A LOGICAL APPROACH TO REASONING ABOUT UNCERTAINTY: A TUTORIAL*

Joseph Y. Halpern
Cornell University
Computer Science Department
4144 Upson Hall
Ithaca, NY 14853
halpern@cs.cornell.edu
http://www.cs.cornell.edu/home/halpern

July 14, 1995

Abstract

I consider a logical framework for modeling uncertainty, based on the use of possible worlds, that incorporates knowledge, probability, and time. This turns out to be a powerful approach for modeling many problems of interest. I show how it can be used to give insights into (among other things) several well-known puzzles.

*This paper will appear in *Discourse, Interaction, and Communication*, X. Arrazola, K. Korta, and F. J. Pelletier, eds., Kluwer, 1997. Much of this work was performed while the author was at IBM Almaden Research Center. IBM's support is gratefully acknowledged.

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. These notes describe a systematic approach for doing so. I have made no attempt to be comprehensive here; I have been guided by my biases and my own research.

Reasoning about uncertainty can be subtle. Perhaps the best way to see this is to consider a number of puzzles, some of them very well-known. These puzzles are presented under the assumption that the uncertainty is quantified in terms of probability, but the issues that they bring out arise whatever method we use to represent uncertainty.

The second-ace puzzle [BF82, Fre65, Sha85]: Suppose 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.

Alice then says "I have an ace". Conditioning on this information, 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 Monty Hall Puzzle [Sav91, MCDD91]: 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 1. Before opening door 1, Monty Hall, the host (who knows what is behind each door), opens door 2, which has a goat. He then asks you if you still want to take what's behind door 1, or to take what's behind door 3 instead. Should you switch?

The Two-Coin Problem [FH94a]: 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.)

The Single-Coin Problem [FH94a]: This time both Bob and Alice know that Alice is using the fair coin. Alice tosses the coin and looks at the outcome. What is the probability of heads according to Bob? (Note I now want the probability *after* the coin toss.) 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.

There is certainly far more to representing uncertainty than dealing with puzzles such as these. Nevertheless, the analysis of these puzzles and paradoxes will give us deeper insight into the process of reasoning under uncertainty and the problems involved with getting a good representation.

So how do we represent and reason about uncertainty? I shall use the *possible-worlds* framework. This is the standard approach for giving semantics to modal logic. The intuition is that besides the true state of affairs, there are a number of other possible states of affairs or "worlds", that an agent considers possible. We can view the set of worlds that an agent considers possible as a qualitative way to measure her uncertainty. The more worlds she considers possible, the more uncertain she has as to the true state of affairs, and the less she knows. We can then quantify this uncertainy by adding a probability distribution to the possible worlds (or using some other means of "grading" the uncertainty). This is not quite enough for dealing with the puzzles above. We need to add time to the picture. That means we need to have possible worlds describing not only the current state of affairs, but the state of affairs at each time point of interest. As we shall see, it is also useful to assume that these states have some internal structure. The resulting framework, incorporating knowledge, time and probability in a concrete setting, is a powerful modeling tool.

The rest of this paper is organized as follows. Section 2 reviews modal logic and the possible-worlds framework. Section 3 discusses a concrete framework for modeling knowledge and time in multi-agent systems. Probability is added in Section 4. The various puzzles are analyzed in the resulting framework in Section 5. Other topics are briefly discussed in Section 6.

2 Basic modal logic: knowledge, belief, and time

Let us start by considering a simple model to capture an agent's knowledge, using ideas that go back to Hintikka [Hin62].

Suppose we have an agent with some information, and we want to reason about her beliefs. Given her current information, the agent may not be able to tell which of a number of possible worlds describes the actual state of affairs. We say that the agent *believes* or *knows*[1] a fact φ if φ is true in all the worlds the agent considers possible.

To capture this intuition, we use the language of modal logic. We start with a set Φ of *primitive propositions*, where a primitive proposition $p \in \Phi$ represents a basic fact of interest like "it is raining in Spain". We then close off under conjunction and negation, as in propositional logic, and modal operators $K_1, \ldots, K_n$, where $K_i\varphi$ is read "agent i knows φ". Thus, a statement such as $K_1K_2p \wedge \neg K_2K_1K_2p$ says "agent 1 knows agent 2 knows p, but agent 2 does not know that 1 knows that 2 knows p". More colloquially: "I know that you know it, but you don't know that I know that you know it."

Next we need a semantics. Suppose for simplicity we start with just one agent (and write $K\varphi$ rather than $K_1\varphi$). We define a *simple structure* M to be a triple (W, w_0, π), where W can be thought of as the set of worlds the agent considers possible, w_0 is the actual world, and π associates with each world 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 in our language. 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 once we move to multiple agents, or even to somewhat more sophisticated logics involving just one agent, this cannot be done.

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 $(M, w) \models \psi'$

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

[1] I will use "belief" and "knowledge" interchangeably here, and ignore the differences between them that the philosophical literature has focused on.

$(M, w) \models K\varphi$ 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. We remark that occasionally when M is clear from context, we write $w \models \varphi$ instead of $(M, w) \models \varphi$.

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. This 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. We return to this viewpoint in the next section.

There are times when we want the set of worlds the agent considers possible to depend on the actual world. This can be done in a straightforward way. We define a *Kripke structure* M to be a tuple $(W, \mathcal{K}, \pi)$, where $\mathcal{K}$ is a binary relation on W—that is, a set of pairs $(w, w') \in W \times W$. 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). If we define $\mathcal{K}(w) = \{w' : (w, w') \in \mathcal{K}\}$, then we can think of $\mathcal{K}(w)$ as describing the set of worlds the agent considers possible in world w. We thus define

- $(M, w) \models K\varphi$ if $(M, w') \models \varphi$ for all $w' \in \mathcal{K}(w)$.

Of course, this framework allows us to generalize easily to multiple agents. We simply have one possibility relation for each agent. In particular, if we have n agents, we take a Kripke structure for n agents to be a tuple $(W, \mathcal{K}_1, \ldots, \mathcal{K}_n, \pi)$, where each $\mathcal{K}_i$ is a binary relation on W. Of course, we now define:

- $(M, w) \models K_i\varphi$ if $(M, w') \models \varphi$ for all $w' \in \mathcal{K}_i(w)$.

The simple structures above are essentially Kripke structures for one agent where $\mathcal{K}(w) = \mathcal{K}(w')$ for $w, w' \in W$.

We can maintain the intuition that the worlds that the agent considers possible depend only on the agent's internal state without resorting to simple structures, but by taking the $\mathcal{K}_i$ relation to be transitive and *Euclidean*.[2] This guarantees that the set of worlds the agent considers possible is the same in all worlds that he considers possible. An even

[2]A binary relation $\mathcal{K}$ on W is Euclidean if $(u, v), (u, w) \in \mathcal{K}$ implies $(v, w) \in \mathcal{K}$, for all $u, v, w \in W$.

stronger assumption, that further ensures that what the agent knows is true, is to take the $\mathcal{K}_i$ relation to be an equivalence relation.[3]

Up to now, we have considered knowledge (and belief). In our examples, time also formed a crucial element. Clearly an agent's beliefs change over time. Thus, to capture this, we need to have time in our model. It is easy to construct a simple model of time, along the lines above. We take a temporal structure T to consist of a pair $((w_0, w_1, w_2, \ldots), \pi)$, where, intuitively, w_k is the world at "time" k, and π associates with each world a truth assignment, as before. Thus, we are implicitly assuming that time is discrete, linear (for each time, there is a unique next time), and infinite. While each of these assumptions can be (and have been!) challenged, let us make them for now. The assumption of discreteness certainly seems reasonable in human affairs (we can allow an arbitrarily fine level of granularity), linearity seems reasonable in our applications (as we shall see, if we want to consider different possible futures, we can combine knowledge with time), and if we want to model a situation where there are only finitely many steps, all we have to do is to repeat the last step infinitely often to get an infinite sequence.

The type of modal operators that are typically considered when dealing with time are $\bigcirc$ and $\Box$, where $\bigcirc$ means "at the next time interval" while $\Box$ means "now and at all times in the future". Thus, for example,

$$(T, w_k) \models \bigcirc\varphi \text{ if } (T, w_{k+1}) \models \varphi, \text{ and}$$

$$(T, w_k) \models \bigcirc\varphi \text{ if } (T, w_m) \models \varphi, \text{ for all } m \geq k.$$

Of course, we can also have structures that combine knowledge and time. A concrete framework for doing so is presented in the next section.

3 A concrete framework for multi-agent systems

The possible-worlds framework of the previous section is somewhat abstract. Possible worlds are represented as elements of a set. But where are these possible worlds coming from? I shall now consider a more concrete framework, that incorporates both knowledge and time, and takes an agent's knowledge to be determined by the agent's internal state. The ideas presented here are mainly taken from [HF89, FHMV95].

Suppose we want to analyze a multi-agent system. The phrase "system" is intended to be interpreted rather loosely here. Players in a poker game, agents conducting a bargaining session, robots interacting to clean a house, and processes in a computing system can all be viewed as multi-agent systems. The only assumption I shall make here about a system is that, at all times, each of the agents in the can be viewed as being in some *local* or *internal* state. Intuitively, the local state encapsulates all the relevant

[3]Readers familiar with modal logic will recognize that if we assume that the $\mathcal{K}_i$'s are Euclidean and transitive, we get the modal logic K45; if we further assume that they are equivalence relations, we get S5 [HM92, HC68].

information to which the agent has access. For example, if we are modeling a poker game, a player's state might consist of the cards he currently holds, the bets made by the other players, any other cards he has seen, and any information he may have about the strategies of the other players (for example, Bob may know that Alice likes to bluff, while Charlie tends to bet conservatively).

It is also useful to view the system as a whole as being in a state. The first thought might be to make the system's state be a tuple of the form $(s_1, \ldots, s_n)$, where s_i is agent i's state. But, in general, more than just the local states of the agents may be relevant when doing an analysis of the system. If we are analyzing a message-passing system where agents send messages back and forth along communication lines, we might want to know about messages that are in transit or about the status of each communication line (whether it is up or down). If we are considering a system of sensors observing some terrain, we might need to include features of the terrain in a description of the state of the system. Thus, we conceptually divide a system into two components: the agents and the *environment*, where we view the environment as "everything else that is relevant". In many ways the environment can be viewed as just another agent, but one that we often ignore, since we are not usually interested in what the environment knows. We define a *global state* of a system with n agents or agents to be an $(n+1)$-tuple of the form $(s_e, s_1, \ldots, s_n)$, where s_e is the state of the environment and s_i is the local state of agent i.

A system is not a static entity. It is constantly changing over time. A *run* is a complete description of what happens over time in one possible execution of the system. For definiteness, we take time to range over the natural numbers. Thus, formally, a run is a function from the natural numbers to global states. Given a run r, $r(0)$ describes the initial global state of the system in r, $r(1)$ describes the next global state, and so on. We refer to a pair (r, m) consisting of a run r and time m as a *point*. If $r(m) = (s_e, s_1, \ldots, s_n)$, we define $r_e(m) = s_e$ and $r_i(m) = s_i$, $i = 1, \ldots, n$; thus, $r_i(m)$ is agent i's local state at the point (r, m).

How do we incorporate knowledge into this framework? The basic idea is that a statement such as "agent i does not know φ" means that, as far as i is concerned, the system could be at a point where φ does not hold. The way we capture that "as far as i is concerned, the system could be at a point where φ does not hold" is closely related to the notion of possible worlds in Kripke structures. We think of i's knowledge as being determined by its local state, so that i cannot distinguish between two points in the system in which it has the same local state, and it can distinguish points in which its local state differs. In fact, all that prevents us from viewing a distributed system as a Kripke structure is that we have no primitive propositions, and no function π telling us how to assign truth values to the primitive propositions. We now rectify this problem.

Assume that we have a set Φ of primitive propositions, which we can think of as describing basic facts about the system. These might be such facts as "Alice holds the ace of spades", "there is a goat behind door 3", or "process 1's initial input was 17". An *interpreted system* $\mathcal{I}$ consists of a pair $(\mathcal{R}, \pi)$, where $\mathcal{R}$ is a system and π associates with each point in $\mathcal{R}$ a truth assignment to the primitive propositions in Φ. We say that the point (r, m) is in the interpreted system $\mathcal{I} = (\mathcal{R}, \pi)$ if $r \in \mathcal{R}$.

We can associate with an interpreted system $\mathcal{I} = (\mathcal{R}, \pi)$ a Kripke structure $M_{\mathcal{I}} = (S, \mathcal{K}_1, \ldots, \mathcal{K}_n, \pi)$ in a straightforward way. The set S of states in $M_{\mathcal{I}}$ consists of the points in $\mathcal{I}$. We can define $\mathcal{K}_i$ so that $\mathcal{K}_i((r, m)) = \{(r', m') : r_i(m) = r'_i(m')\}$. Thus, $\mathcal{K}_i((r, m))$ consists of all points indistinguishable from (r, m) by agent i, where we think of two points as being by agent i if agent i has the same local state in both.

We can now define what it means for a formula φ in our language of knowledge to be true at a point (r, m) in an interpreted system $\mathcal{I}$ by applying the definitions of the previous section to the related Kripke structure $M_{\mathcal{I}}$. Thus we say that $(\mathcal{I}, r, m) \models \varphi$ exactly if $(M_{\mathcal{I}}, s) \models \varphi$, where $s = (r, m)$. For example, we have

- $(\mathcal{I}, r, m) \models p$ (for $p \in \Phi$) if $\pi(r, m)(p) = \mathbf{true}$
- $(\mathcal{I}, r, m) \models K_i\varphi$ iff $(\mathcal{I}, r', m') \models \varphi$ for all (r', m') such that $r_i(m) = r'_i(m')$.

Interpreted systems also allow us to reason about time, since each run provides a temporal model. Thus, for example, we have

- $(\mathcal{I}, r, m) \models \bigcirc\varphi$ if $(\mathcal{I}, r, m + 1) \models \varphi$.

We can also combine temporal and epistemic statements; a formula such as $K_1\bigcirc\neg K_2\Box p$ makes perfect sense.

Where do the runs in the system come from? Typically, they are generated by means of a *protocol*. I do not want to go into the formal definitions here (these can be found in [HF89, FHMV95]), but intuitively, a protocol is a description of the actions that an agent takes as a function of her local state. We shall see some examples of protocols later in the paper.

4 Adding probability

Up to now we have considered time and knowledge. The next step is to add probability. The idea now is to assume that besides having a set of worlds that she considers possible, an agent puts probability on these worlds. To reason about the probability, we need to enrich the language. I shall do so by allowing formulas of the form $\mathrm{pr}_i(\varphi) \geq \alpha$, $\mathrm{pr}_i(\varphi) \leq \alpha$, and $\mathrm{pr}_i(\varphi) = \alpha$, where φ is a formula in the language and α is a real number in the interval $[0, 1]$. A formula such as $\mathrm{pr}_i(\varphi) \geq \alpha$ can be read as "the probability of φ, according to agent i, is at least α". It is easy to allow even more complicated formulas such as $\mathrm{pr}_i(\varphi) + 2\mathrm{pr}_i(\psi) \leq \alpha$ (and this, in fact, is done in [FHM90]), but all the basic ideas should already be clear with the simple language we are considering.

To give semantics to such formulas, we need to augment the Kripke structures used in Section 2 with a probability distribution. To bring out the main ideas, let's start by assuming there is one agent (so that we can temporarily drop the subscript on pr). We take a *simple probability structure* M to be a tuple of the form (W, Pr, π), where Pr is a *discrete probability distribution* on W. A discrete probability distribution Pr just maps worlds in W to numbers in the interval $[0, 1]$, with the constraint that $\sum_{w \in W} \mathrm{Pr}(w) = 1$. We extend Pr to subsets of W by taking $\mathrm{Pr}(A) = \sum_{w \in A} \mathrm{Pr}(w)$.[4] We can now define satisfaction ($\models$) in simple probability structures; the only interesting case comes in dealing with formulas such as $\mathrm{pr}(\varphi) \geq \alpha$. Such a formula is true if the set of worlds where φ is true has probability at least α:

$$(M, w) \models \mathrm{pr}(\varphi) \geq \alpha \text{ if } \mathrm{Pr}(\{w : (M, w) \models \varphi\}) \geq \alpha.$$

Of course, the treatment of $\mathrm{pr}(\varphi) \leq \alpha$ and $\mathrm{pr}(\varphi) = \alpha$ is analogous.

Like the simple structures of Section 2, simple probability structures implicitly assume that the agent's probability distribution is independent of the state. We can generalize simple probability structures analogously to the way we extended simple structures, by having the probability distribution Pr depend on the world, and by allowing different agents to have different probability distributions. We thus define a probabilistic Kripke structure M to be a tuple of the form $(W, \mathcal{PR}_1, \ldots, \mathcal{PR}_n, \pi)$, where for each agent i and world w, we take $\mathcal{PR}_i(w)$ to be a discrete probability distribution, denoted $\mathrm{Pr}_{i,w}$, over W. (Actually, as we shall see, it is sometimes useful to view $\mathrm{Pr}_{i,w}$ as a distribution, not on W, but on some subset of W; that is, its domain is some $W' \subseteq W$ rather than all of W. Of course, we can identify a distribution over $W' \subseteq W$ with a distribution over W by just making all the worlds in $W - W'$ have probability 0.) To evaluate the truth of a formula such as $\mathrm{pr}_i(\varphi) \geq \alpha$ at the world w we use the distribution $\mathrm{Pr}_{i,w}$:

$$(M, w) \models \mathrm{pr}_i(\varphi) \geq \alpha \text{ if } \mathrm{Pr}_{i,w}(\{w : (M, w) \models \varphi\}) \geq \alpha.$$

We can easily combine reasoning about knowledge with reasoning about probability. A Kripke structure for knowledge and probability is a combination of a Kripke structure (for knowledge) and a Kripke structure for probability; thus, it is a tuple of the form $(W, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathcal{PR}_1, \ldots, \mathcal{PR}_n, \pi)$. Such a structure can be used to give semantics to a language that has both knowledge operators and probability operators. A natural assumption in that case is that, in world w, agent i assigns probability only to the worlds that he considers possible, namely $\mathcal{K}_i(w)$. We can model this either by assuming that $\mathrm{Pr}_{i,w}$ is defined only on $\mathcal{K}_i(w)$ or by taking $\mathrm{Pr}_{i,w}(w')$ to be 0 if $w' \notin \mathcal{K}_i(w)$. There may be times, however, when this is not appropriate.

Example 4.1: Let us reconsider the two-coin problem from the introduction. Recall that Alice has two coins, one of which is fair, the other biased. She chooses one of them.

[4] An arbitary—not necessarily discrete—probability distribution is not necessarily defined on all subsets of W, but only on a σ-algebra of subsets of W; that is, a set of subsets that is closed under countable union and complementation. There are added complexities in dealing with arbitrary probability distributions, which I would rather avoid here; see [FHM90] for details.

There are four possible worlds, which we can denote (F, H), (F, T), (B, H), (B, T): the fair coin is chosen and will land heads, the fair coin is chosen and will land tails, and so on.[5] Bob cannot distinguish any of these worlds; in any one of them, he considers all four possible. (His internal state is the same in all four, since he does not know which coin is chosen.) On the other hand, if a world of the form (F, x), Alice considers only worlds of the form (F, y) possible, while in a world of the form (B, x), Alice considers only worlds of the form (B, y) possible. Describing Alice's probability distributions $\Pr_{A,w}$ is straightforward: In a world of the form (F, x), she knows the fair coin is being used, so her probability space consists of the two worlds (F, H) and (F, T), each of which gets probability $1/2$. Similarly, in a world of the form (B, x), she knows the biased coin is being used, so her probability space consists of the two worlds (B, H) and (B, T); the former gets probability $2/3$, while the latter gets probability $1/3$. Suppose we take as primitive propositions f, b, h, and t, representing that Alice chooses the fair coin, Alice chooses the biased coin, the coin will land heads, and the coin will land tails, respectively. Then, for example, $(F, T) \models K_A f \wedge K_A(\mathrm{pr}_A(h) = 1/2)$: in the world where the fair coin is chosen and will in fact land tails, Alice knows that the fair coin is chosen and knows that the probability that it will land heads is $1/2$.

What about Bob? Clearly we have $(F, T) \models \neg K_B f$: Bob does not know which coin has been chosen. But what probability distribution should Bob use? As I said above, if Bob knew the probability of Alice's picking the coin was α (or, for that matter, if Bob's subjective probability that Alice would pick the fair coin were α), then there would be no problem: (F, H) and (F, T) would both get probability $\frac{1}{2}\alpha$, (B, H) would get probability $\frac{2}{3}(1-\alpha)$, and (B, T) would get probability $\frac{1}{3}(1-\alpha)$. Unfortunately, the problem statement does not give us α. A Bayesian would say that Bob should just go ahead and choose some α (and perhaps go on to say that in the absence of any information, $\frac{1}{2}$ would be the best choice for α).

This is not the place to get into a discussion about the merits of the Bayesian approach. Whatever its merits, we certainly want the model to be rich enough to allow us to model the fact that Bob does not know the probability that Alice chooses the fair coin. This is, after all, supposed to be an approach with which we can represent Bob's knowledge (or lack of it). One way to do this is to consider a family of models, one for each choice of α. But doing this negates the point of using the possible-worlds framework: We want to be able to represent Bob's uncertainty within one model, not by considering a family of models. Perhaps the simplest approach to doing this is not to take Bob's probability space to consist of all the four worlds he considers possible, but to split up these four worlds into two separate probability spaces: W_F consisting of the two worlds (F, H) and (F, T) and W_B consisting of the two worlds (B, H) and (B, T). W_F consists of the worlds where Alice chose the fair coin and W_B consists of the worlds where Alice chose the biased coin. In each of these worlds, Bob's probability distribution is the obvious one: For example, $\Pr_{B,(B,H)}$ assigns the world (B, H) probability $2/3$ and the world (B, T)

[5] Actually, it would be cleaner to represent this as a two-step process: first the coin is chosen, and then it is tossed. This is done in the next section, using runs and systems.

probability 1/3.

It is easy to see that, at each world, Bob's probability distribution is the same as Alice's: $\Pr_{A,w} = \Pr_{B,w}$ for each world $w \in W$. The difference between Bob and Alice is not captured in the probability distributions they use, but in the set of worlds they consider possible. We already observed that in the world (F, H), the formula $K_A(\mathrm{pr}_A(h) = 1/2)$ is true. This is because in both of the worlds that Alice considers possible—namely, (F, H) and (F, T)—the formula $\mathrm{pr}_A(h) = 1/2$ holds. Now since Bob is using the same probability distribution as Alice, it is also the case that at both (F, H) and (F, T), the formula $\mathrm{pr}_B(h) = 1/2$ holds. However, the formula $K_B(\mathrm{pr}_B(h) = 1/2)$ does *not* hold at the world (F, H) since Bob, unlike Alice, also considers the worlds (B, H) and (B, T) possible: Bob does not know whether Alice chose the fair coin or the biased coin. In the latter two worlds, the formula $\mathrm{pr}_B(h) = 2/3$ is true. Thus, we have $(F, H) \models K_B(\mathrm{pr}_B(h) = 1/2 \vee \mathrm{pr}_B(h) = 2/3)$. All Bob can say is that the probability of heads is either 1/2 or 2/3, but he does not know which. I would argue that this indeed is a reasonable representation of Bob's ignorance regarding which coin was chosen. ∎

Example 4.2: Armed with the insights from the previous example, we can now also consider the single-coin problem from the introduction. Despite its apparent simplicity, it involves a number of subtleties. Using the notation from the previous example, since Bob now knows that Alice used the fair coin, the only worlds that he considers possible are (F, H) and (F, T). Alice also knows the outcome of the coin toss, so in world (F, H), the only world she considers possible is (F, H), while in (F, T), the only world she considers possible is (F, T). Obviously she assigns probability 1 to the only world that she considers possible, so, for example, we now have $(F, H) \models K_A(h) \wedge K_A(\mathrm{pr}_A(h) = 1)$. How about Bob? What probability distribution should he place on the two worlds he considers possible. One obvious choice would be to make them equally likely, as in the previous example. After all, he does not know the outcome of the coin toss. Thus, if we take M_1 to be the structure where Bob's probability functions $\Pr_{B,(F,H)}$ and $\Pr_{B,(F,T)}$ give each of the two worlds (F, H) and (F, T) probability 1/2, then we have $(M_1, (F, H)) \models K_B(\mathrm{pr}_B(h) = 1/2)$. This seems like a reasonable answer: Bob still believes that heads has probability 1/2.

On the other hand, as we saw in the previous example, the framework does not force us to assign probability to all of Bob's possible worlds. We can divide them up. In particular, just as in the previous example, we can take Bob's probability distribution to be identical to Alice's: let $\Pr_{B,(F,H)}$ assign probability 1 to the world (F, H) (and hence probability 0 to (F, T)) and let $\Pr_{B,(F,T)}$ assign probability 1 to the world (F, T) (and hence probability 0 to (F, H)). If M_2 is the structure where $\mathcal{PR}_B$ is defined in this way, then $(M_2, (F, H)) \models K_B(\mathrm{pr}_B(h) = 1 \vee \mathrm{pr}_B(h) = 0)$, since at the world (F, H), $\mathrm{pr}_B(h) = 1$ holds, while at the world (F, T), $\mathrm{pr}_B(h) = 0$ holds.

I was careful in this example to include the structure. M_1 and M_2 are identical in all respects except for $\mathcal{PR}_B$, which determines Bob's probability distribution at each of the worlds. This is precisely the crucial distinction though. M_1 and M_2 capture formally the two competing intutions we discussed in the introduction: in M_1, Bob assigns probability 1/2 to heads; in M_2, the probability of heads is either 0 or 1, but Bob does not know which. It is beyond the scope of these notes to delve into which of the two approaches is "right". (This issue is discussed in some detail in [HT93], where the point is made that a useful way of approaching this issue is to think in terms of the nature of the adversary against which Bob is playing.) ∎

5 Combining knowledge, probability, and time

We can now put all the pieces together, and combine knowledge, probability, and time, using the framework of runs and systems. The basic idea is to put a probability on runs, and then condition. But the subtleties that we observed before we considered time carry over here.

Where does the probability come from? Typically, if we are given a protocol that generates the set of runs, then some of the transitions are probabilistic. The probability of the transitions then determines the probability of the runs. For example, suppose we first toss a biased coin (with heads having probability 2/3) and then toss a fair coin. There are four runs, depending on the outcome of the coin tosses. These runs are best thought of as being the paths in the tree in Figure 1, where we go left in the tree if the outcome of coin toss is heads, and go right if the outcome is tails. The branches of the tree are labeled by the probability of the outcome, so, for example, the top left branch is labeled by 2/3, since the probability of getting heads on the first coin toss is 2/3. The probability of a run is then just the product of the probabilities of the path in the tree that determines it, so, for example, the probability of getting two heads is 1/3, just as we would expect.

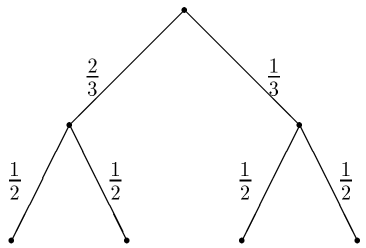

Figure 1: Tossing two coins

The first problem in this approach comes if not all transitions are probabilistic. Some may be nondeterministic, with no probabilities attached. This is precisely what happens if we model the two-coin example using runs. At the first step Alice chooses a coin. She can choose either a fair coin (go left) or a biased coin (go right). This step is nonprobabilistic; the problem does not give us probabilities with which to label this transition. The second step—tossing the coin—is probabilistic, as shown in Figure 2. The solution we take is just the same as that in Example 4.1: we separate the four runs into two probability spaces, one consisting of the runs corresponding to the fair coin (the paths labeled r_1 and r_2 in Figure 2) and one consisting of the runs corresponding to the biased coin (r_3 and r_4 in Figure 2). In the space consisting of $\{r_1, r_2\}$, each run gets probability 1/2. In the space consisting of $\{r_3, r_4\}$, the run r_3 gets probaility 2/3 and r_4 gets probability 1/3.

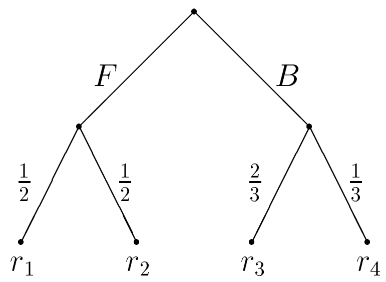

Figure 2: The two-coin example

Putting a probability on the runs (or on various subsets of runs) is not enough. We want to know what probability each agent assigns to various events at each point. In general, the probability that an agent assigns to an event like "the coin landed (or will land) heads" changes over time, depending on what information the agent obtains. To capture this, we need to have, for each agent at each point in the system, a probability. Clearly, we would like that probability to be related to the probability on runs. The obvious way to do this is to condition, but what do we condition on? There is not enough space here to go into all the details (see [HT93] for those) but the general answer is to condition on the agent's knowledge.

To make this clearer, let us consider the two-coin example in a little more detail, by putting in the local states for Alice and Bob. At time 0, we assume that Alice's local state is 0; this essentially captures the fact that she knows the time. At time 1, she is either in state F or B, depending on which coin she has chosen. At time 2, she is in one of the states (F, H), (F, T), (B, H), or (B, T), depending on the outcome of the coin toss. (The fact that Alice's state also contains either B or F encodes the fact that Alice does not forget which coin she has tossed.) Call the four runs where Alice ends up in these states $r_1, \ldots, r_4$, respectively. Bob never finds out which coin was tossed nor how it landed, but we assume that he does know which step of the protocol is currently being followed. Thus, we take his local state at time m, $m \leq 2$, to be just m.

What are Alice's probability spaces? At time 0, she considers all the points $(r_1, 0), \ldots, (r_4, 0)$ possible. (She knows the time is 0.) As in Example 4.1, we partition these points into two subspaces, one corresponding to the fair coin being chosen, and one corresponding to the biased coin being chosen. At time 1, Alice knows what coin she chose, so, for example, at the point $(r_1, 1)$, she only considers $(r_2, 1)$ possible. She puts the obvious probability on these two points, judging them to be equally likely. At time 2, Alice knows the outcome of the coin toss. Her probability spaces are singletons, and she puts probability 1 on the one point in each probability space.

Bob never learns what happens. At each time m, he considers all the points $(r_1, m), \ldots, (r_4, m)$ possible. He can divide these points up into two probability spaces, as we have seen, but he also has the possibility of dividing it up into four probability spaces, just like Alice, at time 2. This will mean that, after the coin toss, he believes that the probability of heads is either 0 or 1, although he does not know which.

We can apply the same methodology to the Second-Ace problem from the introduction. As Shafer points out [Sha85], the ambiguities in the problem become clearer when we specify the protocol that Alice and Bob are following. One possible protocol that is consistent with the story is that, at step 1, Alice is dealt two cards. At step 2, Alice tells Bob whether or not she has an ace. Then, at step 3, Alice tells Bob she has the ace of spades if she has it and otherwise she says she hasn't got it. In this case, the analysis described in the introduction is correct. There are six possible pairs of cards that Alice could have been dealt; each one determines a unique run. Since we assume a fair deal, each of these runs has probability 1/6. Bob conditions on his information, as discussed above. When Alice tells Bob that she has an ace in step 2, then at time 2, Bob can elimate the run where Alice was not dealt an ace, and his conditional probability that Alice has two aces is indeed 1/5, as suggested in the story. At time 3, Bob can eliminate two more runs (the runs where Alice does not have the ace of spades), and he assesses the probability that Alice has both aces as 1/3. Notice, however, the concern as to what happens if Alice had told Bob that she has the ace of hearts does not arise. This cannot happen, according to the protocol.

Now suppose we consider a different protocol (although, again, one consistent with the story). Again, at step 2, Alice tells Bob whether or not she has an ace. However, now, at step 3, Alice tells Bob which ace she has if she has an ace (otherwise we can assume she says nothing). This still does not completely specify the protocol. What does Alice tell Bob at step 3 if she has both aces? One possible response is to say that she says "I have the ace of hearts" and "I have the ace of spades" with equal probability in this case. With this protocol, there are seven runs: each of the six possible pairs of cards that Alice could have been dealt determines a unique run, with the exception of the case where Alice is dealt two aces, for which there are two possible runs (depending on which ace Alice tells Bob she has). Each run has probability 1/6 except for the two runs where Alice was dealt two aces, which each have probability 1/12.

Again, Bob conditions on the information he receives, and again, at time 2, his conditional probability that Alice has two aces is 1/5. What is the situation at time 3, after Alice says she has the ace of spades? In this case Bob considers three points possible, those in the two runs where Alice has the ace of spades and a deuce, and the point in the run where Alice has both aces and tells Bob she has the ace of spades. Notice, however, that after conditioning, the probability of the point on the run where

Alice has both aces is $1/5$, while the probability of each of the other two points is $2/5$! This is because the probability of the run where Alice holds both aces and tells Bob she has the ace of spades is $1/12$, half the probability of the runs where Alice holds only one ace. Thus, Bob's probability that Alice holds both aces at time 3 is $1/5$, not $1/3$, if this is the protocol. The fact that Alice says she has the ace of spades does not change Bob's assessment of the probability that she has two aces. Similarly, if Alice says that she has the ace of hearts at step 3, the probability that she has two aces remains at $1/5$.

Suppose we modify Alice's protocol so that, if she has both aces, the probability that she says she has the ace of spades is α. In this case, a similar analysis shows that Bob's probability that Alice holds both aces at time 3 is $\alpha/(\alpha+2)$. In the special case we started with, we have $\alpha = 1/2$, and $\alpha/(\alpha+2)$ reduces to $1/5$. If $\alpha = 0$, then Alice never says "I have the ace of spades" if she has both aces. In this case, Bob's probability that Alice has both aces is 0, as we would expect. If $\alpha = 1$, which corresponds to Alice saying "I have the ace of spades" either if she has only the ace of spades or if she has both aces, Bob's probability that Alice has both aces is $1/3$.

What if Alice does not choose which ace to say probabilistically, but uses some deterministic protocol which Bob does not know? In this case, all Bob can say is that the probability that Alice holds both aces is either 0 or $1/3$, depending on which protocol Alice is following.

Finally, let's consider the Monty Hall puzzle. The standard argument says that you ought to switch: you lose by switching if the goat is behind the door you've picked; otherwise you gain. Thus, the probability of gaining is $2/3$. Is this argument reasonable? It depends. I'll just sketch the analysis here, since it's so similar to that of the second-ace problem.

What protocol describes the situation? We assume that at the first step Monty places a car behind one door and a goat behind the other two. For simplicity, let's assume that the car is equally likely to be placed behind any door. At step 2, you choose a door. At step 3, Monty opens a door (one with a goat behind it other than the one you chose). Finally, at step 4, you must decide if you'll take what's behind your door or what's behind the other unopened door. Again, to completely specify the protocol, we have to say what Monty does if the door you choose has a car behind it (since then he can open either of the other two doors). Suppose we take the probability of him opening door j if you choose door i to be α_{ij} (where α_{ii} is 0—Monty never opens the door you've chosen). Computations similar to those above show that, if you initially take door i and Monty then opens door j, the probability of you gaining by switching is $1/(\alpha_{ij}+1)$. If $\alpha_{ij} = 1/2$, then we get $2/3$, just as in the standard analysis. Intuitively, the standard analysis presumes that learning which door Monty opens does not affect your prior probability that the car is equally likely to be behind each door. If $\alpha_{ij} = 0$, then you are certain that the car can't be behind the door you opened once Monty opens door j. Not surprisingly, you certainly should switch; you are certain to win in this case. On the other hand, if $\alpha_{ij} = 1$, you are just as likely to win by switching as not. Since, with any choice of α_{ij}, you are at least as likely to win by switching as by not switching, it seems that you ought to switch.

However, as pointed out in [MCDD91], this analysis is carried out under the assumption that, at step 3, Monty must open another door. If we instead modify Monty's protocol so that he is far more likely to open another door if the door that you chose has a car behind it, then this analysis is no longer correct. In particular, if we assume that Monty only opens another door if the door that you chose has the car behind it (in order to tempt you away from the "good" door), then you should clearly stick with your original choice.

6 Conclusion

Uncertainty is pervasive, and it is unlikely that any one approach can handle all of its complexities. Nevertheless, I hope I have made the case that thinking in terms of protocols, runs and systems can go a long way towards clarifying things. I believe the framework is rich enough to capture much of the reasoning that goes on, natural enough to allow us to model easily many interesting examples, while forcing us to make explicit issues that, when left implicit or unsettled, cause many of the ambiguities and difficulties in reasoning about uncertainty.

As I said in the introduction, I could cover only a small selection of topics here. Let me close with a few words about some related issues that I would have liked to cover, given more space and time:

- Since the focus here has been on semantics and modeling issues, nothing has been said about axiomatizations. Complete axiomatizations and decision procedures are known for most of the logics presented here. In particular, axiomatizations for modal (multi-agent) logics of knowledge and belief are presented in [HM92], the case of knowledge and time is considered in [HV89], the case of probability is considered in [FHM90], and the combination of knowledge and probability is considered in [FH94a].

- As Pearl and others have stressed, one of the most important aspects of probabilistic reasoning involves reasoning about dependencies, independences, and causality. Indeed, Pearl makes a strong case that much human reasoning about uncertainty involves this type of reasoning. *Bayesian networks* have been shown to be a powerful tool for representing (in)dependencies graphically, and many algorithms have been developed for computing with Bayesian networks and learning Bayesian networks. This is an active area of research that holds a great deal of promise; I would encourage the interested reader to consult [Pea88] for an overview.

- The focus here has been on representing uncertainty with probability. There is nothing to stop us from using different, perhaps more qualititative, representations. For example, we might consider a *preferential ordering* on worlds [KLM90], where, intuitively, one world is preferred to another if it is significantly more likely. Alternatively, we could use Dempster-Shafer belief functions [Sha76], possibility measures [DP90], comparative probability [Fin73], or ordinal rankings [Spo87]. Recently, Nir Friedman and I have introduced *plausibility measures* [FH95b], which generalize all of these approaches. A plausibility measure associates with each set its *plausibility*, which is just some element of a partially ordered set. By taking the set to be $[0,1]$ (with the usual ordering) and imposing some extra requirements on the plausibility, we can get back probability measures. A discussion of how to use preferential orderings rather than probability in the context of systems can be found in [FH94b]; in [FH95a], this discussion is redone using plausibility. The changes required are all quite straightforward.

References

[BF82] M. Bar-Hillel and R. Falk. Some teasers concerning conditional probabilities. *Cognition*, 11:109–122, 1982.

[DP90] D. Dubois and H. Prade. An introduction to possibilistic and fuzzy logics. In G. Shafer and J. Pearl, editors, *Readings in Uncertain Reasoning*. Morgan Kaufmann, San Francisco, Calif., 1990.

[FH94a] R. Fagin and J. Y. Halpern. Reasoning about knowledge and probability. *Journal of the ACM*, 41(2):340–367, 1994.

[FH94b] N. Friedman and J. Y. Halpern. A knowledge-based framework for belief change. Part I: foundations. In R. Fagin, editor, *Theoretical Aspects of Reasoning about Knowledge: Proc. Fifth Conference*, pages 44–64. Morgan Kaufmann, San Francisco, Calif., 1994.

[FH95a] N. Friedman and J. Y. Halpern. Modeling belief in dynamic systems. Part I: foundations. Technical Report RJ9965, IBM, 1995. Available by anonymous ftp from `starry.stanford.edu/pub/nir` or via WWW at `http://robotics.stanford.edu/users/nir`.

[FH95b] N. Friedman and J. Y. Halpern. Plausibility measures: a user's manual. In P. Besnard and S. Hanks, editors, *Proc. Eleventh Conference on Uncertainty in Artificial Intelligence (UAI '95)*. Morgan Kaufmann, San Francisco, Calif., 1995.

[FHM90] R. Fagin, J. Y. Halpern, and N. Megiddo. A logic for reasoning about probabilities. *Information and Computation*, 87(1/2):78–128, 1990.

[FHMV95] R. Fagin, J. Y. Halpern, Y. Moses, and M. Y. Vardi. *Reasoning about Knowledge*. MIT Press, Cambridge, Mass., to appear, 1995.

[Fin73] T. L. Fine. *Theories of Probability.* Academic Press, New York, 1973.

[Fre65] J. E. Freund. Puzzle or paradox? *American Statistician*, 19(4):29–44, 1965.

[HC68] G. E. Hughes and M. J. Cresswell. *An Introduction to Modal Logic.* Methuen, London, 1968.

[HF89] J. Y. Halpern and R. Fagin. Modelling knowledge and action in distributed systems. *Distributed Computing*, 3(4):159–179, 1989. 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".

[Hin62] J. Hintikka. *Knowledge and Belief.* Cornell University Press, Ithaca, N.Y., 1962.

[HM92] J. Y. Halpern and Y. Moses. A guide to completeness and complexity for modal logics of knowledge and belief. *Artificial Intelligence*, 54:319–379, 1992.

[HT93] J. Y. Halpern and M. R. Tuttle. Knowledge, probability, and adversaries. *Journal of the ACM*, 40(4):917–962, 1993.

[HV89] J. Y. Halpern and M. Y. Vardi. The complexity of reasoning about knowledge and time, I: lower bounds. *Journal of Computer and System Sciences*, 38(1):195–237, 1989.

[KLM90] S. Kraus, D. Lehmann, and M. Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44:167–207, 1990.

[MCDD91] J. P. Morgan, N. R. Chaganty, R. C. Dahiya, and M. J. Doviak. Let's make a deal: the player's dilemma (with commentary). *The American Statistician*, 45(4):284–289, 1991.

[Pea88] J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, San Francisco, Calif., 1988.

[Sav91] M. vos Savant. Ask Marilyn. *Parade Magazine*, Sept. 9, 1990; Dec. 2, 1990; Feb. 17, 1991.

[Sha76] G. Shafer. *A Mathematical Theory of Evidence.* Princeton University Press, Princeton, N.J., 1976.

[Sha85] G. Shafer. Conditional probability. *International Statistical Review*, 53(3):261–277, 1985.

[Spo87] W. Spohn. Ordinal conditional functions: a dynamic theory of epistemic states. In W. Harper and B. Skyrms, editors, *Causation in Decision, Belief Change and Statistics*, volume 2, pages 105–134. Reidel, Dordrecht, Holland, 1987.