i$ otherwise. If V is a set of truth assignments, let φ_V be the disjunction of the formulas φ_α for $\alpha \in V'$. Define a binary relation $\rhd$ on 2^W as follows: $V \rhd V'$ iff $AX \vdash \sigma \Rightarrow (\varphi_V \gg \varphi_{V'})$. Show that $\rhd$ is a modular strict partial order on $\mathcal{F}$. (Hint: the fact that $\rhd$ is irreflexive follows easily from RL1; the fact that it is orderly follows from RL3; the fact that it is qualitative follows from RL2; transitivity follows from the fact that $\rhd$ is qualitative and orderly, by Lemma 2.3.3; modularity follows from RL4.)

(e) By Exercise 2.20, there is a total preorder $\succeq$ on W such that $\rhd = \succ^s$. Let $M = (W, \succeq, \pi)$, where $\pi(\alpha) = \alpha$. Let $w \in W$ be such that have $\pi(w)(p_i) = \textbf{true}$ iff p_i is one of the conjuncts in σ. Show that for each subformula φ_j of φ, we have $(M, w) \models \varphi_j$ iff φ is a conjunct of σ. (Hint: proceed by induction on the structure of formulas.)

Thus, (M, w) satisfies σ, and hence φ.

2.25 Prove Lemma 2.3.10.

2.26 Prove Theorem 2.3.11.

Notes

An excellent introduction to propositional logic can be found in [Enderton 1972], which is the source of the example about borogroves. Numerous alternatives to classical logic have been proposed over the years. Perhaps the best known include *multi-valued logics* [Rescher 1969; Rine 1984], *intuitionistic logic* [Heyting 1956], and *relevance logics* [Anderson and Belnap 1975].

A simple complete axiomatization for propositional logic is given by the two axioms

$$\varphi \Rightarrow (\psi \Rightarrow \varphi)$$
$$(\varphi_1 \Rightarrow (\varphi_2 \Rightarrow \varphi_3)) \Rightarrow ((\varphi_1 \Rightarrow \varphi_2) \Rightarrow (\varphi_1 \Rightarrow \varphi_3)).$$

A completeness proof can be found in [Popkorn 1994], which is also a good introduction to recent work in modal logic. Other good introductions are provided by Chellas [1980] and Hughes and Cresswell [1968, 1984].

Modal logic was originally viewed as the logic of possibility and necessity. (That is, the only modal operators considered were operators for necessity and possibility.) More recently, it has become common to view knowledge, belief, time, and so on, as modalities, and the term "modal logic" has encompassed logics for reasoning about these notions as well.

The presentation of Section 2.2 is largely taken from [Fagin, Halpern, Moses, and Vardi 1995], which explores the topic of reasoning about knowledge, and its applications to artificial intelligence, distributed systems, and game theory, in much more detail. A proof of Theorem 2.2.3 can also be found there, as well as a discussion of approaches to giving semantics to knowledge that distinguish between logically equivalent formulas (in that one of two logically equivalent formulas may be known while the other is not). Finally, the book contains an extensive bibliography of the literature on the subject. One particularly useful reference is [Lenzen 1978], which discusses in detail the justifications of various axioms for knowledge.

What I have called here an Aumann structure is a special case of what is called a *frame* in the modal logic literature, that is, a Kripke structure without the interpretation π. Frames were introduced by Lemmon and Scott [Lemmon 1977], who called them "world systems"; the term "frame" is due to Segerberg [1968].

The material in Section 2.3 is taken from [Halpern 1997a]. All the results in Section 2.3 are proved in [Halpern 1997a], as are the results discussed in Exercise 2.13. Most of the ideas in this section go back to Lewis [1973], but he focused on the case of total preorders.

Bibliography

Adams, E. (1975). *The Logic of Conditionals*. Dordrecht, Netherlands: D. Reidel.

Anderson, A. and N. D. Belnap (1975). *Entailment: The Logic of Relevance and Necessity*. Princeton, N.J.: Princeton University Press.

Ash, R. B. (1970). *Basic Probability Theory*. New York: John Wiley & Sons.

Aumann, R. J. (1976). Agreeing to disagree. *Annals of Statistics 4*(6), 1236–1239.

Bacchus, F., A. J. Grove, J. Y. Halpern, and D. Koller (1996). From statistical knowledge bases to degrees of belief. *Artificial Intelligence 87*(1–2), 75–143.

Bacchus, F., H. E. Kyburg, and M. Thalos (1990). Against conditionalization. *Synthese 85*, 475–506.

Bar-Hillel, M. and R. Falk (1982). Some teasers concerning conditional probabilities. *Cognition 11*, 109–122.

Bonanno, G. and K. Nehring (1996). How to make sense of the common prior assumption under incomplete information. Unpublished manuscript.

Boole, G. (1854). *An Investigation Into the Laws of Thought on Which are Founded the Mathematical Theories of Logic and Probabilities.* London: Macmillan.

Borel, E. (1943). *Les Probabilités et la Vie*. Paris: Presses Universitaires de France. English translation *Probabilities and Life* (1962), New York: Dover.

Burgess, J. (1981). Quick completeness proofs for some logics of conditionals. *Notre Dame Journal of Formal Logic 22*, 76–84.

Cheeseman, P. (1985). In defense of probability. In *Proc. Ninth International Joint Conference on Artificial Intelligence (IJCAI '85)*, pp. 1002–1009.

Chellas, B. F. (1980). *Modal Logic.* Cambridge, U.K.: Cambridge University Press.

Choquet, G. (1953). Theory of capacities. *Annales de l'Institut Fourier (Grenoble) 5*, 131–295.

Cox, R. (1946). Probability, frequency, and reasonable expectation. *American Journal of Physics 14*(1), 1–13.

Davis, M. (1977). *Applied Nonstandard Analysis.* New York: Wiley.

de Campos, L. M., M. T. Lamata, and S. Moral (1990). The concept of conditional fuzzy measure. *International Journal of Intelligent Systems 5*, 237–246.

Dempster, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics 38*, 325–339.

Dempster, A. P. (1968). A generalization of Bayesian inference. *Journal of the Royal Statistical Society, Series B 30*, 205–247.

Diaconis, P. (1978). Review of "A Mathematical Theory of Evidence". *Journal of the American Statistical Society 73*(363), 677–678.

Diaconis, P. and S. L. Zabell (1982). Updating subjective probability. *Journal of the American Statistical Society 77*(380), 822–830.

Diaconis, P. and S. L. Zabell (1986). Some alternatives to Bayes's rule. In B. Grofman and G. Owen (Eds.), *Proc. Second University of California, Irvine, Conference on Political Economy*, pp. 25–38.

Dubois, D. and H. Prade (1982). On several representations of an uncertain body of evidence. In M. M. Gupta and E. Sanchez (Eds.), *Fuzzy Information and Decision Processes*, pp. 167–181.

Dubois, D. and H. Prade (1990). An introduction to possibilistic and fuzzy logics. See Shafer and Pearl [1990], pp. 742–761.

Dubois, D. and H. Prade (1991). Possibilistic logic, preferential models, non-monotonicity and related issues. In *Proc. Twelfth International Joint Conference on Artificial Intelligence (IJCAI '91)*, pp. 419–424.

Enderton, H. B. (1972). *A Mathematical Introduction to Logic.* New York: Academic Press.

Fagin, R. and J. Y. Halpern (1991a). A new approach to updating beliefs. In P. Bonissone, M. Henrion, L. Kanal, and J. Lemmer (Eds.), *Uncertainty in Artificial Intelligence: Volume VI*, pp. 347–374. Amsterdam: Elsevier Science Publishers.

Fagin, R. and J. Y. Halpern (1991b). Uncertainty, belief, and probability. *Computational Intelligence 7*(3), 160–173.

Fagin, R. and J. Y. Halpern (1994). Reasoning about knowledge and probability. *Journal of the ACM 41*(2), 340–367.

Fagin, R., J. Y. Halpern, and N. Megiddo (1990). A logic for reasoning about probabilities. *Information and Computation 87*(1/2), 78–128.

Fagin, R., J. Y. Halpern, Y. Moses, and M. Y. Vardi (1995). *Reasoning about Knowledge*. Cambridge, Mass.: MIT Press.

Fariñas del Cerro, L. and A. Herzig (1991). A modal analysis of possibilistic logic. In *Symbolic and Quantitative Approaches to Uncertainty*, Lecture Notes in Computer Science, Vol. 548, pp. 58–62. Berlin/New York: Springer-Verlag.

Feinberg, Y. (1995). A converse to the Agreement Theorem. Technical Report Discussion Paper #83, Center for Rationality and Interactive Decision Theory.

Feinberg, Y. (1996). Characterizing common priors in the form of posteriors. Unpublished manuscript.

Feldman, Y. (1984). A decidable propositional probabilistic dynamic logic with explicit probabilities. *Information and Control 63*, 11–38.

Feller, W. (1957). *An Introduction to Probability Theory and its Applications* (2nd ed.), Volume 1. New York: John Wiley & Sons.

Finetti, B. d. (1931). Sul significato soggestivo del probabilità. *Fundamenta Mathematicae 17*, 298–329.

Finetti, B. d. (1937). La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré 24*, 17–24. Translated as "Foresight: its logical laws, its subjective sources" in [Kyburg and Smokler 1964].

Finetti, B. d. (1972). *Probability, Induction and Statistics*. New York: John Wiley & Sons, Inc.

Fischer, M. J. and N. Immerman (1986). Foundations of knowledge for distributed systems. In J. Y. Halpern (Ed.), *Theoretical Aspects of Reasoning about Knowledge: Proc. 1986 Conference*, pp. 171–186. San Francisco, Calif.: Morgan Kaufmann.

Fischer, M. J. and L. D. Zuck (1988). Reasoning about uncertainty in fault-tolerant distributed systems. Technical Report YALEU/DCS/TR–643, Yale University.

Freund, J. E. (1965). Puzzle or paradox? *American Statistician 19*(4), 29–44.

Friedman, N. and J. Y. Halpern (1995). Plausibility measures: a user's manual. In *Proc. Eleventh Conference on Uncertainty in Artificial Intelligence (UAI '95)*, pp. 175–184.

Friedman, N. and J. Y. Halpern (1998). Plausibility measures and default reasoning. *Journal of the ACM ?* Accepted for publication. Also available at `http://www.huji.ac.il/~nir`. A preliminary version appeared in *Proc., 13'th National Conference on Artificial Intelligence*, pp. 1297–1304, 1996.

Fudenberg, D. and J. Tirole (1991). *Game Theory*. Cambridge, Mass.: MIT Press.

Gabbay, D. (1985). Theoretical foundations for nonmonotonic reasoning in expert systems. pp. 459–476. Berlin: Springer-Verlag.

Gaifman, H. (1986). A theory of higher order probabilities. In J. Y. Halpern (Ed.), *Theoretical Aspects of Reasoning about Knowledge: Proc. 1986 Conference*, pp. 275–292. San Francisco, Calif.: Morgan Kaufmann.

Gardner, M. (1961). *Second Scientific American Book of Mathematical Puzzles and Diversions*. Simon & Schuster.

Geffner, H. (1992a). *Default Reasoning*. Cambridge, Mass.: MIT Press.

Geffner, H. (1992b). High probabilities, model preference and default arguments. *Mind and Machines 2*, 51–70.

Gilboa, I. and D. Schmeidler (1993). Updating ambiguous beliefs. *Journal of Economic Theory 59*, 33–49.

Goldszmidt, M., P. Morris, and J. Pearl (1993). A maximum entropy approach to nonmonotonic reasoning. *IEEE Transactions of Pattern Analysis and Machine Intelligence 15*(3), 220–232.

Goldszmidt, M. and J. Pearl (1992). Rank-based systems: A simple approach to belief revision, belief update and reasoning about evidence and actions. In *Proc. Third International Conference on Principles of Knowledge Representation and Reasoning (KR '92)*, pp. 661–672. San Francisco: Morgan Kaufmann.

Gordon, J. and E. H. Shortliffe (1984). The Dempster-Shafer theory of evidence. In B. G. Buchanan and E. H. Shortliffe (Eds.), *Rule-based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, Chapter 13. New York: Addison-Wesley.

Grove, A. J. and J. Y. Halpern (1997). Probability update: conditioning vs. cross-entropy. In *Proc. Thirteenth Conference on Uncertainty in Artificial Intelligence (UAI '97)*, pp. 208–214.

Hacking, I. (1975). *The Emergence of Probability*. Cambridge, UK: Cambridge University Press.

Hagashi, M. and G. J. Klir (1983). Measures of uncertainty and information based on possibility distributions. *International Journal of General Systems 9*(2), 43–58.

Halmos, P. (1950). *Measure Theory*. Van Nostrand.

Halpern, J. Y. (1990). Let many flowers bloom: a response to "An inquiry into computer understanding". *Computational Intelligence 6*, 184–188.

Halpern, J. Y. (1997a). Defining relative likelihood in partially-ordered preferential structures. *Journal of A.I. Research 7*, 1–24.

Halpern, J. Y. (1997b). On ambiguities in the interpretation of game trees. *Games and Economic Behavior 20*, 66–96.

Halpern, J. Y. (1998a). Characterizing the common prior assumption. In *Theoretical Aspects of Rationality and Knowledge: Proc. Seventh Conference*, San Francisco, pp. 133–146. Morgan Kaufmann.

Halpern, J. Y. (1998b). A logical approach for reasoning about uncertainty: a tutorial. In *Discourse, Interaction, and Communication*, pp. 141–155. Kluwer.

Halpern, J. Y. (1999a). A counterexample to theorems of Cox and Fine. *Journal of A.I. Research 10*, 76–85.

Halpern, J. Y. (1999b). Cox's theorem revisited. Available at http://www.cs.cornell.edu/home/halpern.

Halpern, J. Y. and R. Fagin (1989). Modelling knowledge and action in distributed systems. *Distributed Computing 3*(4), 159–179. A preliminary version appeared in *Proc. 4th ACM Symposium on Principles of Distributed Computing*, 1985, with the title "A formal model of knowledge, action, and communication in distributed systems: preliminary report".

Halpern, J. Y. and Y. Moses (1990). Knowledge and common knowledge in a distributed environment. *Journal of the ACM 37*(3), 549–587. A preliminary version appeared in *Proc. 3rd ACM Symposium on Principles of Distributed Computing*, 1984.

Halpern, J. Y. and M. R. Tuttle (1993). Knowledge, probability, and adversaries. *Journal of the ACM 40*(4), 917–962.

Halpern, J. Y., R. van der Meyden, and M. Y. Vardi (1997). Complete axiomatizations for reasoning about knowledge and time. Submitted for publication.

Halpern, J. Y. and M. Y. Vardi (1989). The complexity of reasoning about knowledge and time, I: lower bounds. *Journal of Computer and System Sciences 38*(1), 195–237.

Harsanyi, J. (1968). Games with incomplete information played by 'Bayesian' players, parts I-III. *Management Science 14*, 159–182, 320–334, 486–502.

Hart, S. and M. Sharir (1984). Probabilistic temporal logics for finite and bounded models. In *Proc. 16th ACM Symp. on Theory of Computing*, pp. 1–13.

Heyting, A. (1956). *Intuitionism: An Introduction*. Amsterdam: North-Holland.

Hisdal, E. (1978). Conditional possibilities—independence and noninteractivity. *Fuzzy Sets and Systems 1*, 283–297.

Hughes, G. E. and M. J. Cresswell (1968). *An Introduction to Modal Logic*. London: Methuen.

Hughes, G. E. and M. J. Cresswell (1984). *A Companion to Modal Logic*. London: Methuen.

Jaffray, J.-Y. (1992). Bayesian updating and belief functions. *IEEE Transactions on Systems, Man, and Cybernetics 22*(5), 1144–1152.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review 106*(4), 620–630.

Jeffrey, R. C. (1968). Probable knowledge. In I. Lakatos (Ed.), *International Colloquium in the Philosophy of Science: The Problem of Inductive Logic*, pp. 157–185. North Holland Publishing Co.

Jelinek, F. (1997). *Statistical Methods for Speech Recognition*. Cambridge, Mass.: MIT Press.

Kemeny, J. G. (1955). Fair bets and inductive probabilities. *Journal of Symbolic Logic 20*(3), 263–273.

Keynes, J. M. (1921). *A Treatise on Probability*. London: Macmillan.

Klir, G. J. and T. A. Folger (1988). *Fuzzy Sets, Uncertainty, and Information*. Englewood Cliffs, New Jersey: Prentice-Hall.

Klir, G. J. and M. Mariano (1987). On the uniqueness of possibilistic measure of uncertainty and information. *Fuzzy Sets and Systems 24*, 197–219.

Kouvatsos, D. D. (1994). Entropy maximisation and queueing network models. *Annals of Operations Research 48*, 63–126.

Kozen, D. (1985). Probabilistic PDL. *Journal of Computer and System Sciences 30*, 162–178.

Kraus, S., D. Lehmann, and M. Magidor (1990). Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence 44*, 167–207.

Kreps, D. (1988). *Notes on the Theory of Choice*. Boulder, Colorado: Westview Press.

Kries, J. v. (1886). *Die Principien der Wahrscheinlichkeitsrechnung und Rational Expectation*. Freiburg.

Kullback, S. and R. A. Leibler (1951). On information and sufficiency. *Annals of Mathematical Statistics 22*, 76–86.

Kyburg, Jr., H. E. and H. Smokler (Eds.) (1964). *Studies in Subjective Probability*. New York: John Wiley & Sons.

Lambalgen, M. v. (1987). *Random Sequences*. Ph. D. thesis, University of Amsterdam.

Lehmann, D. and S. Shelah (1982). Reasoning about time and chance. *Information and Control 53*, 165–198.

Lemmon, E. J. (1977). *The "Lemmon Notes": An Introduction to Modal Logic*. Oxford, U.K.: Basil Blackwell. Written in collaboration with Dana Scott; edited by Krister Segerberg. American Philosophical Quarterly Monograph Series. Monograph No. 11.

Lenzen, W. (1978). Recent work in epistemic logic. *Acta Philosophica Fennica 30*, 1–219.

Lewis, D. K. (1973). *Counterfactuals*. Cambridge, Mass.: Harvard University Press.

Makinson, D. (1989). General theory of cumulative inference. Lecture Notes in Artificial Intelligence, Vol. 346, pp. 1–18. Berlin: Springer-Verlag.

Manna, Z. and A. Pnueli (1992). *The Temporal Logic of Reactive and Concurrent Systems*, Volume 1. Berlin/New York: Springer-Verlag.

Marek, W. and M. Truszczyński (1993). *Nonmonotonic Logic*. Berlin/New York: Springer-Verlag.

Maurer, S. B. and A. Ralston (1991). *Discrete Algorithmic Mathematics*. Reading, Mass: Addison-Wesley.

May, S. (1976). Probability kinematics: a constrained optimization problem. *Journal of Philosophical Logic 5*, 395–398.

McCarthy, J. (1980). Circumscription—a form of non-monotonic reasoning. *Artificial Intelligence 13*, 27–39.

McDermott, D. and J. Doyle (1980). Non-monotonic logic I. *Artificial Intelligence 13*(1,2), 41–72.

Monderer, D. and D. Samet (1989). Approximating common knowledge with common beliefs. *Games and Economic Behavior 1*, 170–190.

Moore, R. C. (1985). Semantical considerations on nonmonotonic logic. *Artificial Intelligence 25*, 75–94.

Morgan, J. P., N. R. Chaganty, R. C. Dahiya, and M. J. Doviak (1991). Let's make a deal: the player's dilemma (with commentary). *The American Statistician 45*(4), 284–289.

Morris, S. (1994). Trade with heterogeneous prior beliefs and asymmetric information. *Econometrica 62*, 1327–1348.

Morris, S. (1995). The common prior assumption in economic theory. *Economics and Philosophy 11*, 227–253.

Mosteller, F. (1965). *Fifty Challenging Problems in Probability with Solutions*. Reading, Mass.: Addison-Wesley.

Nilsson, N. (1986). Probabilistic logic. *Artificial Intelligence 28*, 71–87.

Parikh, R. and R. Ramanujam (1985). Distributed processing and the logic of knowledge. In R. Parikh (Ed.), *Proc. Workshop on Logics of Programs*, pp. 256–268.

Paris, J. B. (1994). *The Uncertain Reasoner's Companion*. Cambridge, U.K.: Cambridge University Press.

Pearl, J. (1989). Probabilistic semantics for nonmonotonic reasoning: a survey. In R. J. Brachman, H. J. Levesque, and R. Reiter (Eds.), *Proc. First International Conference on Principles of Knowledge Representation and Reasoning (KR '89)*, pp. 505–516. Reprinted in *Readings in Uncertain Reasoning*, G. Shafer and J. Pearl (eds.), Morgan Kaufmann, San Francisco, Calif., 1990, pp. 699–710.

Pearl, J. (1990). System Z: A natural ordering of defaults with tractable applications to nonmonotonic reasoning. In *know90*, pp. 121–135. San Francisco: Morgan Kaufmann.

Popkorn, S. (1994). *First Steps in Modal Logic*. Cambridge; New York: Cambridge University Press.

Popper, K. (1968). *The Logic of Scientific Discovery (revised edition)*. London: Hutchison.

Prior, A. N. (1957). *Time and Modality.* Oxford, U.K.: Oxford University Press.

Rabin, M. O. (1980). Probabilistic algorithm for testing primality. *Journal of Number Theory 12*, 128–138.

Rabin, M. O. (1982). N-process mutual exclusion with bounded waiting by $4 \cdot \log n$-valued shared variable. *Journal of Computer and System Sciences 25*(1), 66–75.

Ramsey, F. P. (1931). Truth and probability. In R. B. Braithwaite (Ed.), *The Foundations of Mathematics and other Logical Essays*, pp. 156–198. London: Routledge and Kegan Paul.

Reiter, R. (1980). A logic for default reasoning. *Artificial Intelligence 13*, 81–132.

Reiter, R. (1984). Towards a logical reconstruction of relational database theory. In M. L. Brodie, J. Mylopoulos, and J. W. Schmidt (Eds.), *On Conceptual Modelling*, pp. 191–233. Berlin/New York: Springer-Verlag.

Rescher, N. (1969). *Many-valued Logic.* New York: McGraw-Hill.

Rine, D. C. (Ed.) (1984). *Computer Science and Multiple-Valued Logics: Theory and Applications.* North-Holland.

Rivest, R. L., A. Shamir, and L. Adelman (1978). A method for obtaining digital signatures and public-key cryptosystems. *Communications of the ACM 21*(2), 120–126.

Rosenschein, S. J. (1985). Formal theories of AI in knowledge and robotics. *New Generation Computing 3*, 345–357.

Rosenschein, S. J. and L. P. Kaelbling (1986). The synthesis of digital machines with provable epistemic properties. In J. Y. Halpern (Ed.), *Theoretical Aspects of Reasoning about Knowledge: Proc. 1986 Conference*, pp. 83–97. San Francisco, Calif.: Morgan Kaufmann.

Ruspini, E. H. (1987). The logical foundations of evidential reasoning. Research Note 408, revised version, SRI International, Menlo Park, Calif.

Samet, D. (1997). Bayesianism without learning. unpublished manuscript.

Samet, D. (1998). Quantified beliefs and believed quantities. In I. Gilboa (Ed.), *Theoretical Aspects of Rationality and Knowledge: Proc. Seventh Conference*, pp. 263–272. Morgan Kaufmann.

Samet, D. (1998, to appear). Common priors as separation of convex sets. *Games and Economic Behavior*.

Savage, L. J. (1954). *Foundations of Statistics*. New York: John Wiley & Sons.

Savant, M. v. (1990a, Sept. 9,). Ask Marilyn. *Parade Magazine*, 15.

Savant, M. v. (1990b, Dec. 2,). Ask Marilyn. *Parade Magazine*, 25.

Savant, M. v. (1991, Feb. 17,). Ask Marilyn. *Parade Magazine*, 12.

Segerberg, K. (1968). *Results in Nonclassical Logic*. Lund, Sweden: Berlingska Boktryckeriet.

Shackle, G. L. S. (1969). *Decision, Order, and Time in Human Affairs* (2nd ed.). Cambridge, U.K.: Cambridge University Press.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton, N.J.: Princeton University Press.

Shafer, G. (1979). Allocations of probability. *Annals of Probability 7*(5), 827–839.

Shafer, G. (1985). Conditional probability. *International Statistical Review 53*(3), 261–277.

Shafer, G. (1990). Perspectives on the theory and practice of belief functions. *International Journal of Approximate Reasoning 4*, 323–362.

Shafer, G. and J. Pearl (Eds.) (1990). *Readings in Uncertain Reasoning*. San Francisco: Morgan Kaufmann.

Shannon, C. and W. Weaver (1949). *The Mathematical Theory of Communication*. University of Illinois Press.

Shimony, A. (1955). Coherence and the axioms of confirmation. *Journal of Symbolic Logic 20*(1), 1–26.

Shoham, Y. (1987). A semantical approach to nonmonotonic logics. In *Proc. 2nd IEEE Symp. on Logic in Computer Science*, pp. 275–279. Reprinted in M. L. Ginsberg (Ed.), *Readings in Nonmonotonic Reasoning*, Morgan Kaufman, San Francisco, Calif., 1987, pp. 227–250.

Shore, J. E. and R. W. Johnson (1980). Axiomatic derivation of the principle of maximum entropy and the principle of minimimum cross-entropy. *IEEE Transactions on Information Theory IT-26*(1), 26–37.

Smets, P. and R. Kennes (1989). The transferable belief model: comparison with Bayesian models. Technical Report 89-1, IRIDIA, Université Libre de Bruxelles.

Smith, C. A. B. (1961). Consistency in statistical inference and decision. *Journal of the Royal Statistical Society, Series B 23*, 1–25.

Solovay, R. and V. Strassen (1977). A fast Monte Carlo test for primality. *SIAM Journal on Computing 6*(1), 84–85.

Spohn, W. (1988). Ordinal conditional functions: a dynamic theory of epistemic states. In W. Harper and B. Skyrms (Eds.), *Causation in Decision, Belief Change, and Statistics*, Volume 2, pp. 105–134. Dordrecht, Netherlands: Reidel.

Stalnaker, R. C. (1968). A theory of conditionals. In N. Rescher (Ed.), *Studies in logical theory*, Number 2 in American Philosophical Quarterly monograph series. Blackwell, Oxford. Also appears in *Ifs* (Ed. W. L. Harper, R. C. Stalnaker and G. Pearce), Reidel, Dordrecht, Netherlands, 1981.

Stalnaker, R. C. (1992). Notes on conditional semantics. In Y. Moses (Ed.), *Theoretical Aspects of Reasoning about Knowledge: Proc. Fourth Conference*, pp. 316–328. San Francisco: Morgan Kaufmann.

Stalnaker, R. C. and R. Thomason (1970). A semantical analysis of conditional logic. *Theoria 36*, 246–281.

Teller, P. (1973). Conditionalisation and observation. *Synthese 26*, 218–258.

Uffink, J. (1995). Can the maximum entropy principle be explained as a consistency requirement? *Studies in the History and Philosophy of Modern Physics 26*(3), 223–261.

van Fraassen, B. C. (1981). A problem for relative information minimizers. *British Journal for the Philosophy of Science 32*, 375–379.

van Fraassen, B. C. (1984). Belief and the will. *Journal of Philosophy 81*, 235–245.

Vardi, M. Y. (1985). Automatic verification of probabilistic concurrent finite-state programs. In *Proc. 26th IEEE Symp. on Foundations of Computer Science*, pp. 327–338.

von Mises, R. (1957). *Probability, Statistics, and Truth*. London: George Allen and Unwin. English translation of third German edition, 1951.

Walley, P. (1981). Coherent lower (and upper) probabilities. Manuscript, Dept. of Statistics, University of Warwick.

Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*, Volume 12 of *Monographs on Statistics and Applied Probability*. London: Chapman and Hall.

Weber, S. (1991). Uncertainty measures, decomposability and admissibility. *Fuzzy Sets and Systems 40*, 395–405.

Yager, R. R. (1983). Entropy and specificity in a mathematical theory of evidence. *International Journal of General Systems 9*, 249–260.

Zadeh, L. A. (1975). Fuzzy logics and approximate reasoning. *Synthese 30*, 407–428.

Zadeh, L. A. (1978). Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems 1*, 3–28.