# A Knowledge-Based Framework for Belief Change Part I: Foundations

Nir Friedman
Stanford University
Dept. of Computer Science
Stanford, CA 94305-2140
nir@cs.stanford.edu

Joseph Y. Halpern
IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120–6099
halpern@almaden.ibm.com

**Abstract**

We propose a general framework in which to study belief change. We begin by defining belief in terms of knowledge and plausibility: an agent believes $\varphi$ if he knows that $\varphi$ is true in all the worlds he considers most plausible. We then consider some properties defining the interaction between knowledge and plausibility, and show how these properties affect the properties of belief. In particular, we show that by assuming two of the most natural properties, belief becomes a KD45 operator. Finally, we add time to the picture. This gives us a framework in which we can talk about knowledge, plausibility (and hence belief), and time, which extends the framework of Halpern and Fagin [HF89] for modeling knowledge in multi-agent systems. We show that our framework is quite expressive and lets us model in a natural way a number of different scenarios for belief change. For example, we show how we can capture an analogue to prior probabilities, which can be updated by "conditioning". In a related paper, we show how the two best studied scenarios, *belief revision* and *belief update*, fit into the framework.

## 1 Introduction

The study of *belief change* has been an active area in philosophy and in artificial intelligence [Gär88, KM91] and, more recently, in game theory [Bic88, Sta92]. The focus of this research is to understand how an agent should revise his beliefs as a result of getting new information. In the literature, two instances of this general phenomenon have been studied in detail: *Belief revision* [AGM85, Gär88] attempts to describe how an agent should accommodate a new belief (possibly inconsistent with his other beliefs) about a static world. *Belief update* [KM91], on the other hand, attempts to describe how an agent should change his beliefs as a result of learning about a change in the world. Belief revision and belief update describe only two of the many scenarios in which beliefs change. Our goal is to construct a framework to reason about belief change in general. This paper describes the details of that framework. In a companion paper [FH93a] we consider the special cases of belief revision and update in more detail.

Perhaps the most straightforward approach to belief change is to simply represent an agent's beliefs as a closed set of formulas in some language and then put constraints on how the beliefs can change. This is essentially the approach taken in [AGM85, Gär88]; as these papers show, much can be done with this framework. The main problem with this approach is that it does not provide a good semantics for belief. As we hope to show in this paper and in [FH93a], such a semantics can give us a much deeper understanding of how and why beliefs change.

One standard approach to giving semantics to belief is to put a plausibility ordering on a set of worlds (intuitively, the worlds the agent considers possible). Using plausibility orderings, we can interpret statements such as "it typically does not rain in San Francisco in the summer". Roughly speaking, a statement such as "$\varphi$ typically implies $\psi$" is true at a given world if $\psi$ is true in the most plausible worlds where $\varphi$ is true. Various authors [Spo87, KM91, Bou92] have then interpreted "the agent believes $\varphi$" as "$\varphi$ is true in the most plausible worlds that the agent considers possible". Under this interpretation, the agent believes $\varphi$ if *true* typically implies $\varphi$.[1]

By modeling beliefs in this way, there is an assumption that the ordering is part of the agent's epistemic state. (This assumption is actually made explicitly in [KLM90, Bou92].) This implies that the ordering is *subjective*, that is, it describes the agent's estimate of what the plausible states are. But actually, an even stronger assumption is being made: namely, that the agent's epistemic state is characterized by a *single* plausibility ordering. We feel this latter assumption makes the models less expressive than they need to be. In particular, they cannot represent a situation where the agent is not sure about what is plausible, such as "Alice does not know that it typically does not rain in San Francisco in the summer". To do this, we need to allow Alice to consider several orderings possible; in some it typically does not rain and in others it typically does.[2] As we shall see, this extra expressive power is necessary to capture some interesting scenarios of belief change.

To capture such situations, in addition to plausibility orderings, we add a standard accessibility relation to represent *knowledge*. Once we have knowledge in the picture, we define belief by saying that an agent *believes* $\varphi$ if she *knows* that $\varphi$ is typically true. That is, according to all the orderings she considers possible, $\varphi$ is true in the most plausible worlds. Notice that in the special case where the agent considers only one ordering possible, our definition of belief is equivalent to the definition of belief as truth in the most plausible worlds.

The properties of belief depend on how the plausibility ordering interacts with the accessibility relation that defines knowledge. We study these interactions, keeping in mind the interpretation of plausibility in terms of qualitative probability [Pea89]. In view of this interpretation, it is perhaps not surprising that many of the issues studied by Fagin and Halpern [FH88] when considering the interaction of knowledge and probability also arise in our framework. There are, however, a number of new issues that arise in our framework due to the interaction between knowledge and belief. As we shall see, if we take what are perhaps the most natural restrictions on this interaction, our notion of belief is characterized by the axioms of the modal logic KD45 (where an agent has complete introspective knowledge about her beliefs, but may have false beliefs). Moreover, the interaction between knowledge and belief satisfies the standard properties considered by Kraus and Lehmann [KL88]. Although our major goal is not an abstract study of the properties of knowledge and belief, we view the fact that we have a concrete interpretation under which these properties can be studied to be an important side-benefit of our approach.

Having a notion of belief is not enough in order to study belief change. We want a framework that captures the beliefs of the agent before and after the change. This is achieved by introducing *time* explicitly into the framework. The resulting framework is an extension of the framework of [HF89] for modeling knowledge in multi-agent systems, and allows to to talk about knowledge, plausibility (and hence belief), and time.

As we show by example, with knowledge, plausibility and time represented explicitly in the framework we have a powerful and expressive framework for capturing belief change. One important feature of our approach is that it gives us the tools to study how plausibility changes over time. We focus here on one particular way this can happen, which is an analogue to the Bayesian approach of updating prior

[1] The technique of putting an ordering on worlds has also been used to model counterfactuals, conditionals and nonmonotonic inference [Lew73, Sho87, KLM90, Pea89]. We focus here on its application to modeling belief.

[2] In fact, this issue is discussed by Boutilier [Bou92], although his framework does not allow him to represent such a situation.

probabilities by conditioning. In this process the plausibility ordering before the change dictates the plausibility ordering after the change. Thus, the prior encodes all the plausibility orderings that can arise in the system. As we show, many situations previously studied in the literature, such as *diagnostic reasoning* [Kle90] and the *prisoner's dilemma* from game theory, can be easily captured by using such prior plausibilities.

The rest of this paper is organized as follows. In the next section, we review the syntax and semantics of the standard approach to modeling knowledge using Kripke structures and show how plausibility can be added to the framework. Much of our technical discussion of axiomatizations and decision procedures is closely related to that of [FH88]. In Section 3, we present our full framework which adds plausibility to the framework of [HF89] for modeling knowledge (and time) in multi-agent systems. In Section 4 we introduce prior plausibilities and show how they can be used. We conclude in Section 5 with some discussion of the general approach. In Appendix A we provide complete axiomatizations for a number of variants of the logic of knowledge and plausibility. In Appendix B, we examine the relationships between *ranked* plausibility orderings, which assume a total ordering on the set of worlds, and various probabilistic approaches to dealing with the problem of conditioning on events of measure 0.

## 2 Knowledge and plausibility

In this section, we briefly review the standard models for knowledge, describe a notion of plausibility, and then show how to combine the two notions. Finally, we compare the derived notion of belief with previous works on the subject.

### 2.1 The logic of knowledge

The syntax for the logic of knowledge is simple: we start with primitive propositions and close off under conjunction, negation, and the modal operators $K_1, \dots, K_n$. A formula such as $K_i\varphi$ is read "agent $i$ knows $\varphi$". We denote the resulting language as $\mathcal{L}^K$.

The semantics for this language is given by means of *Kripke structures*. A *Kripke structure for knowledge* is a tuple $(W, \pi, \mathcal{K}_1, \dots, \mathcal{K}_n)$, where $W$ is the set of worlds that can be thought of as distinct situations, or different ways that the world can be, $\pi(w)$ is a truth assignment for primitive propositions in each world $w \in W$, and $\mathcal{K}_i$ are equivalence relations among worlds.[3] For convenience, we define $\mathcal{K}_i(w) = \{w' \mid (w, w') \in \mathcal{K}_i\}$.

We now assign truth values to formulas at each world in the structure. We write $(M, w) \models \varphi$ if the formula $\varphi$ is true at a world $w$ in the Kripke structure $M$.

- $(M, w) \models p$ for primitive proposition $p$ if $\pi(w)(p) = \mathit{true}$.
- $(M, w) \models \neg\varphi$ if not $(M, w) \models \varphi$.
- $(M, w) \models \varphi \land \psi$ if $(M, w) \models \varphi$ and $(M, w) \models \psi$.
- $(M, w) \models K_i\varphi$ if for all $w' \in \mathcal{K}_i(w)$, $(M, w') \models \varphi$.

The last clause captures the intuition that $\varphi$ is known exactly when it is true in all possible worlds.

Let $\mathcal{M}^K$ be the class of Kripke structures described. We say that $\varphi \in \mathcal{L}^K$ is valid in some $M \in \mathcal{M}^K$ if for all $w$, $(M, w) \models \varphi$. We say that $\varphi \in \mathcal{L}^K$ is *valid* in $\mathcal{M}^K$ if it is valid in all models $M \in \mathcal{M}^K$.

[3] In general, we may not want to require the $\mathcal{K}_i$'s to be equivalence relations. We focus on this case here, since it is of most interest to us. Many of our technical results have natural analogues for other assumptions on the $\mathcal{K}_i$ relations.

We say that $\varphi$ is satisfiable in $\mathcal{M}^K$ if there is a model $M \in \mathcal{M}^K$ and $w$, such that $(M, w) \models \varphi$. It is well known (see, for example, [HM92]) that the valid formulas in $\mathcal{L}^K$ over $\mathcal{M}^K$ are characterized by the modal logic S5 (which is defined formally in Section 2.4).

## 2.2 Plausibility spaces

We want to extend the logic of knowledge by adding plausibility. To do this, we must first introduce *plausibility spaces*, which can be viewed as a qualitative analogue of probability spaces [Pea89]. For now, we discuss these structures in the abstract; in the next section, we combine them with knowledge. A plausibility space describes a qualitative measure of plausibility over some set of alternatives (one can think of them as possible worlds). This measure is qualitative in the sense that it compares the plausibility of alternatives but does not provide an exact degree of plausibility. Formally, a plausibility space is a pair $(\Omega, \preceq)$, where $\Omega$ is a set and $\preceq$ is a *pre-order* on $\Omega$, that is, a reflexive and transitive relation over $\Omega$. As usual, we write $s \prec s'$ if $s \preceq s'$ and it is not the case that $s' \preceq s$. Intuitively, $s \prec s'$ if $s$ is strictly more plausible than $s'$.[4]

Given two subsets $S$ and $T$ of $\Omega$ (which we can think of as representing events), we would like to define $S \rightarrow T$ to hold in $(\Omega, \preceq)$ if $T$ is plausible, given $S$. Intuitively, we want $S \rightarrow T$ to hold if all the minimal points (with respect to $\preceq$) of $S$ are in $T$. Unfortunately, if $S$ is infinite it may not have any minimal points. We do not necessarily want $S \rightarrow T$ to hold if $S$ has no minimal points (since this would give $\rightarrow$ some properties not in accord with our intuitions). Thus, we follow the standard technique [Lew73, Bur81] of saying that a plausibility structure $(\Omega, \preceq)$ satisfies $S \rightarrow T$ if for every point $s \in S$ there is a point $t \in T \cap S$ such that $t \preceq s$, and there is no point $u \preceq t$ such that $u \in S - T$. Note that if $S$ has no infinite descending chain (that is, if there is no sequence of points $s_1, s_2, s_3, \ldots$ in $S$ such that $\ldots s_3 \prec s_2 \prec s_1$) then this definition reduces to saying that the minimal points of $S$ are in $T$.

As noted above, plausibility spaces can be viewed as a qualitative analogue of probability spaces (see [Pea89, Gef92]). The intuition is that $s_1 \prec s_2$ holds if $s_1$ is much more probable than $s_2$; and that $S \rightarrow T$ holds whenever $\Pr(T|S)$ has high probability. A naive way of capturing this intuition is by fixing a small $\epsilon$ and defining $s_1 \prec s_2$ to hold if $\Pr(s_2)/\Pr(s_1) \leq \epsilon$. The problem is that, with this definition, we can easily construct examples where $S \rightarrow T$ holds, and yet $\Pr(T|S)$ can be arbitrarily small. The standard way to overcome this problem [Pea89] is to consider, not one $\epsilon$, but a sequence of $\epsilon$'s converging to 0. More formally, consider a family $\{\Pr_\epsilon : \epsilon > 0\}$ of probability distributions on $\Omega$, parameterized by $\epsilon$. We then define $s_1 \prec s_2$ to hold if $\lim_{\epsilon \rightarrow 0} \Pr_\epsilon(s_2)/\Pr_\epsilon(s_1) = 0$. It is easy to see that with this definition, if $\Omega$ is finite, we have that $S \rightarrow T$ holds if and only if $\lim_{\epsilon \rightarrow 0} \Pr_\epsilon(T|S) = 1$. Thus, we can think of $S \rightarrow T$ as saying that the probability of $T$ given $S$ is arbitrarily close to, but not necessarily equal to, 1. We remark that it can be shown that any plausibility ordering on a finite space $\Omega$ can be characterized in this way. That is, given $\preceq$, we can choose a family of probability distributions parameterized by $\epsilon$, such that $\lim_{\epsilon \rightarrow 0} \Pr_\epsilon(s_2)/\Pr_\epsilon(s_1) = 0$ exactly when $s_1 \prec s_2$.

**Example 2.1:** The circuit diagnosis problem has been well studied in the literature (see [DH88] for an overview). Consider a circuit that contains $n$ logical components $c_1, \ldots, c_n$. Our target is to construct a plausibility ordering over the possible failures of the circuit. A *failure* is taken to be a set of faulty components. We assume that failures of individual components are independent of one another. If we also assume that the probability of each component failing is the same, we can construct a plausibility ordering as follows: Let $\epsilon$ be the probability that a single component fails. Then the probability of

[4] We follow the standard notion for plausibility [Lew73, KLM90, Pea89], which uses the (perhaps confusing) convention of placing the more plausible event on the left of the $\prec$ operator.

a failure $f = \{c_{i_1}, \ldots, c_{i_k}\}$ is $\Pr_\epsilon(f) = \epsilon^{|f|}(1-\epsilon)^{(n-|f|)}$. For two failures $f_1$ and $f_2$, we have that $\lim_{\epsilon\to 0} \Pr_\epsilon(f_1)/\Pr_\epsilon(f_2) = 0$ if and only if $|f_2| < |f_1|$. Thus, when the probability of component failure is small and unknown, it is reasonable to use a plausibility ordering that compares failures by their cardinality.

In some situations it might be unreasonable to assume that all components have the same probability of failure. Thus, we might assume that for each component $c_i$ there is a probability $\epsilon_i$ of failure. If we assume independence, then given $\vec{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$, the probability of a failure $f$ is $\Pr_{\vec{\epsilon}}(f) = \Pi_{c_i \in f} \epsilon_i \Pi_{c_i \notin f}(1-\epsilon_i)$. If we examine the probabilities as the different $\epsilon_i$ get smaller we notice that $\lim_{\vec{\epsilon}\to 0} \Pr_{\vec{\epsilon}}(f_1)/\Pr_{\vec{\epsilon}}(f_2) = 0$ if and only if $f_2 \subset f_1$, i.e., if $f_1$ contains all the failures in $f_2$ and more. In this case, since we do not assume any relations among probability of failure of different components, it is not possible to compare failures unless one is a subset of the other. If we introduce constraints such as $\epsilon_1 \leq \epsilon_2$ (i.e., $c_1$ is more reliable than $c_2$), then we can construct more informative plausibility orderings. ▌

Given a probability space $\mu$ on $\Omega$ and a subset $S$ of $\Omega$ such that $\mu(S) \neq 0$, we can define the conditional probability measure $\mu|_S$ on $S$ in the standard way. Given a plausibility space $P = (\Omega, \preceq)$, we want to similarly define a *conditional plausibility space* $P|_S = (S, \preceq_S)$. We do this by taking $\preceq_S$ to be the restriction of $\preceq$ to $S$. Thus, if $s_1, s_2 \in S$, then $s_1 \preceq_S s_2$ if and only if $s_1 \preceq s_2$. This notion of conditionalization is closely tied with that of conditionalization in probability: Assume that $P = (\Omega, \preceq)$ corresponds (in the manner described above) to the family $\{\Pr_\epsilon\}$. Then, it is not hard to verify that $P|_S$ corresponds to the family $\{\Pr_\epsilon|_S\}$.

In the literature a special class of plausibility spaces received special attention. A plausibility space is *ranked* if $\preceq$ is a *total* pre-order, i.e., for any $s_1, s_2 \in \Omega$ either $s_1 \preceq s_2$ or $s_2 \preceq s_1$. It is known [GMP93] that $P$ is ranked if and only if it corresponds to a family $\{\Pr_\epsilon\}$ such that for any $S$ and $T$ the limit $\lim_{\epsilon\to 0} \Pr_\epsilon(T|S)$ is defined. Ranked plausibility spaces are closer in spirit to probability spaces than arbitrary plausibility spaces since they correspond to a linear ordering on events. In Appendix B we examine the relationship between ranked plausibility spaces and a number of other probabilistic approaches to dealing with the problem of conditioning on events of measure 0, including nonstandard probability functions that can assign infinitesimal values [LM92], Popper functions [Fra76], and the lexicographic probabilities approach of Blume, Brandenburger, and Dekel [BBD91].

### 2.3 Combining knowledge and plausibility

We now define a logic that combines knowledge and plausibility. Let $\mathcal{L}^{KP}$ be the language obtained by starting with primitive propositions, and closing off under conjunction, negation, and the operators $K_i$ and $\rightarrow_i$, $i = 1, \ldots, n$. Note that we have a different plausibility operator for each agent. We read $\varphi \rightarrow_i \psi$ as "under agent $i$'s plausibility ordering, $\varphi$ typically implies $\psi$".

We define a *(Kripke) structure (for knowledge and plausibility)* to be a tuple $(W, \pi, \mathcal{K}_1, \ldots, \mathcal{K}_n, \mathcal{P})$ where $W$, $\pi$ and $\mathcal{K}_i$ are just as in Kripke structures for knowledge, while $\mathcal{P}$ is a function that assigns a plausibility space to each agent at each world. Intuitively, the structure $\mathcal{P}(w,i) = (\Omega_{(w,i)}, \preceq_{(w,i)})$ captures agent $i$'s plausibility ordering in the world $w$. For now we allow $\Omega_{(w,i)}$ to be an arbitrary subset of $W$. We discuss some possible restrictions on $\Omega_{(w,i)}$ below. It is reasonable to ask at this point where the plausibility spaces $\mathcal{P}(w,i)$ are coming from, and why we need a different one for each agent at each world? The answer to this question depends very much on the intended application. We defer further discussion of this issue until later.

We can now give semantics to formulas in $\mathcal{L}^{KP}$ in Kripke structures for knowledge and plausibility. The only new feature to deal with are formulas of the form $\varphi \rightarrow_i \psi$. Intuitively, this is true at a world

$w$ if $\psi$ is true at all the most plausible $\varphi$-worlds in $\Omega_{(w,i)}$. Given a world $w$ and a formula $\varphi$, define $S^{\varphi}_{(w,i)} = \{w' \in \Omega_{(w,i)} \mid (M,w') \models \varphi\}$; this is the set of worlds in $\Omega_{(w,i)}$ where $\varphi$ is true. We then define

$$(M,w) \models \varphi \rightarrow_i \psi \text{ if } S^{\varphi}_{(w,i)} \rightarrow S^{\psi}_{(w,i)} \text{ holds in } \mathcal{P}(w,i).$$

We say that an agent *believes* $\varphi$ if he knows that $\varphi$ is true in the most plausible worlds. Thus, we define $B_i\varphi$, read "agent $i$ believes $\varphi$", as an abbreviation for $K_i(\mathit{true} \rightarrow_i \varphi)$. This definition matches the intuition since, at world $w$, we have that $\mathit{true} \rightarrow_i \varphi$ holds only if $\varphi$ is true at the most plausible worlds according to $\mathcal{P}(w,i)$.

The following example illustrates some of the expressive power of this language.

**Example 2.2:** Consider an agent that performs *diagnosis* of an electrical circuit. The agent can set the values of input lines of the circuit and observe the output values. The agent then compares the actual output values to the expected output values and attempts to locate faulty components. We model the situation as follows. Each possible world $w$ is composed of two parts: $\mathit{fault}(w)$, the set of faulty components in $w$, and $\mathit{obs}(w)$, the input-output relation the agent has observed. We consider only worlds where the test results are consistent with the set of faulty components. The agent knows which test he performed and the results he observed. Therefore, we have $(w,w') \in \mathcal{K}_1$ if $\mathit{obs}(w) = \mathit{obs}(w')$. Assuming that there is always more than one explanation for the observed faulty behavior (as is typically the case), the agent can never *know* exactly which components are faulty, but he may have *beliefs* on that score.

Using the plausibility orderings described in Example 2.1 we can construct two possible structures $M_1$ and $M_2$. In both structures we set $\Omega_{(w,1)} = \mathcal{K}_1(w)$. We define $\preceq$ in $M_1$ so that $w_1 \preceq_{(w,1)} w_2$ if $|\mathit{fault}(w_1)| \leq |\mathit{fault}(w_2)|$ and in $M_2$ so that $w_1 \preceq_{(w,1)} w_2$ if $\mathit{fault}(w_1) \subseteq \mathit{fault}(w_2)$. It is easy to see that, in both structures, if the test results correspond to the predicted results, then the agent believes that the circuit is faultless. If the agent detects an error, he believes that it is caused by one of the *minimal explanations* of his observations, where the notion of minimality differs in the two structures. It is interesting to note that this description captures, albeit somewhat simplistically, the assumptions made in model-based diagnostics. In particular, $M_1$ captures the assumptions made in [Kle90], while $M_2$ captures the assumptions made in [Rei87]. ∎

Kripke structures for knowledge and plausibility are quite similar to the Kripke structures for knowledge and probability introduced in [FH88]. The only difference is that in Kripke structures for knowledge and probability, $\mathcal{P}(w,i)$ is a probability space rather than a plausibility space. In [FH88], various natural restrictions on the interactions between the probability spaces $\mathcal{P}(w,i)$ and the accessibility relations $\mathcal{K}_i$ are investigated. Here we similarly investigate restrictions on the interaction between the plausibility spaces and the accessibility relations. Not surprisingly, some of these conditions are exact analogues to conditions investigated in [FH88]. We also examine some conditions on plausibility ordering without regard to the knowledge. Such conditions were described before in [Lew73] and other works on conditionals. For consistency, we use a naming scheme similar to those used in [Lew73] and [FH88].

$\Omega_{(w,i)}$ consists of all worlds to which agent $i$ assigns some degree of plausibility in world $w$. We would not expect the agent to place a positive probability on worlds that he considers impossible. Similarly, he would not want to consider as plausible (even remotely) a world he knows to be impossible. This intuition leads us to the following condition, called CONS for *consistency* (following [FH88]):

**CONS**: For all worlds $w$, $\Omega_{(w,i)} \subseteq \mathcal{K}_i(w)$.[5]

[5] We remark that CONS is inappropriate if we use $\rightarrow$ to model, not plausibility, but counterfactual conditions, as is done

A consequence of assuming CONS is a stronger connection between knowledge and belief. Since CONS implies that the most plausible worlds are in $\mathcal{K}_i(w)$, it follows that if the agent knows $\varphi$ he also believes $\varphi$. (Indeed, as we shall see, this condition characterizes CONS.)

While CONS ensures that the agent does not assign plausibility to impossible situations, We may also want the agent to consider *some* worlds as plausible. Otherwise, the agent does not have most plausible worlds and this amount to saying that the agent does not believe that any world is the real world. We call this condition NORM for *normality* (following [Lew73]):

**NORM**: For all worlds $w$, $\Omega_{(w,i)} \neq \emptyset$.

We can strengthen this condition somewhat to one that says that the agent always considers the real world possible. We call this condition REF for *reflexiveness* (following [Lew73]):

**REF**: For all worlds $w$, $w \in \Omega_{(w,i)}$.

As we said in the introduction, many previous works using conditionals assumed (implicitly or explicitly) that the agent considers only one plausibility ordering possible. This is captured by an assumption called SDP (following [FH88]) for *state determined plausibilities*:

**SDP**: For all $w$ and $w'$, if $(w, w') \in \mathcal{K}_i$ then $\mathcal{P}(w, i) = \mathcal{P}(w', i)$.

It is easy to see that SDP implies that an agent knows his plausibility ordering. In particular, as we shall see, under SDP, we have that $\varphi \rightarrow_i \psi$ implies $K_i(\varphi \rightarrow_i \psi)$.

It is easy to verify that the structures described in Example 2.2 satisfy CONS, REF, and SDP. As mentioned in the introduction, SDP is not appropriate in all situations; at times we may want to allow the agent to consider possible several plausibility orderings. To capture this, we need to generalize SDP. The following example might help motivate the formal definition.

**Example 2.3:** This is a variation on the Liar's Paradox. On a small Pacific island there are two tribes, the Rightfeet and the Leftfeet. The Rightfeet are known to usually tell the truth, while the Leftfeet are known to usually lie. Alice is a visitor to the island. She encounters a native, Bob, and discusses with him various aspects of life on the island. Now, Alice does not know what tribe Bob is in. Thus, she considers it possible that both Bob is a Rightfoot and that he is a Leftfoot. In the first case, she should believe what he tells her and in the second she should be skeptical.

Thus, we can imagine Alice's worlds partitioned into two sets according to Bob's tribe. Alice has a plausibility orderings on the set $W_R$ of worlds she considers possible where Bob is a Rightfoot and another on the set $W_L$ of worlds she considers possible where Bob is a Leftfoot. For each $w' \in W_j$, $j \in \{L, R\}$, $\Omega_{(w',Alice)} = W_j$ and for each $w'' \in W_j$, $\mathcal{P}(w', Alice) = \mathcal{P}(w'', Alice)$. In such a structure, the formula $\neg K_{Alice} \neg(tell(\varphi) \rightarrow_{Alice} \neg\varphi) \wedge \neg K_{Alice} \neg(tell(\varphi) \rightarrow_{Alice} \varphi)$ is satisfiable. On the other hand, in structures satisfying SDP this formula is satisfiable only when $tell(\varphi)$ is false in all the worlds Alice considers plausible. ∎

While this example may seem contrived, in many situations it is possible to extract parameters such as Leftfoot and Rightfoot that determine which conditional statements are true. If the agent does not

by Lewis [Lew73]. If CONS holds, then it is easy to see that $K_i\varphi \Rightarrow K_i(\neg\varphi \rightarrow_i \psi)$ is valid, for all $\psi$. That is, if agent $i$ knows $\varphi$, then he knows that in the most plausible worlds where $\neg\varphi$ is true, $\psi$ is vacuously true, because there are no plausible worlds where $\neg\varphi$ is true. On the other hand, under the counterfactual reading, it makes perfect sense to say "I know the match is dry, but it is not the case that if it were wet, then it would light if it were struck."

know the value of these parameters, she will not necessarily know which conditionals are true at a given world (as was the case in the example above).

This example motivates the condition called *uniformity* in [FH88].

**UNIF**: For all $w$, if $w' \in \Omega_{(w,i)}$ then $\mathcal{P}(w,i) = \mathcal{P}(w',i)$.[6]

Note that SDP and CONS together imply UNIF. If both UNIF and CONS hold, then we can partition $\mathcal{K}_i(w)$ into disjoint clusters, each of which is a separate plausibility space, as in the example above.

When we model uncertainty about the relative plausibility of different worlds this way it is reasonable to demand that any specific plausibility structure can compare all possible worlds, i.e., it is ranked. The RANK assumption is:

**RANK**: For all $w$ and $i$, $\mathcal{P}(w,i)$ is ranked.

While ranked orders are quite natural, they have often been rejected as being too inexpressive [Gin86]. The standard argument for partial orders is as follows: In general, an agent may not be able to determine the relative plausibility of $a$ and $b$. If the plausibility ordering is ranked, the agent is forced to make this determination; with a partial order, he is not. This argument loses much of its force in our framework, once we combine knowledge and plausibility. As we said above, the agent's ignorance can be modeled by allowing him to consider (at least) two total orders possible, one in which $a$ is more plausible than $b$, and one in which $b$ is more plausible that $a$. The agent then believes neither that $a$ is more plausible than $b$ nor that $b$ is more plausible than $a$.

### 2.4 Knowledge and belief

How reasonable is the notion of belief we have defined? In this section, we briefly compare it to other notions considered in the literature; we provide more details in the full paper.

Let $\mathcal{L}^B$ be the language where the only modal operators are $B_1, \ldots, B_n$, and let $\mathcal{L}^{KB}$ be the language where we have $K_1, \ldots, K_n$ and $B_1, \ldots, B_n$ (but no $\rightarrow_i$ operators). Let $\mathcal{M}$ be the set of all Kripke structures for knowledge and plausibility as defined in the previous section, and let $\mathcal{M}^{CONS}$ (resp. $\mathcal{M}^{CONS,NORM}$) be the structures satisfying CONS (resp. CONS and NORM).

Work on belief and knowledge in the literature [Hin62, Lev84, HM92] has focused on the modal systems S5, KD45, D45, and K. We briefly describe these systems here; more details can be found in, for example [Che80, HM92]. The system S5 is composed of the following axioms K1–K5 and rules RK1 and RK2:

**K1.** All substitution instances of propositional tautologies

**K2.** $K_i\varphi \wedge K_i(\varphi \Rightarrow \psi) \Rightarrow K_i\psi$

**K3.** $K_i\varphi \Rightarrow \varphi$

**K4.** $K_i\varphi \Rightarrow K_iK_i\varphi$

**K5.** $\neg K_i\varphi \Rightarrow K_i\neg K_i\varphi$

**RK1.** From $\varphi$ and $\varphi \Rightarrow \psi$ infer $\psi$

**RK2.** From $\varphi$ infer $K_i\varphi$

[6] Note that this condition is not the same as uniformity as defined in [Lew73]; rather, it corresponds in the Lewis terminology to absoluteness.

The system KD45 for belief consists of axioms B1–B5 and rules of inference RB1 and RB2. We obtain all but B3 by replacing the $K_i$ in the description of S5 above by $B_i$. The axiom B3 is a weakening of K3:

**B3.** $\neg B_i \mathit{false}$

By dropping B3 we get the system K45; if, in addition, we drop B4 and B5, we get the system known as K.

When restricting our attention to $\mathcal{L}^B$, the modal system K precisely characterizes the valid formulas of $\mathcal{L}^B$ in the class $\mathcal{M}$. However, in the literature, belief has typically been taken to be characterized by the modal system K45 or KD45, not K. We get K45 by restricting to models that satisfy CONS, and KD45 by further restricting to models that satisfy NORM. Thus, the two requirements that are most natural, at least if we have a probabilistic intuition for plausibility, are already enough to make $B_i$ a KD45 operator.

**Theorem 2.4:** *K (resp., K45, KD45) is a sound and complete axiomatization for $\mathcal{L}^B$ with respect to $\mathcal{M}$ (resp., $\mathcal{M}^{CONS}$, $\mathcal{M}^{CONS,NORM}$).*

Considering knowledge and belief together, Kraus and Lehmann [KL88] have argued that the following axioms are appropriate:

**KB1.** $B_i\varphi \Rightarrow K_iB_i\varphi$

**KB2.** $K_i\varphi \Rightarrow B_i\varphi$

KB1 holds in $\mathcal{M}$ and KB2 is a consequence of CONS. In fact, we have:

**Theorem 2.5:**

*(a) The S5 axioms for the operators $K_i$, the K axioms for the operators $B_i$, together with KB1 gives a sound and complete axiomatization for $\mathcal{L}^{KB}$ with respect to $\mathcal{M}$.*

*(b) The S5 axioms for the operators $K_i$, the K45 axioms for the operators $B_i$, together with KB1 and KB2 gives a sound and complete axiomatization for $\mathcal{L}^{KB}$ with respect to $\mathcal{M}^{CONS}$.*

*(c) The S5 axioms for the operators $K_i$, the KD45 axioms for the operators $B_i$, together with KB1 and KB2 gives a sound and complete axiomatization for $\mathcal{L}^{KB}$ with respect to $\mathcal{M}^{CONS,NORM}$.*

As a corollary, we can show that there is a close relationship between our framework and that of [KL88]. Let $KL$ be the logic of Kraus and Lehmann:

**Corollary 2.6:** *For any $\varphi \in \mathcal{L}^{KB}$, $KL \models \varphi$ if and only if $\mathcal{M}^{CONS,NORM} \models \varphi$.*

Shoham and Moses [SM89] also view belief as being derived from knowledge. The intuition that they try to capture is that once the agent makes a defeasible assumption the rest of his beliefs should follow from his knowledge. In this sense Shoham and Moses concentrate on the implications of an assumption and not on how it was obtained. Indeed, we can even understand their notion as saying that $\varphi$ is believed if it is known to be true in the most plausible worlds. But for them, plausibility is not defined by an ordering. Rather, it is defined in terms of a formula, which can be thought of as characterizing the most plausible worlds. More formally, for a fixed formula $\alpha$, they define $B_i^\alpha\varphi$ to be an abbreviation for $K_i(\alpha \Rightarrow \varphi)$.[7]

[7] Shoham and Moses also examine two variants of this definition.

We can show the connection in a formal manner. Suppose we assume that there are only finitely many primitive propositions, and we restrict attention to ***propositional structures***, ones in which there is at most one world satisfying a given truth assignment to the primitive propositions. We say that a propositional formula $\alpha$ *characterizes* a set $W'$ of worlds in a propositional structure if it is equivalent to the disjunction of the truth assignments at these worlds. Thus, if $\alpha$ characterizes $W'$, then $(M, w) \models \alpha$ if and only if $w \in W'$.

**Lemma 2.7:** *Let $M$ be a propositional Kripke structure of knowledge and plausibility satisfying CONS and SDP. Given $w$ and $i$, suppose $\alpha$ characterizes the most plausible worlds in $\mathcal{P}(w,i)$. Then $(M,w) \models B_i\varphi$ if and only if $(M,w) \models B_i^\alpha\varphi$.*

Voorbraak [Voo92] distinguishes two notions of knowledge: *objective* and *true justified belief*. He then studies the interaction of both notions of knowledge with beliefs. The intuition we assign to knowledge is similar to his notion of objective knowledge. However, Voorbraak objects to the axiom $K_i\varphi \Rightarrow B_i\varphi$, and suggests $B_i\varphi \Rightarrow B_iK_i\varphi$. The difference lies in the interpretation of belief. Voorbraak's notion of belief is stronger than ours. His view is that the agent cannot distinguish what he believes from what he knows (indeed, he believes that what he believes is the same as what he knows). Our notion of belief is weaker, in that we allow agents to be aware of the defeasibility of their beliefs.

### 2.5 Axiomatizing the full language

Up to now, we have considered just the restricted language $\mathcal{L}^{KB}$. In Appendix A, we present sound and complete axiomatization for the full language $\mathcal{L}^{KP}$ for each of the classes of structures described above.

The technical details are much in the spirit of the axiomatizations presented in [FH88] for knowledge and probability. Our complete axiomatization for $\mathcal{M}$ consists of two "modules": a complete axiomatization for knowledge (i.e., S5) and a complete axiomatization for conditionals (for example, the one given by Burgess in [Bur81]). There are no axioms connecting knowledge and plausibility in this case. In the other cases, for each of the conditions we consider, we provide an axiom that characterizes it. The axioms characterizing NORM, REF, RANK and UNIF are taken from [Lew73] and [Bur81], while the axioms for CONS and SDP (and also UNIF) correspond directly to the axioms suggested in [FH88] for their probabilistic counterparts. We also provide complete characterizations of the complexity of the validity problem for all the logics considered, based on complexity results for knowledge [HM92] and for conditionals [FH93b].

## 3 Knowledge and plausibility in multi-agent systems

Having a good model of knowledge and belief is not enough in order to study how beliefs change. Indeed, if we are mainly interested in agents' beliefs, the additional structure of plausibility spaces does not play a significant role in a static setting. However, if we introduce an explicit notion of time, we expect the plausibility ordering to (partially) determine how the agent changes his beliefs. As we shall see, this gives a reasonable notion of belief change.

A straightforward approach to adding time is by introducing another relation among worlds that signifies which worlds temporally follow any given world (see for example [KL88]). We prefer to introduce more structure into the description by adopting the framework of [HF89] for modeling multi-agent systems. This structure gives a natural definition of knowledge and an intuitive way to describe agents' interactions with their environment.

The key assumption in this framework is that we can characterize the system by describing it in terms of a *state* that changes over time. This is a powerful and natural way to model systems. Formally, we assume that at each point in time, each agent is in some *local state*. Intuitively, this local state encodes the information the agent has observed thus far. In addition, there is also an *environment*, whose state encodes relevant aspects of the system that are not part of the agents' local states.

A *global state* is a tuple $(s_e, s_1, \ldots, s_n)$ consisting of the environment state $s_e$ and the local state $s_i$ of each agent $i$. A *run* of the system is a function from time (which, for ease of exposition, we assume ranges over the natural numbers) to global states. Thus, if $r$ is a run, then $r(0), r(1), \ldots$ is a sequence of global states that, roughly speaking, is a complete description of what happens over time in one possible execution of the system. We take a *system* to consist of a set of runs. Intuitively, these runs describe all the possible sequences of events that could occur in a system.

Given a system $\mathcal{R}$, we refer to a pair $(r, m)$ consisting of a run $r \in \mathcal{R}$ and a time $m$ as a *point*. If $r(m) = (s_e, s_1, \ldots, s_n)$, we define $r_i(m) = s_i$, for each component; thus, $r_i(m)$ is agent $i$'s local state at the point $(r, m)$. We say two points $(r, m)$ and $(r', m')$ are *indistinguishable* to agent $i$, and write $(r, m) \sim_i (r', m')$, if $r_i(m) = r'_i(m')$, i.e., if the agent has the same local state at both points. Finally, we define an *interpreted plausibility system* to be a tuple $(\mathcal{R}, \pi, \mathcal{P})$ consisting of a system $\mathcal{R}$ together with a mapping $\pi$ that associates a truth assignment with the primitive propositions at each point and a function $\mathcal{P}$ such that $\mathcal{P}(r, m, i) = (\Omega_{(r,m,i)}, \preceq_{(r,m,i)})$ is a plausibility space for each point $(r, m)$ and agent $i$.

An interpreted plausibility system can be viewed as a Kripke structure for knowledge and plausibility: the points are the possible worlds, and $\sim_i$ plays the role of the $\mathcal{K}_i$ relations. $\mathcal{P}$ is a function that maps for each agent a point to a plausibility structure over points. We give semantics to sentences in interpreted systems just as in Kripke structures.

We add to the language temporal modalities in the standard fashion (see [HF89]). These include $\bigcirc\varphi$ for "Next time step $\varphi$ will be true" and $\varphi U \psi$ that stands for "$\varphi$ is true until the first time that $\psi$ is true". We call this language $\mathcal{L}^{KPT}$. Evaluation of temporal modalities at a point $(r, m)$ is done by examining the future points on the run $r$. This framework is clearly a temporal extension of the logic of knowledge and plausibility described in the previous section.

## 4 Prior plausibilities

The formal framework of knowledge and plausibility described in the previous section raises a serious problem: While it is easy to see where the $\sim_i$ relations that define knowledge come from, the same cannot be said for the plausibility spaces $\mathcal{P}(r, m, i)$. We now present one possible answer to this question, inspired by the view of plausibility as qualitative probability.

Bayesians assume that agents start with priors on events. If we were thinking probabilistically, we could imagine the agents in a multi-agent system starting with priors on the runs in the system. Since a run describes a complete history over time, this means that the agents are putting a prior probability on the sequences of events that could happen. We would then expect the agent to modify his prior by conditioning on whatever information he has learned. This is essentially the approach taken in [HT89] to defining how the agents' probability distribution changes in a multi-agent system.

Here we assume that the agents start with a prior plausibility ordering on *runs*. In this discussion, we assume for simplicity that we are dealing with *synchronous* systems. Intuitively, this means that the agents know what the time is. Following [HV89], we model this by assuming $(r, m) \sim_i (r', m')$ only if $m = m'$. Notice that by restricting to synchronous systems, if we further assume that the plausibility ordering $\mathcal{P}(r,m,i)$ satisfies CONS, we never have to compare the plausibilities of two different points on the same run.

Suppose that agent $i$'s prior plausibility ordering on the runs in the system is $\mathcal{P}_i = (\mathcal{R}, \preceq_i)$. We can then define $\mathcal{P}(r,m,i) = (\mathcal{K}_i(r,m), \preceq_{(r,m,i)})$, where $(r',m) \preceq_{(r,m,i)} (r'',m)$ if and only if $r' \preceq_i r''$ (for $(r',m),(r'',m) \in \mathcal{K}_i(r,m)$). Notice that $\mathcal{P}(r,m,i)$ is the result of projecting $\mathcal{P}_i$ onto the points in $\mathcal{K}_i(r,m)$ in the obvious way. Thus, in a precise sense, $\mathcal{P}(r,m,i)$ is obtained from $\mathcal{P}_i$ by conditioning on agent $i$'s knowledge at $(r,m)$. Clearly $\mathcal{P}(r,m,i)$ satisfies CONS, REF, and SDP. If, rather than considering one prior on all runs, we partition the runs and have a separate prior on each cell, then $\mathcal{P}(r,m,i)$ would satisfy UNIF instead of SDP.

Formally, we say that $\mathcal{I} = (\mathcal{R}, \pi, \mathcal{P})$ satisfies PRIOR if $\mathcal{I}$ is synchronous and for each run $r$ and agent $i$ there is a prior plausibility $\mathcal{P}_{r,i}$ on some subset of the runs such that $\mathcal{P}(r,m,i)$ is the result of projecting $\mathcal{P}_{r,i}$ onto $\mathcal{K}_i(r,m)$. Notice that PRIOR implies CONS. Furthermore, if $\mathcal{P}_{r,i}$ is independent of $r$, so that agent $i$'s prior is independent of the run he is in, then the resulting system satisfies SDP. If, instead, the set of runs can be partitioned into disjoint subsets $\mathcal{R}_1, \ldots, \mathcal{R}_k$ such that for $r, r' \in \mathcal{R}_j$, we have $\mathcal{P}_{r,i} = \mathcal{P}_{r',i}$, then the resulting system satisfies UNIF.

The conditioning process is especially natural in systems satisfying *perfect recall* [HV89]. A system satisfies prefect recall if the agents' local states encode all the information known in previous states. Systems satisfying both perfect recall and synchronicity satisfy a simple technical condition: the set of runs considered possible at time $m+1$ is a subset of the runs considered possible at time $m$. In this setting, the conditionalization process is a local condition: the plausibility space at $(r,m)$ determines the plausibility at $(r,m+1)$. Moreover, the ordering makes the agent's beliefs as persistent as possible: If $(r',m)$ is minimal in $\mathcal{P}(r,m,i)$, then $(r',m')$ will continue to be minimal in $\mathcal{P}(r,m',i)$ for $m' > m$ until agent $i$ considers $r'$ impossible (i.e., $(r',m') \notin \mathcal{K}_i(r,m')$).[8]

In synchronous systems satisfying PRIOR and perfect recall, the agent can use the plausibility ordering to consider how his belief might change: If his knowledge at the next step is characterized by $\varphi$, then $\psi$ will be believed exactly if $K_i(\bigcirc\varphi \rightarrow_i \bigcirc\psi)$ holds right now. Note that if we assume SDP and that propositions do not change their values along the run, then this reduces to the conditional $\varphi \rightarrow_i \psi$. Interestingly, in several recent papers [Bou92, LS93], this conditional is given semantics similar to ours but described as "given evidence $\varphi$, $\psi$ is believed".

In a precise sense, in systems satisfying CONS, REF, and SDP, we can assume plausibilities are generated via priors in this way without loss of generality: We can effectively transform a system satisfying CONS, REF, and SDP to another system where the plausibilities are generated by a prior and the same formulas are satisfied.

**Theorem 4.1:** *Let $M$ be a synchronous system satisfying CONS, REF and SDP (resp. UNIF). Then there is a system $M'$ satisfying PRIOR, REF and SDP (resp. UNIF) and a mapping $f : \mathcal{R} \mapsto \mathcal{R}'$ such that $(M,r,m) \models \varphi$ if and only if $(M', f(r), m) \models \varphi$.*

By using prior plausibility orderings, we have reduced the question of where the plausibility ordering at each point comes from to the simpler question of where the prior comes from. While this question is far from trivial, it is analogous to a question that needs to be addressed by anyone using a Bayesian approach. Just as with probability theory, in many applications, there is a natural prior (or class of priors) we can use. Interestingly, as shown in [FH93a], we can capture belief update [KM91] by considering systems in which the agent has a prior that satisfies certain restrictions. Thus, we can understand belief update by examining these restrictions.

[8] A similar property of persistence also arises in belief revision; see [FH93a] for further discussion.

By conditioning on plausibility rather than probability, we can deal with a standard problem in the Bayesian approach, that of conditioning on an event of measure 0: Notice that whenever a prior assigns an event a probability measure of 0 it is not possible to condition on that event. The standard solution in the Bayesian school is to give every event of interest, no matter how unlikely, a small positive probability.[9] We may well discover that a formula $\varphi$ that we believed to be true, i.e., one that was true in all the most plausible worlds, is in fact false. Under the probabilistic interpretation of plausibility, this means that we are essentially conditioning on an event $(\neg\varphi)$ of measure 0. The plausibility approach has no problem with this, since the plausibility ordering is still well defined on the worlds that are less plausible, and the conditioning process described above still makes perfect sense.

We conclude this section by considering several examples that show how the assumption of prior plausibilities can be used to naturally model a number of situations.

**Example 4.2:** We extend Example 2.2 to incorporate time, allowing the agent to perform a sequence of tests. We assume that the faults are persistent and do not change during the tests. We model this system as follows: The agent's state is the sequence of input-output relations observed. The environment's state describes the faulty components of the circuit. Each run describes the results of a specific series of tests the agent performs.

The plausibility ordering described in Example 2.2 is now captured as a prior over all runs: We define $r \preceq_1 r'$ if, for every time $m$, $|r_e(m)| \leq |r'_e(m)|$, that is, the number of faulty components in $(r, m)$ is not larger than the number of those in $(r', m')$. At a point $(r, m)$, the agent considers possible all the points where he performed the same tests up to time $m$ and observed the same results. As before, the agent believes that the faulty components are one of the minimal explanations of his observations. As the agent performs more tests, his knowledge increases and his beliefs might change. This process is very similar to the use of Bayesian updating in diagnosis [Kle90]. ■

**Example 4.3:** Our framework lets us easily capture the process of learning a conditional. Consider Alice of Example 2.3. Now, suppose Alice learns that Bob is a Leftfoot. This does not cause Alice to modify any of her orderings. Rather, she conditions on her new knowledge, so that runs where Bob's statements are typically true are considered impossible. As a result, Alice gains knowledge about the plausibility ordering. We claim that this simple example fits into neither the framework of belief revision nor belief update (both of which essentially assume that Alice's beliefs at any time are completely characterized by a single plausibility ordering). Nevertheless, it is certainly a phenomenon we would like to capture, and one that can be easily captured in our framework.[10] ■

**Example 4.4:** *Prisoner's dilemma* is a well-known game where each player can either defect or cooperate. Each player maximizes his payoff by defecting no matter what the other player does, but the utility when both cooperate is higher than when both defect. More precisely, taking $u_j$ to denote player $j$'s utility assignment, then $u_1(c,d) < u_1(d,d) < u_1(c,c) < u_1(d,c)$ (where $(c,d)$ denotes the situation where player 1 cooperates and player 2 defects, and so on); $u_2$ is symmetric, i.e., $u_2(x,y) = u_1(y,x)$, so $u_2(d,c) < u_2(d,d) < u_2(c,c) < u_2(c,d)$. If we assume that a *rational* player is one that chooses a dominating strategy if there is one, then it is clear that rational players playing a one-shot prisoner's dilemma both defect. There is a standard backwards induction argument which seems to show that in any finitely-repeated prisoner's dilemma, if rationality is initially *common knowledge* (so that the players

[9] Of course, this only works if there are only countably many disjoint events of interest.

[10] Such a change was also studied by Boutilier and Goldszmidt in [BG93]. However, since they do not have knowledge in their framework, their models are not expressive enough to represent the agent's beliefs. Rather, they have to use a set of models (i.e., orderings).

both know that the players both know ... that each of them is rational), then the players must play the unique Nash equilibrium solution, which is to always defect. This seems somewhat paradoxical. How do we account for the fact that rational players do much worse in repeated prisoner's dilemma than supposedly irrational players who cooperate? There has been a great deal of effort in the game-theoretic literature to construct models of prisoner's dilemma where rational players can cooperate (see [KMRW82] for one of the best-known examples).

More recently, there has been intense scrutiny of the assumption of common knowledge of rationality. Indeed, it has argued variously (a) that common knowledge of rationality is an inconsistent assumption [Bic89, Ren92], (b) that it is consistent and it indeed implies the backwards induction solution [Aum93], and (c) that, while consistent, it does not necessarily imply the backwards induction solution [Ben92, Sta92]. While a comparison of these arguments is beyond the scope of this paper, we note that the subtleties typically arise when knowledge is treated as "believed to hold with probability 1" (as is often the case in the game theory literature). Under this definition, it is not clear how a player should update her beliefs if the other player cooperates on the first move. The problem is that if we take the standard Bayesian approach of conditioning, then we are conditioning on a measure 0 event.

As we mentioned above, if we use our definition of belief, we avoid the problem of conditioning on measure 0 events. As we now show, under our definitions, while initial common *belief* of rationality is consistent with the backwards induction argument, it does not imply it. On the other hand, if we assume perfect recall (so that the players do not forget what they have seen), then common *knowledge* of rationality does imply the backwards induction argument (using our definition of knowledge as truth in all accessible worlds). Thus, the distinction between knowledge and belief in our framework plays a crucial role here.

We construct a system that distinguishes the knowledge and belief of the players during the game. This construction is similar in spirit to Ben-Porath's and Stalnaker's models [Ben92, Sta92]. We assume two players are playing $n$ iterations of prisoners dilemma. We model this game as a a two agent system, where each player is an agent. Recall that a strategy for player $i$ is a function that returns a move given a sequence of moves (intuitively the ones that have been made up to that time). Since the moves in our case are cooperate and defect, we can view a strategy as a function from a string of $c$'s and $d$'s to $\{c, d\}$. (We allow the empty string as an argument; this represents the initial move.) Each player has a *type* that determines the strategy played by the player and a ranking over the opponent's types. This ranking describes the player's prior beliefs about his opponent. Formally, we consider a set $T$ of types. With each $t \in T$ we associate a strategy $s_t$ and a ranking $\preceq_t$ over $T$. We assume that a player's type does not change throughout a run; let $\mathit{type}(r, i)$ denote player $i$'s type in run $r$. An player's local state at time $m$ contains the player's type and the moves taken up to time $m$. This means that players know their own type and have perfect recall of the game. The set $\mathcal{R}$ of runs consists of all the runs where both players observe the same history and act according to their strategies. We define $\mathcal{P}(r, m, i)$ according to player $i$'s type: $(r', m) \preceq_{(r,m,i)} (r'', m)$ if $\mathit{type}(r', j) \preceq_{\mathit{type}(r,i)} \mathit{type}(r'', j)$, where $j$ is $i$'s opponent. It is easy to verify that this system satisfies PRIOR, REF and SDP.

Let $\mathcal{S}$ be the set of possible strategies in the game, and for each $s \in \mathcal{S}$ let the proposition $p_{i,s}$ denote that player $i$ plays strategy $s$. Using these propositions we can analyze the player's beliefs about the opponent's strategy. Let $\mathcal{S}_i(r, m)$ denote the set of strategies player $i$ believes her opponent may be using. Formally, $\mathcal{S}_i(r, m) = \{s \in \mathcal{S} : (r, m) \models \neg B_i \neg p_{j,s}\}$, where $j$ is $i$'s opponent. These beliefs change from step to step: If the opponent's move at step $m$ is consistent with $i$'s beliefs (i.e., if there are strategies in $\mathcal{S}_i(r, m)$ consistent with the observed move), then $\mathcal{S}_i(r, m + 1)$ is just the subset of $\mathcal{S}_i(r, m)$ consisting of all the strategies consistent with the observed move. If the opponent's move is inconsistent with $i$'s beliefs, then $\mathcal{S}_i(r, m + 1)$ is determined by $i$'s prior ranking, but there is no necessary connection between

$\mathcal{S}_i(r,m)$ and $\mathcal{S}_i(r,m+1)$. Intuitively, if the opponent's move is inconsistent with $i$'s beliefs, then $i$ is surprised; this case essentially corresponds to the occurrence of an event of measure 0.

Now we can define rationality. Of course, we expect that a rational player would not use a strategy that is dominated by another one, assuming that the other player is using a strategy in $\mathcal{S}(r,0)$. However, this condition is not sufficient, since the player's beliefs may change during the game and a strategy that was not dominated initially might be dominated at time $m$ with respect to the player's beliefs at $m$. We expect a rational player to always use a rational strategy. This motivates the following definition: If $x$ is a string of moves and $s$ is a strategy, then $s^x$ is the strategy such that $s^x(y) = s(xy)$ (where $xy$ is the concatenation of the moves in $x$ and $y$). Let $x_j$ be the sequence of moves that have been performed by player $j$ up to the point $(r,m)$ and let $t$ be $\mathit{type}(r,i)$. We say that player $i$ is *locally rational* at the point $(r,m)$ if $s_t^{x_j}$ is not dominated by any strategy when played against strategies $s^{x_i}$ for $s \in \mathcal{S}_i(r,m)$; i.e., there is no strategy that has at least as good a payoff as $s_t^{x_j}$ when played against $s^{x_i}$ for each $s \in \mathcal{S}_i(r,m)$, and does better that $s_t^{x_j}$ when played against at least one strategy $s^{x_i}$ for $s \in \mathcal{S}_i(r,m)$. We say that player $i$ is *rational* at $(r,m)$ if player $i$ is locally rational at $(r',m')$ for all $m' \geq m$ and for all $r'$ such that $r(m) \sim_i r'(m)$.[11]

We now show that there are runs, such that both players cooperate until the last $k+2$ rounds and a common belief that both players are rational holds throughout the run, where $k$ is the minimal natural number such that $k(u_1(c,c) - u_1(d,d)) > u_1(d,c) + u_1(c,d) - 2u_1(d,d)$. In particular, for appropriate choices of utilities, there are runs such that both players cooperate in all but the last two rounds and there is a common belief that both players are rational throughout the run.

We start by defining three families of strategies $s_l^h$, where $h \in \{1,2,3\}$ and $l \leq n$. In these definitions, $x$ and $y$ represent (possible empty) strings of moves; $c^k$ is the string composed of $k$ $c$'s, and $|x|$ is the length of the string $x$.

$$
\begin{aligned}
s_l^1(x) &= \begin{cases} c & \text{if } x = c^j \text{ and either } 0 \leq j < \min(l, n-k-2) \text{ or } l < j < n-1, \text{ or} \\ & \text{if } x = c^j d c^{j'},\ 0 \leq j < \min(l, n-k-2), \text{ and } |x| < n-2 \\ d & \text{otherwise} \end{cases} \\
s_l^2(x) &= \begin{cases} c & \text{if } x = c^j \text{ and } 0 \leq j < \min(l, n-k-2) \text{ or} \\ & \text{if } x = c^j d c^{j'},\ 0 \leq j < \min(l, n-k-2), \text{ and } |x| < n-2 \\ d & \text{otherwise} \end{cases} \\
s_l^3(x) &= \begin{cases} c & \text{if } x = y c^j \text{ and } |y| \leq l \\ d & \text{otherwise} \end{cases}
\end{aligned}
$$

Let $t_l^h$ be types whose associated strategy is $s_l^h$; in the case that $h$ is 1 or 2, we define the ordering $\preceq_{t_l^h}$ below in such a way as to make a player of type $t_l^h$ rational. We do not bother to define the ordering for $t_l^3$ since it is not rational. The beginning of the ordering for $t_l^1$ is

$$t_0^1 \prec_{t_l^1} \cdots \prec_{t_l^1} t_{l-1}^1 \prec_{t_l^1} t_l^2 \prec_{t_l^1} t_{l+1}^3 \prec_{t_l^1} \cdots$$

The beginning of the ordering for $t_l^2$ is the same, except that it does not include the type $t_{l+1}^3$:

$$t_0^1 \prec_{t_l^2} \cdots \prec_{t_l^1} t_{l-1}^2 \prec_{t_l^1} t_l^2 \prec_{t_l^2} \cdots$$

[11] This definition is similar to Stalnaker's definition [Sta92], but has the advantage that it can be represented in our language without introducing counterfactuals. Indeed, it is straightforward (although tedious) to write a formula $\mathit{rational}_i$ in $\mathcal{L}^K PT$ such that $(\mathcal{I}, r, m) \models \mathit{rational}_i$ if player $i$ is rational at the point $(r,m)$ in the interpreted plausibility system $\mathcal{I}$.

Thus, if player 1 is of type $t_l^1$ then he starts out believing that player 2 is of type $t_0^1$. If player 2's first move is $c$, then player 1 is surprised. He revises his beiiefs so that he now believes that player 2 is of type $t_1^1$. If he sees another $c$, he is again surprised, and now believes her type is $t_2^1$, and so on. The initial part of the ordering explains how revision works if a sequence of moves of the form $c^j d^k$ is seen. It does not explain what happens if a different sequence is seen. The rest of the ordering (which we describe in the full paper) ensures that if player 1 observed a sequence that does not correspond to any type in the beginning of the ordering (as defined above), then player 1 believes that player 2 will play $d$ for the rest of the run. In the full paper, we show that the types $t_l^h$ are indeed rational for $h = 1, 2$.

It is easy to see that if $r$ is a run where both players are of type $t_{n-k-2}^1$, then they will cooperate for the first $n-k-2$ rounds, and defect for the last $k+2$ rounds. Moreover, we can show that the players share a common belief that they are rational throughout the run. After $m$ rounds where $m < n-k-2$, both believe that the other is of type $t_m^1$. Since a player of type $t_m^1$ believes that the other player is of type $t_m^2$, and a player of type $t_m^2$ believes that the other is also of type $t_m^2$, at this point common belief of rationality holds. After $m$ rounds where $n-k-2 \leq m \leq n$ both players believe that the other is of type $t_{n-k-2}^2$. Since a player of type $t_{n-k-2}^2$ believes that the other is also of type $t_{n-k-2}^2$, we again have common belief of rationality. Thus, in all rounds a common belief of rationality holds.

We note that the only incentive for player 1 to cooperate in this example is the belief that it will surprise player 2, leading her to believe that he irrational. More precisely, at time $m < l$ player 1 of type $t_l^1$ believes that player 2 is of type $t_m^1$ and that after he will cooperate she will believe him to be of type $t_{m+1}^3$. Thus, cooperation is maintained because both players surprise each other at every round. ∎

## 5 Conclusion

What have we gained by using this formal framework? And how natural is it?

The framework, as presented, give us an expressive modeling tool, enabling us to capture many interesting situations in a natural way. The reader might wonder whether the framework is too expressive for the purposes of belief change. For example, do we really need a different plausibility space for each agent at each point? As we show in [FH93a], in order to capture the notion of belief revision [AGM85] in the most natural way, we do, precisely because the AGM theory puts so few constraints on how beliefs can be revised. On the other hand, it is clear that in many applications there are reasonable constraints on the plausibility spaces. In this paper, we focused on one possible constraint that captures a natural class of belief change situations: prior plausibilities. We are currently exploring other constraints that lead to different notions of belief change. We believe that one of the advantages of our approach is that it gives us the tools both to model and to analyze reasonable constraints. As the examples given above show, this in turn gives us the ability to capture scenarios of belief change in a natural way.

### Acknowledgements

The authors are grateful to Craig Boutilier, Ronen Brafman, Steve Cousins, Ron Fagin, Moises Goldszmidt and particularly Daphne Koller and Moshe Vardi, for comments on drafts of this paper and useful discussions relating to this work.

## A Axiomatizing knowledge and plausibility

We now describe a sound and complete axiomatization for the logic of knowledge and conditionals. The completeness proofs combine techniques used in epistemic logic [HM92] and conditional logics [Bur81, FH93b] are much in the spirit of those in [FH88], so we omit details here.

As in the case of probability [FH88], the axiom system can be modularized into components: propositional reasoning, reasoning about knowledge and reasoning about conditionals. The component for propositional reasoning consists of K1 and RK1 (from Section 2.4); the component for reasoning about knowledge consists of K2–K5 and RK2. The component for reasoning about conditionals consists of the following axioms and rule of inference, taken from [Bur81]:

**C1.** $\varphi \rightarrow_i \varphi$

**C2.** $((\varphi \rightarrow_i \psi_1) \wedge (\varphi \rightarrow_i \psi_2)) \Rightarrow (\varphi \rightarrow_i (\psi_1 \wedge \psi_2))$

**C3.** $(\varphi \rightarrow_i (\psi_1 \wedge \psi_2)) \Rightarrow (\varphi \rightarrow_i \psi_1)$

**C4.** $((\varphi_1 \rightarrow_i \psi) \wedge (\varphi_1 \rightarrow_i \varphi_2)) \Rightarrow (\varphi_1 \wedge \varphi_2 \rightarrow_i \psi)$

**C5.** $((\varphi_1 \rightarrow_i \psi) \wedge (\varphi_2 \rightarrow_i \psi)) \Rightarrow (\varphi_1 \vee \varphi_2 \rightarrow_i \psi)$

**RC1.** From $\varphi_1 \equiv \varphi_2$ and $\psi$ infer $\psi'$ the result of replacing $\varphi_1$ by $\varphi_2$ in $\psi$ (replacement of equivalent subformulas)

Let AX consist of K1–K5, C1–C5, RK1,RK2, and RC1.

**Theorem A.1:** *AX is a sound and complete axiomatization for $\mathcal{L}^{KP}$ with respect to $\mathcal{M}$.*

We now capture the conditions described above axiomatically. CONS, NORM, REF, SDP, UNIF and RANK correspond to the following axioms, respectively:

**A1.** $K_i\varphi \Rightarrow (\neg\varphi \rightarrow_i \mathit{false})$

**A2.** $\neg(\mathit{true} \rightarrow_i \mathit{false})$

**A3.** $(\varphi \rightarrow_i \mathit{false}) \Rightarrow \neg\varphi$

**A4.** $(\varphi \rightarrow_i \psi) \Rightarrow K_i(\varphi \rightarrow_i \psi)$

**A5.** $[(\varphi \rightarrow_i \psi) \Rightarrow (\neg(\varphi \rightarrow_i \psi) \rightarrow_i \mathit{false})] \wedge [\neg(\varphi \rightarrow_i \psi) \Rightarrow ((\varphi \rightarrow_i \psi) \rightarrow_i \mathit{false})]$

**A6.** $((\varphi_1 \vee \varphi_2) \rightarrow_i \neg\varphi_2) \Rightarrow ((\varphi_2 \vee \psi) \rightarrow_i \neg\varphi_2) \vee ((\varphi_1 \vee \psi) \rightarrow_i \neg\psi)$

We show that adding the appropriate axioms to AX gives a sound and complete axiomatization of the logic with respect to the class of structures satisfying the corresponding conditions.

**Theorem A.2:** *Let $C$ be a subset of { CONS, NORM, REF, SDP, UNIF, RANK } and let $A$ be the corresponding subset of { A1, A2, A3, A4, A5, A6 }. Then AX$\cup$A is a sound and complete axiomatization with respect to the structures in $\mathcal{M}$ satisfying $C$.*

We now consider the complexity of the validity problem. Our results are based on a combination of results for complexity of epistemic logics [HM92] and conditional logics [FH93b]. Again, the technical details are much in the spirit of those in [FH88]. We presume that the reader is familiar with standard complexity-theoretic notions such as NP, co-NP, polynomial space, and exponential time (see [HU79] for details).

**Theorem A.3:** *Let $C$ be a subset of { CONS, NORM, REF, SDP, UNIF, RANK }. If CONS $\in C$, but it is not the case that UNIF or SDP is in $C$, then the validity problem with respect to structures satisfying $C$ is complete for exponential time. Otherwise, the validity problem is complete for polynomial space.*

If we restrict attention to the case of one agent and structures satisfying CONS and either UNIF or SDP, then we can do better.

**Theorem A.4:** *Let $C$ be a subset of { CONS, NORM, REF, SDP, UNIF, RANK } containing CONS and either UNIF or SDP. For the case of one agent, the validity problem is co-NP-complete.*

## B Ranked plausibility spaces and nonstandard approaches to probability

We want to relate the conditional $S \rightarrow T$ to the conditional probability $\Pr(T|S) = 1$. This is impossible to do in standard probability theory since it does not handle conditioning on measure 0 events. In this appendix, we show how ranked plausibility spaces can be related to a number of probabilistic approaches to dealing with the problem of conditioning on events of measure 0, including nonstandard probability functions that can assign infinitesimal values [LM92], Popper functions [Fra76], and the lexicographic probability approach of Blume, Brandenburger, and Dekel [BBD91].

*Popper functions* take the notion of a conditional probability as primitive. Formally, a Popper function takes two arguments and returns a value in $[0,1]$, in a way that satisfies a number of axioms described below. Intuitively, if $f$ is a Popper function, then we think of $f(A,B)$ as the conditional probability $\Pr(A|B)$. To emphasize this intuition, we actually use the latter notation when talking about Popper functions. As we would expect, if we fix the context $B$ then $\Pr_B(A) = \Pr(A|B)$ satisfies the usual properties of absolute probabilities (i.e., Kolmogorov axioms). Formally, a Popper function satisfies the following axioms [Fra76]:

**P1.** $0 \leq \Pr(A|B) \leq \Pr(B|B) = 1$

**P2.** If $\Pr(\bar{B}|B) \neq 1$ then $\Pr(-|B)$ is a probability function

**P3.** $\Pr(A \cap B|C) = \Pr(A|C)\Pr(B|C \cap A)$

These conditions match our intuitions about conditional probabilities, except that it is possible to condition on an event which as a prior probability of 0. There are some events that cannot be conditioned on (such as the empty set). These are called *abnormal* events and are exactly those where $\Pr(\bar{B}|B) = 1$. A *normal* Popper function is one for which the empty set is the only abnormal event.

It is not hard to prove the following connection between ranked plausibility spaces and Popper functions:

**Theorem B.1:** *Let $P = (\Omega, \preceq)$ be a ranked plausibility space, such that $\Omega$ is countable. Then, there is a normal Popper function $\Pr_P$ over $2^\Omega$ such that $\Pr_P(T|S) = 1$ if and only if $S \rightarrow T$ holds in $P$. Furthermore, for each normal Popper function $\Pr$ over $2^\Omega$ there is a plausibility space $P$ such that $\Pr_P = \Pr$.*

Another possible way to handle conditioning by 0 measure events is to allow infinitesimal probabilities. An infinitesimal is larger than 0 but smaller than all positive reals. The idea is to consider an extension $\mathcal{R}^*$ of the reals that satisfies all the properties of the reals. Then it is possible to define non-standard probability as a mapping into $[0,1]^*$, the extended interval. Probability still satisfies the usual Kolmogorov axioms. Lehmann and Magidor [LM92] show the following correspondence:

**Theorem B.2:** *[LM92] For every non-standard probability function $\Pr^*$, there is a ranked plausibility space $P_{\Pr^*}$ such that $S \rightarrow T$ holds in $P_{\Pr^*}$ if and only if $1 - \Pr^*(T|S)$ is infinitesimal. Furthermore, for every ranked plausibility space $P$ with countable domain there is a non-standard probability probability function $\Pr^*$ such that $P = P_{\Pr^*}$.*

Although Popper functions and non-standard probability functions are essentially equivalent in the context of rankings, they differ once we consider more detailed quantitative information. In this case, non-standard probability functions are strictly more general than Popper functions. The intuition is that Popper functions can be mapped to non-standard probabilities by defining $\Pr^*$ such that if $\Pr(A|B) = 0$

then $\Pr^*(A \cap B)/\Pr^*(B)$ is infinitesimal. However, there are non-standard probability functions that cannot be represented by a Popper function, for example a function $\Pr*$ such that, for some primitive event $A$ and some context $B$, $\Pr^*(A|B) = x + \epsilon$ where $x$ is a positive real number and $\epsilon$ is infinitesimal.

Blume, Brandenburger, and Dekel [BBD91] consider some related issues from a decision-theoretic viewpoint. Savage, in his seminal book [Sav54], gave a number of axioms characterizing preference orderings, and showed that any preference ordering could, in a precise sense, be represented by a probability function. In [BBD91], one of Savage's axioms (the so-called Archimedian Axiom) is replaced with a weaker axiom $AX$. It is then shown that the resulting preference order can be represented in terms of a *lexicographic probability system (LPS).* (We omit the details of the definition here.) The key point is that, as shown in [BBD91], the representation could have been done equally well using an extended probability function. In addition, they also consider a stronger version of $AX$ that is still weaker than Savage's Archimedean axiom, and show that the resulting class of preference order can be represented in terms of what they call *lexicographic conditional probability systems (LCPS).* As we show in the full paper, these preference orders can be represented by Popper functions.

## References

[AGM85] C. E. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: partial meet functions for contraction and revision. *Journal of Symbolic Logic*, 50:510–530, 1985.

[Aum93] R. J. Aumann. Backwards induction and common knowledge of rationality. Presented at the Summer workshop of the Stanford Institute for Theoretical Economics, 1993.

[BBD91] L. Blume, A. Brandenburger, and E. Dekel. Lexicographic probabilities and choice under uncertainty. *Econometrica*, 59(1):61–79, 1991.

[Ben92] E. Ben-Porath. Rationality, Nash equilibrium and backward induction in perfect information games. Working paper, The Sackler Institute of Economic Studies, Tel-Aviv University, 1992.

[BG93] C. Boutilier and M. Goldszmidt. Revising by conditional beliefs. In *Proc. National Conference on Artificial Intelligence (AAAI-93)*, pages 648–654, 1993.

[Bic88] C. Bicchieri. Strategic behavior and counterfactuals. *Synthese*, 76:135–169, 1988.

[Bic89] C. Bicchieri. Self refuting theories of strategic interaction: A paradox of common knowledge. *Erkenntnis*, 30:69–85, 1989.

[Bou92] C. Boutilier. Normative, subjective and autoepistemic defaults: Adopting the Ramsey test. In *Principles of Knowledge Representation and Reasoning: Proc. Third International Conference (KR '92)*, 1992.

[Bur81] J. Burgess. Quick completeness proofs for some logics of conditionals. *Notre Dame Journal of Formal Logic*, 22:76–84, 1981.

[Che80] B. F. Chellas. *Modal Logic.* Cambridge University Press, Cambridge, UK, 1980.

[DH88] R. Davis and W. Hamscher. Model-based reasoning: troubleshooting. In H. Shrobe and The American Association for Artificial Intelligence, editors, *Exploring AI*, pages 297–346. Morgan Kaufmann, San Mateo, CA., 1988.

[FH88] R. Fagin and J. Y. Halpern. Reasoning about knowledge and probability: preliminary report. In M. Y. Vardi, editor, *Proc. Second Conference on Theoretical Aspects of Reasoning about Knowledge*, pages 277–293. Morgan Kaufmann, San Mateo, CA, 1988. An expanded version of this paper appears as IBM Research Report RJ 6020, 1990; to appear in *Journal of the ACM.*

[FH93a] N. Friedman and J. Y. Halpern. A knowledge-based framework for belief change. Part II: revision and update. Technical report, 1993. Submitted, KR'94.

[FH93b] N. Friedman and J. Y. Halpern. On the complexity of conditional logics. Technical report, 1993. Submitted, KR'94.

[Fra76] B. C. van Fraasen. Representation of conditional probabilities. *Journal of Philosophical Logic*, 5:417–430, 1976.

[Gär88] P. Gärdenfors. *Knowledge in Flux.* Cambridge University Press, Cambridge, UK, 1988.

[Gef92] H. Geffner. *Default Reasoning.* MIT Press, Cambridge, MA, 1992.

[Gin86] M. L. Ginsberg. Counterfactuals. *Artificial Intelligence*, 30:35–79, 1986.

[GMP93] M. Goldszmidt, P. Morris, and J. Pearl. A maximum entropy approach to nonmonotonic reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15:220–231, 1993.

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

[Hin62] J. Hintikka. *Knowledge and Belief.* Cornell University Press, Ithaca, NY, 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.

[HT89] J. Y. Halpern and M. R. Tuttle. Knowledge, probability, and adversaries. In *Proc. 8th ACM Symp. on Principles of Distributed Computing*, pages 103–118, 1989. To appear in *Journal of the ACM.*

[HU79] J. E. Hopcroft and J. D. Ullman. *Introduction to Automata Theory, Languages and Computation.* Addison-Wesley, New York, 1979.

[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.

[KL88] S. Kraus and D. J. Lehmann. Knowledge, belief, and time. *Theoretical Computer Science*, 58:155–174, 1988.

[Kle90] J. de Kleer. Using crude probability estimates to guide diagnosis. *Artificial Intelligence*, 45:381–392, 1990.

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

[KM91] H. Katsuno and A. Mendelzon. On the difference between updating a knowledge base and revising it. In *Principles of Knowledge Representation and Reasoning: Proc. Second International Conference (KR '91)*, pages 387–394, 1991.

[KMRW82] D. Kreps, P. Milgrom, J. Roberts, and R. Wilson. Rational cooperation in finitely repeated Prisoners' Dilemma. *Journal of Economic Theory*, 27(2):245–252, 1982.

[Lev84] H. J. Levesque. A logic of implicit and explicit belief. In *Proc. National Conference on Artificial Intelligence (AAAI '84)*, pages 198–202, 1984.

[Lew73] D. K. Lewis. *Counterfactuals.* Harvard University Press, Cambridge, MA., 1973.

[LM92] D. Lehmann and M. Magidor. What does a conditional knowledge base entail? *Artificial Intelligence*, 55:1–60, 1992.

[LS93] P. Lamarre and Y. Shoham. Knowledge, certainty, belief, and conditionalizition. 1993.

[Pea89] J. Pearl. Probabilistic semantics for nonmonotonic reasoning: A survey. In R. J. Brachman, H. J. Levesque, and R. Reiter, editors, *Proc. First International Conference on Principles of Knowledge Representation and Reasoning (KR '89)*, pages 505–516, 1989. Reprinted in *Readings in Uncertain Reasoning*, G. Shafer and J. Pearl (eds.), Morgan Kaufmann, San Mateo, CA, 1990, pp. 699–710.

[Rei87] R. Reiter. A theory of diagnosis from first principles. *Artificial Intelligence*, 32:57–95, 1987. Reprinted in in *Readings in Nonmonotonic Reasoning*, M. L. Ginsberg (ed.), Morgan Kaufman, San Mateo, CA. 1987, pp. 352–371.

[Ren92] P. Reny. Rationality in extensive form games. *Journal of Economic Perspectives*, 6:103–118, 1992.

[Sav54] L. J. Savage. *Foundations of Statistics.* John Wiley & Sons, New York, 1954.

[Sho87] Y. Shoham. A semantical approach to nonmonotonic logics. In *Proc. 2nd IEEE Symp. on Logic in Computer Science*, pages 275–279, 1987. Reprinted in in *Readings in Nonmonotonic Reasoning*, M. L. Ginsberg (ed.), Morgan Kaufman, San Mateo, CA. 1987, pp. 227–250.

[SM89] Y. Shoham and Y. Moses. Belief as defeasible knowledge. In *Proc. Eleventh International Joint Conference on Artificial Intelligence (IJCAI '89)*, pages 1168–1173, 1989.

[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.

[Sta92] R. C. Stalnaker. Knowledge, belief and counterfactual reasoning in games. Forthcoming in *Proceedings of the Second Castiglioncello Conference*, edited by C. Bicchieri and B. Skyrms., 1992.

[Voo92] F. Voorbraak. Generalized Kripke models for epistemic logic. In *Theoretical Aspects of Reasoning about Knowledge: Proc. Fourth Conference*, pages 214–228, 1992.