

Moshe Y. Vardi

CSLI, Ventura Hall
Stanford University
Stanford, CA 94305, USA

Abstract

We present a semantic model for knowledge with the following properties: (1) Knowledge is necessarily correct, (2) agents are logically omniscient, i.e., they know all the consequences of their knowledge, and (3) agents are positively introspective, i.e., they are aware of their knowledge, but not negatively introspective, i.e., they may not be aware of their ignorance. We argue that this is the appropriate model for implicit knowledge. We investigate the properties of the model, and use it to formalize the notion of circumscribed knowledge.

1. Introduction

Planning sequences of actions and reasoning about their effects is one of the most thoroughly studied areas in AI. Realistic planning requires, however, not only the ability to reason about the real world, but also the ability of the planning agent to reason about the information it has about the real world. In other words, intelligent agents need to be able to reason about their own *knowledge* as well as other agents' knowledge. (Planning, in fact, requires a theory of knowledge and action. See [Mo80], [Mo84].)

A theory of knowledge requires a robust mathematical model of knowledge. Starting with Hintikka [Hi62], knowledge was traditionally modelled by *Kripke structures* [Kr63] (e.g., [Mo80, Mo84, MSH78, Sa76]). More recently, however, the adequacy of Kripke structures to model *propositional attitudes*, and in particular *epistemic attitudes*, has been questioned on several grounds [FHV84, FV84]. Fagin et al. introduced *modal structures* [FV84], called *knowledge structures* when applied to knowledge [FHV84], which captures Hintikka's *possible worlds semantics* in a set-theoretic way, rather than a graph-theoretic way as in Kripke structures. It turns out that modal structures not only a fortiori justify the use of Kripke structures to model possible worlds, but they also enable a much more refined analysis of several notions [FHV84, FV84]. We shall use modal structure in our model-theoretic analysis of knowledge.

It is well known that there is not really such a thing as *the* concept of knowledge, but rather a whole continuum of "concepts of knowledge" (cf. [Le78, OC68]). Fagin et al. [FHV84] chose to use a notion of knowledge where the agents are *logically omniscient*, i.e., they know all the consequences of their knowledge, and are *fully introspective*, they are both *positively introspective* (aware of their knowledge) and *negatively introspective* (aware of their ignorance). One way to explain the logical omniscience of these agents is to view their knowledge as *implicit*, i.e., the agents implicitly know everything that logically follows from their *explicit* knowledge [HM84a, Le84, RP85]. From this point of view, however, it is hard to justify the assumption that the agents have negative introspection.

The above argument motivates a model-theoretic study of a notion of *positively introspective* knowledge, where the agents are introspective of their knowledge but not of their ignorance. Such a concept of knowledge was advocated, on other grounds, by Hintikka [Hi62] and Moore [Mo84]. A fully introspective knowledge is *non-monotonic*, since ignorance leads to knowledge (by introspection). This lead to several difficulties in trying to circumscribe this kind of knowledge (cf. [HM84b, Pa84, St81]). In contrast, the concept of knowledge studied here is *monotonic*. In terms of modal logic, fully introspective knowledge correspond to the modal logic S5, while positively introspective knowledge correspond to the modal logic S4.

Our treatment here of positively introspective knowledge closely resemble the treatment of fully introspective knowledge in [FHV84]. Nevertheless, there are several significant differences, upon which we elaborate in the paper. The most important point is that

in our framework we can study *comparative knowledge* and *circumscribed knowledge*. Our framework enables us to assign precise meaning to the notion of "to know more", which in turn enables us to circumscribe knowledge, i.e., to assign precise meaning to the notion of "all one knows".

2. Knowledge Structures

We now define *knowledge structures* that capture the essence of Hintikka's [Hi62] *possible worlds* approach towards modelling knowledge.

Before we formally define knowledge structures, let us discuss them informally. The basic idea underlying knowledge structures is that there are different levels of knowledge. Consider the following example, taken from [FHV84]. Assume there are two agents, Alice and Bob, and that there is only one atomic proposition p . At the 0th level ("nature"), assume that p is true. The 1st level tells each player's knowledge about nature. For example, Alice's knowledge at the 1st level could be "I (Alice) don't know whether p is true or false", and Bob's could be "I (Bob) know that p is true". The 2nd level tells each player's knowledge about the other player's knowledge about nature. For example, Alice's knowledge at the 2nd level could be "I know that I don't know about p , but I know that Bob knows whether p is true or false", and Bob's could be "I know that I know p , but I don't know whether Alice knows p ". Thus, Alice knows that either p is true and Bob knows it, or else p is false and Bob knows it. At the 3rd level, Alice's knowledge could be "I know that Bob does not know whether I know about p ". This can continue for arbitrarily many levels.

We assume a fixed finite set P of atomic propositions, and a fixed finite set A of agents. Intuitively, a knowledge structure has various *levels*, where the 0th level is a truth assignment to the atomic propositions, and where the k th level contains a set of "possible" k -ary worlds for each agent. Formally, we define a *0th-order assignment*, f_0 , to be a truth assignment to the atomic propositions. We call $\langle f_0 \rangle$ a *1-ary world* (since its "length" is 1). Assume inductively that k -ary worlds $\langle f_0, \dots, f_{k-1} \rangle$ have been defined. Let W_k be the set of all k -ary worlds. A *kth-order assignment* is a function $f_k: A \rightarrow \text{PowerSet}(W_k)$. Intuitively, f_k associates with each agent a set of "possible" k -ary worlds. There is a "compatibility" restriction on f_k 's, which we shall discuss shortly. We call $\langle f_0, \dots, f_k \rangle$ a $(k+1)$ -ary world. An infinite sequence $\langle f_0, f_1, f_2, \dots \rangle$ is called a *knowledge structure* if each *prefix* $\langle f_0, \dots, f_{k-1} \rangle$ is a k -ary world for each k .

Before we list the restriction that we mentioned, let us reconsider the above example. In that example, f_0 is the truth assignment that makes p true. Also, $f_1(\text{Alice}) = \{p, \bar{p}\}$ (where by p (respectively, $\bar{p}$) we mean the 1-ary world $\langle f_0 \rangle$ where f_0 is the truth assignment that makes p true (respectively, false)), and $f_1(\text{Bob}) = \{p\}$. Saying $f_1(\text{Alice}) = \{p, \bar{p}\}$ means that Alice does not know whether p is true or false. We can write the 2-ary world $\langle f_0, f_1 \rangle$ as $\langle p, (\text{Alice} \rightarrow \{p, \bar{p}\}, \text{Bob} \rightarrow \{p\}) \rangle$. Let us denote this 2-ary world by w_1 . Let w_2 be the 2-ary world $\langle \bar{p}, (\text{Alice} \rightarrow \{p, \bar{p}\}, \text{Bob} \rightarrow \{p\}) \rangle$, let w_3 be $\langle p, (\text{Alice} \rightarrow \{p\}, \text{Bob} \rightarrow \{p\}) \rangle$, let w_4 be $\langle \bar{p}, (\text{Alice} \rightarrow \{p\}, \text{Bob} \rightarrow \{p\}) \rangle$. In the example above $f_2(\text{Alice}) = \{w_1, w_2, w_3, w_4\}$, since Alice thinks both w_1 (where p is true, Bob knows it, but she does not), w_2 (where p is false, Bob knows it, but she does not), w_3 (where p is true, Bob knows it, and she knows it), and w_4 (where p is false, Bob knows it, and she knows it), are possible worlds. Similarly, $f_2(\text{Bob}) = \{w_1, w_3\}$, since Bob thinks both w_1 (where p is true, he knows it, but Alice does not know it) and w_3 (where p is true, he knows it, and Alice knows it) are possible worlds.

The restriction that we mentioned earlier, is:

(K1) $\langle g_0, \dots, g_{k-2} \rangle \in f_{k-1}(a)$ iff there is a $(k-1)$ st-order assignment g_{k-1} such that $\langle g_0, \dots, g_{k-2}, g_{k-1} \rangle \in f_k(a)$, for $a \in A$ and $1 < k$.

This restriction says that the $(k-1)$ -ary worlds that agent a thinks are possible are prefixes of the k -ary worlds that a thinks are possible. That is, a 's higher-order knowledge extends a 's lower-order knowledge.

Restriction K1 is not really particular to knowledge and actually applies to any normal modality [FV84]. Indeed, the structures that we have defined so far are exactly the *modal structures* of Fagin and Vardi [FV84]. Since we are trying to model knowledge, we have to impose certain semantic restrictions on these structures in order to capture the properties of knowledge that we have in mind. The properties of knowledge in which we are interested are those relevant to planning and acting.

The first property that we are trying to capture is that knowledge is always correct, i.e., anything that someone knows is true. While it is possible to have false *beliefs*, it is impossible to have false knowledge. The reason that this distinction is important to planning and acting is simply that, for an agent to achieve her goals, the beliefs on which she bases her actions must generally be true. To capture this property we impose the following restriction on worlds [FHV84]:

(K2) $\langle f_0, \dots, f_{k-1} \rangle \in f_k(a)$, for $a \in A$ and $k \geq 1$.

That is, the real k -ary world is one of the possibilities for each agent.

The second property of knowledge that is important to planning is that, if an agent knows something, then she knows that she knows it. That is, the agents are aware of their knowledge. We call this *positive introspection*. This property is important for reasoning about plans that requires gathering information or plans that consists of several actions [Mo84]. To capture this property we impose the following restriction on worlds:

(K3) If $\langle g_0, \dots, g_{k-1} \rangle \in g_k(a)$, then $g_{k-1}(a) \subseteq f_{k-1}(a)$, for $a \in A$ and $k > 1$.

Let us see why this captures positive introspection. Knowledge is captured in our model by sets of possibilities. The *more* possibilities an agent has, the *less* she knows. Since an agent knows what she knows but does not necessarily know what she does not know, it is conceivable to her that she knows more than she is aware of. Thus if $w = \langle f_0, \dots, f_{k-1} \rangle$ represents the *actual* state of knowledge, then $w' = \langle g_0, \dots, g_{k-1} \rangle$ is a *conceivable* possibility for a with respect to w only if $g_{k-1}(a) \subseteq f_{k-1}(a)$. Thus if $w' = \langle g_0, \dots, g_{k-1} \rangle$ is one of a 's possible worlds, then w' must be conceivable to her with respect to w , i.e., $g_{k-1}(a) \subseteq f_{k-1}(a)$.

The last property of knowledge that we consider is that if an agent does not know something, then she knows that she does not know it. That is, the agents are aware of their ignorance. We call this *negative introspection*. This property is philosophically controversial (see [Le78]), and was rejected by Hintikka [Hi62] and Moore [Mo84]. We also reject it here but on different grounds.

In every model of knowledge in which knowledge is captured by a set of possibilities, the agent turns out to be *logically omniscient* [Hi75]. This is because these models satisfy the following formula, where $K_a \phi$ means that a knows ϕ :

$$K_a p / K_a (p \supset q) \supset K_a q$$

Namely, the agents know all the consequences of their knowledge. This is of course unrealistic, not only for humans but for any physical agent. One way to explain the logical omniscience of these agents is to view their knowledge as *implicit* rather than *explicit*, i.e., an agent implicitly knows all the consequences of her explicit knowledge [HM84a, Le84, Re74, RP85]. The following analysis shows that an agent is implicitly aware of her implicit knowledge: Suppose she explicitly knows p and $p \supset q$, so she implicitly knows q . Then she explicitly knows that she knows p and she explicitly knows that she knows $p \supset q$, and therefore she implicitly knows that she knows q . On the other hand, there is no reason to accept that an agent is implicitly aware of her ignorance. Thus we also reject negative introspection.

We note that another approach to the logical omniscience problem is to consider the agents as idealized [Mo84]. In that case, we believe that it does make sense to accept negative introspection. Indeed, Fagin et al. [FHV84] did accept negative introspection, and they captured full introspection by the following restriction:

(K3') If $\langle g_0, \dots, g_{k-1} \rangle \in g_k(a)$, then $g_{k-1}(a) = f_{k-1}(a)$, for $a \in A$ and $k > 1$.

That is, in all the worlds that an agent thinks possible she has exactly the same knowledge as in the real world.

Notice that the only difference between K3 and K3' is that the inclusion is replaced by an equality. As we shall see later this subtle difference carries substantial implication. We call knowledge structures that satisfy K3' *fully introspective* knowledge structures.

A natural question that one may ask is whether worlds can be viewed as partial specifications of structures. This would be very desirable since worlds are finite while structures are infinite. McCarthy [Mc84] posed essentially this question as an open problem in 1975. It turns out surprisingly, that not every world can be extended to a knowledge structures. The following theorem characterizes the extendible worlds.

Theorem 1. A world $\langle f_0, \dots, f_k \rangle$ is a prefix of some knowledge structure if and only if either $k \leq 1$ or whenever $\langle g_0, \dots, g_{k-1} \rangle \in f_k(a)$ and $\langle h_0, \dots, h_{k-2} \rangle \in g_{k-1}(a)$, then there is a $(k-1)$ st-order assignment h_{k-1} such that $\langle h_0, \dots, h_{k-1} \rangle \in f_k(a)$ and $h_{k-1}(a) \subseteq g_{k-1}(a)$.

We note that, in the context of fully introspective structures, every world is a prefix of some structure.

The intuition behind the condition of the theorem will be explained in the full paper. From now on, we deal only with worlds that satisfy the condition of Theorem 1.

3. Syntax and Semantics

The set of *formulas* is the smallest set that contains the atomic propositions, is closed under Boolean connectives and contains $K_a \phi$ if it contains ϕ . The *depth* of a formula ϕ is the depth of nesting of the K_a 's in ϕ :

- $\text{depth}(p) = 0$, if p is atomic propositions.
- $\text{depth}(\neg \phi) = \text{depth}(\phi)$.
- $\text{depth}(\phi / \psi) = \max\{\text{depth}(\phi), \text{depth}(\psi)\}$.
- $\text{depth}(K_a(\phi)) = 1 + \text{depth}(\phi)$.

We are almost ready to define what it means for a knowledge structure to *satisfy* a formula. We begin by defining what it means for an $(r+1)$ -ary world $\langle f_0, \dots, f_r \rangle$ to satisfy formula ϕ , written $\langle f_0, \dots, f_r \rangle \models \phi$, if $r \geq \text{depth}(\phi)$.

- $\langle f_0, \dots, f_r \rangle \models p$, where p is a primitive proposition, if p is true under the truth assignment f_0 .
- $\langle f_0, \dots, f_r \rangle \models \neg \phi$ if $\langle f_0, \dots, f_r \rangle \not\models \phi$.
- $\langle f_0, \dots, f_r \rangle \models \phi_1 / \phi_2$ if both $\langle f_0, \dots, f_r \rangle \models \phi_1$ and $\langle f_0, \dots, f_r \rangle \models \phi_2$.
- $\langle f_0, \dots, f_r \rangle \models K_a \phi$ if $\langle g_0, \dots, g_{r-1} \rangle \models \phi$ for each $\langle g_0, \dots, g_{r-1} \rangle \in f_r(a)$.

Note that we do not define satisfaction of "deep" formulas in "short" worlds. The actual arity of the world does not quite matter, as long as it is "long" enough.

Lemma 2: Assume that $\text{depth}(\phi) = k$ and $r \geq k$. Then $\langle f_0, \dots, f_r \rangle \models \phi$ iff $\langle f_0, \dots, f_k \rangle \models \phi$.

We say that the knowledge structure $\mathbf{f} = \langle f_0, f_1, \dots \rangle$ satisfies ϕ , written $\mathbf{f} \models \phi$, if $\langle f_0, \dots, f_k \rangle \models \phi$, where $k = \text{depth}(\phi)$. This is a reasonable definition, since if $w = \langle f_0, \dots, f_r \rangle$ is an arbitrary prefix of $\mathbf{f}$ such that $r \geq k$, then it then follows from Lemma 2 that $\mathbf{f} \models \phi$ iff $w \models \phi$.

A formula ϕ is *valid* if it is satisfied by all knowledge structures. From Theorem 1 and Lemma 2 it follows that if $\text{depth}(\phi) = k$, then it suffices to check that ϕ is satisfied in all $(k+1)$ -ary worlds.

Theorem 3: The validity problem for knowledge structures is decidable.

We can axiomatize validity in the following way:

Theorem 4: The following formal system is sound and complete for validity in knowledge structure:

- (A1) All substitution instances of propositional tautologies.
- (A2) $K_a \phi / (K_a(\phi \supset \psi) \supset K_a \psi)$.
- (A3) $K_a \phi \supset K_a K_a \phi$.
- (R1) From ϕ and $\phi \supset \psi$ infer ψ .
- (R2) From ϕ infer $K_a \phi$.

Thus the logic of knowledge structures is the generalization of the modal logic S4 [Ch80] to include multiple modalities. We note that to axiomatize validity in fully introspective knowledge structure one has to add the axiom $\neg K_a \phi \supset K_a \neg K_a \phi$ [FHV84]. That is, the logic of fully introspective knowledge structures is a generalization of the modal logic S5.

4. Comparative Knowledge

In this section we want to compare the knowledge state of the same agent in two different knowledge structures. We want to give precise meaning to informal statements such as "I know today *more* than I knew yesterday". The basic idea, which we have already mentioned in the previous sections, is that, since knowledge is captured in our model by sets of possibilities, the *more* possibilities an agent has, the *less* she knows. Thus if $f = \langle f_0, f_1, \dots \rangle$ and $g = \langle g_0, g_1, \dots \rangle$, then a knows in g at least as much as she knows in f , denoted $f \leq_a g$, if $g_k(a) \subseteq f_k(a)$, for all $k \geq 1$, that is if a has in f at least all the possibilities she has in g and possibly more. If we have both $f \leq_a g$ and $g \leq_a f$, then we say that a knows the same in f and g , denoted $f \equiv_a g$. Note that $f \equiv_a g$ if $f_k(a) = g_k(a)$ for all $k \geq 1$.

Lemma 5: $\leq_a$ is a reflexive and transitive relation. $\equiv_a$ is an equivalence relation.

The interpretation that we want to give to "knowing at least as much" is that if a knows in g at least as much as she knows in f , then g is a *possible* state for her in f . The following theorem justifies this interpretation.

Theorem 6:

- (1) $f \leq_a g$ if and only if $\langle f_0, \dots, g_{k-1} \rangle \in f_k(a)$ for all $k \geq 1$.
- (2) $f \models K_a \phi$ if and only if $g \models \phi$ whenever $f \leq_a g$.

Thus, a knows ϕ precisely when ϕ holds in all knowledge structures that a thinks are possible. Theorem 6 shows the equivalence of two notions of truth. The first notion of truth is "internal", this truth is determined by looking inside the structure. The second notion of truth is "external", this truth is determined by considering an infinitude of possible structures. The reader should compare Theorem 6 with Theorem 2.6 in [FHV84].

Now that we have justified our notion of "knowing at least as much", we show that this notion makes sense only for partially introspective agents:

Theorem 7: Let f and g be fully introspective knowledge structure. Then $f \leq_a g$ if and only if $g \equiv_a f$.

Thus for fully introspective agents "knowing at least as much" is equivalent to "knowing the same". The intuition behind it is that when an agent does not know something, she knows that she does not know. Therefore, her ignorance gains her knowledge. Consequently, no two distinct states of knowledge are comparable (cf. [HM84]).

Previous attempts (cf. [Mo80, Mo84, MSHI78, Sa76]) to provide a semantic foundation for reasoning about knowledge have made use of *Kripke structures* [Kr63]. We describe Kripke structures for positively introspective knowledge and explore the relationship between them and knowledge structures.

A Kripke structure for positively introspective knowledge is tuple $M = (W, \Pi, R)$, where W is a set of possible states, $\Pi : P \rightarrow \text{PowerSet}(W)$ assigns meaning to the atomic propositions, and $R : A \rightarrow W^2$ assigns to every agent a a reflexive and transitive relation on W . Intuitively, $(s, t) \in R(a)$ if t is a possible state for a in t . We now define what it means for a formula ϕ to be *satisfied* at a state s of M , written $M, s \models \phi$.

- $M, s \models p$, where p is an atomic proposition, if $s \in \Pi(p)$.
- $M, s \models \neg \phi$ if $M, s \not\models \phi$.
- $M, s \models \phi \wedge \psi$ if $M, s \models \phi$ and $M, s \models \psi$.
- $M, s \models K_a \phi$ if $M, t \models \phi$ for all t such that $(s, t) \in R(a)$.

It is not hard to show that with Kripke semantics, the modality K_a has all the properties discussed in Section 2 [HM85], but that seems to us less than satisfying, since it leaves to many notions unexplained. Both the notion of a state and the notion of possibility are left as primitive notions with no explication. The following theorem clarifies these issues by providing an exact correspondence between knowledge structures and states in Kripke structures.

Theorem 8:

- (1) To every Kripke structure M for positively introspective knowledge and every state s in M , there corresponds a knowledge structure $f_{M,s}$ such that $M, s \models \phi$ if and only if $f_{M,s} \models \phi$, for every formula ϕ ; and if s and t are states in M such that $(s, t) \in R(a)$ then $f_{M,s} \leq_a f_{M,t}$.
- (2) $M_{\text{know}} = (W, \Pi, R)$, where W is the set of all knowledge structures, $\Pi(p) = \{f : f \models p\}$, and $R(a) = \{(f, g) : f \leq_a g\}$, is a Kripke structure for positively introspective knowledge. Moreover, $M, f \models \phi$ if and only if $f \models \phi$, for every formula ϕ .

Theorem 8 shows that knowledge structures and Kripke structures have the same theory (as we already know from Theorem 3 and the results in [HM85] about axiomatization of validity in Kripke structures), but its implications are deeper. It shows that knowledge structures and Kripke structures complement each other in modelling knowledge: knowledge structures model states of knowledge, and Kripke structures model collections of knowledge states. In particular a state t is possible for a in a state s if a knows in t at least as much as it knows in s . Correspondence between fully introspective knowledge structures and Kripke structures for fully introspective knowledge was shown in [FHV84].

5. Circumscribed Knowledge

In principle there is no difference between *positive* information, the facts that are stated to be true, and *negative* information, the facts that are stated to be false. Indeed, in logic, a truth-assignment is a function from the set of atomic propositions to the set $\{\text{true}, \text{false}\}$, and the roles of *true* and *false* are completely symmetric. In common sense reasoning, however, we tend to use mostly the positive facts while ignoring a whole plethora of unstated negative facts. *Circumscription* [Mc80] is a formal tool to capture this paradigm of common sense reasoning. Our goal in this section is to circumscribe knowledge, that is, we want to give precise meaning to informal statements such as "I know that ϕ , and this is *all* I know".

We start by analyzing a simple case. Suppose that we are given a k -ary world w as a description of the agents' state of knowledge, and we are told that this is all the agents know. For example, suppose that we are told about Alice and Bob in Section 2 that this is all they know. Strictly speaking, this cannot be true. Since Bob knows p , he knows that he knows p , he knows that he knows that he knows p , etc. *ad infinitum*. What we want to find is some knowledge structure f , such that w is a prefix of f , i.e., $w = \langle f_0, \dots, f_{k-1} \rangle$, and such that the knowledge contained in w is circumscribed, i.e., there is no knowledge in f unless it follows, in some sense, from the knowledge in w .

The answer to that problem depends in part on the underlying model of knowledge acquisition. For example, is it possible for Alice to know that Bob knows about p without being told so by Bob, say by tapping into Bob's database? We consider here the most "permissive" situation, where agents have no knowledge about how other agents acquire knowledge. In this case $f_k(a)$ should include all the k -ary worlds that are possible for a in w , i.e., all the k -ary worlds where a knows at least as much as he knows in w .

This leads to the following definition, where we use the convention that $f_0(a)$ is the empty set: $f_k(a)$ is the *no-information extension*

of $f_{k-1}(a)$ if $f_k(a) = \{\langle g_0, \dots, g_{k-1} \rangle : g_{k-1}(a) \subseteq f_{k-1}(a)\}$. The *no-information extension of the k -ary world $w = \langle f_0, \dots, f_{k-1} \rangle$, denoted w^* , is the sequence $\langle f_0, \dots, f_{k-1}, f_k, \dots \rangle$, where $f_m(a)$ is the no-information extension of $f_{m-1}(a)$ for $m \geq k$ for each agent $a \in A$.*

Intuitively, the no-information extension of $f_{k-1}(a)$ describes what a knows at level $k+1$, given that she has no information besides that already described at $f_{k-1}(a)$ and given the underlying "permissive" model of knowledge acquisition described above.

To justify our definition of the no-information extension we have first to prove that w^* is indeed a knowledge structure (which is not a priori clear), and then we have to show that the knowledge in w is indeed circumscribed in w^* .

Theorem 9: Let w be a k -ary world. Then

- (1) w^* is a knowledge structure, and
- (2) if f is a knowledge structure such that w is a prefix of f , then $w^* \leq_a f$, for all $a \in A$.

Consider now a k -ary world $w = \langle f_0, \dots, f_{k-1} \rangle$. In such a world the agents have knowledge of depth $k-1$. Thus in w^* we would expect the agents to have knowledge of depth $k-1$ and no more. Now we have to be careful how we define depth of knowledge. If a knows p , then she knows that she knows p , she knows that she knows that she knows p , etc. This does not seem, however, as genuinely deep knowledge. If, on the other hand, a knows that b knows that a knows p , then it seems that a does have deep knowledge. We characterize *depth* of knowledge by the following definitions.

An a -formula, where a is an agent, is a formula that describes the knowledge of a . Formally,

- $K_a \phi$ is an a -formula.
- If ϕ is an a -formula and ψ is an arbitrary formula, then $\neg \phi$, $\phi \wedge \psi$, and $\psi \wedge \phi$ are a -formulas.

We can now define the $depth_a$ of a formula, which is the depth of a 's knowledge described by the formula.

- $depth_a(p) = 0$ if p is atomic propositions.
- $depth_a(\neg\phi) = depth_a(\phi)$.
- $depth_a(\phi/\psi) = \max\{depth_a(\phi), depth_a(\psi)\}$.
- $depth_a(K_a(\phi)) = 1 + \max\{depth_b(\phi) : b \neq a \text{ and } \phi \text{ is a } b\text{-formula}\}$. Otherwise $depth_a(K_a\phi) = depth_a(\phi)$.

Theorem 10: Let w be a k -ary world and let a be an agent. Then $w \models \phi$, where $depth_a(\phi) = k$ if and only if ϕ is a valid formula.

Corollary. Let w be a 1-ary world and let a be an agent. Then $w \models K_a\phi$ if and only if ϕ is a valid formula.

That is, 1-ary worlds consists only of a description of "nature" and they contain no knowledge.

The no-information extension was defined in [FHV84] for fully introspective knowledge. In that context, however, it is not true that w^* has the minimum amount of knowledge among all extensions of w , since, as we have argued before, ignorance leads to knowledge by negative introspection. In particular, even no information extensions of 1-ary worlds contain arbitrarily deep knowledge [FHV84]. Thus the only justification for the definition in [FHV84] is the intuition that lead to the definition in the first place.

We now consider statements of the form "all a knows is ϕ ". The first observation is that such statements need not necessarily make sense even if ϕ is a consistent formula [HM84b]. Consider the formula K_{ap}/K_aq . The statement "all a knows is that either she knows p or she knows q " does not make sense. Either a knows p , and then she knows that she knows p , or she knows q , and then she knows that she knows q . It just cannot be the case that all she knows is that either she knows p or either she knows q . In the terminology of [Mc80], $K_a\phi$ does not have a unique circumscription. When does the statement "all a knows is ϕ " make sense? Only when there is a unique way to circumscribe $K_a\phi$.

This motivates the following definition. A knowledge structure f is an a -minimal model for a formula ϕ if $f \models \phi$ and $f \leq_a g$ for all knowledge structures g such that $g \models \phi$. A formula ϕ is a -honest if $K_a\phi$ has an a -minimal model. The following theorem shows that a -minimal models for $K_a\phi$ indeed circumscribe the knowledge expressed by ϕ (just as minimal predicate circumscribe predicates in [Mc80]).

Theorem 11: Let f be an a -minimal model for $K_a\phi$. Then $f \models K_a\psi$ if and only if $K_a\phi \supset \psi$ is a valid formula.

We can now answer the question "what does a knows if all she knows is ϕ ?" The answer is that a knows exactly all the logical consequences of $K_a\phi$. Note that this is not the case for fully introspective knowledge. There, if all that a knows is p , then a does not know q , so she knows that she does not know q , even though $K_{ap} \supset K_a \neg K_a q$ is not a valid formula for fully introspective knowledge [HM84b]. Indeed, the issue of circumscription for fully introspective knowledge is less clear and less intuitive than it is for positively introspective knowledge (cf. [Ha84, HM84b, Pa84, St81]).

Our treatment here is closely related to Konolige's "circumscriptive ignorance" [Ko82]. Konolige introduced *circumscriptive modalities* $[\phi]$, where ϕ is a formula. Roughly speaking, $[K_a\phi]\psi$ holds if a knows ψ when all she knows is ϕ . Konolige's treatment is completely axiomatic; he adds the inference rule:

(R3) From ϕ/ψ (in the system $\{A1-3, R1-2\}$) infer $\neg[\phi]\psi$.

Since Konolige had no semantics, we view our treatment (Theorem 11) here is as an *ad fortiori* justification for his axiomatic approach.

We want to show now that we can test effectively for honesty of formulas. To do this we extend our definition of $\leq_a$ from structures to worlds in the natural way: $\langle f_0, \dots, f_k \rangle \leq_a \langle g_0, \dots, g_k \rangle$ if $g_k(a) \subseteq f_k(a)$.

Theorem 12: A formula ϕ , such that $depth(\phi) = k$ is a -honest if and only if there is a $(k+2)$ -ary world w such that $w \models K_a\phi$, and if w' is a $(k+2)$ -ary world such that $w' \models K_a\phi$, then $w \leq_a w'$.

Corollary: Honesty is decidable.

6. Concluding Remarks

We have defined a model-theoretic framework for positively introspective knowledge, which we argue is the right model for implicit knowledge. This framework provides precise means to compare knowledge states, which enabled us to define the notion of a minimal model. We have shown that using the notion of minimal models we can circumscribe knowledge.

Acknowledgements. The research reported here was supported by gift from the System Development Foundation. Much of the work was completed on computer equipment provided by the Xerox Corporation. I'd like to thank R. Fagin, Y.O. Moses, J.Y. Halpern, and the anonymous reviewers for their helpful comments on a previous draft of this paper.

References

- [Ch80] Chellas, B.F.: *Modal logic*. Cambridge Univ. Press, 1980.
- [FHV84] Fagin, R., Halpern, J.Y., Vardi, M.Y.: A model-theoretic analysis of knowledge. *Proc. 25th IEEE Symp. on Foundations of Computer Science*, West Palm Beach, 1984, pp. 268-278.
- [FV84] Fagin, R., Vardi, M.Y.: An internal semantics for modal logic. *Proc. 17th ACM Symp. on Theory of Computing*, 1985.
- [Ha84] Halpern, J.Y.: *Special Kripke models*, Unpublished manuscript.
- [Hi62] Hintikka, J.: *Knowledge and belief*. Cornell Univ. Press, 1962.
- [Hi75] Hintikka, J.: Impossible possible worlds vindicated. *J. Philosophy* 4(1975), pp. 475-484.
- [HM84a] Halpern, J.Y., Moses, Y.: Knowledge and common knowledge in distributed environments. *Proc. 3rd ACM Symp. on Distributed Computing*, 1984, pp. 50-61.
- [HM84b] Halpern, J.Y., Moses, Y.: Towards a theory of knowledge and ignorance. *Proc. AAAI Workshop on Nonmonotonic Reasoning*, 1984.
- [HM85] Halpern, J.Y., Moses, Y.O.: *A guide to the modal logics of knowledge and belief*. This volume.
- [Ko82] Konolige, K.: Circumscriptive ignorance. *Proc. 2nd National Conf. on Artificial Intelligence*, Pittsburgh, 1982, pp. 202-204.
- [Kr63] Kripke, S.A.: Semantical analysis of modal logic" *Zeit. fur Math. Logik und Grundlagen der Math.* 9(1963), pp. 67-96.
- [Le78] Lenzen, W.: Recent work in epistemic logic. *Acta Phil. Fenn.* 30(1978), pp. 1-219.
- [Le84] Levesque, H.J.: A logic of implicit and explicit belief. *Proc. 3rd National Conf. on Artificial Intelligence*, Austin, 1984, pp. 198-202.
- [Mc80] McCarthy, J.: Circumscription - A form of non-monotonic reasoning. *Artificial Intelligence* 13(1980), pp. 27-39.
- [Mc84] McCarthy, J.: Personal communication.
- [Mo80] Moore, R.C.: *Reasoning about knowledge and action*, Technical Note 191, SRI International, AI Center, 1984.
- [Mo84] Moore, R.C.: *A formal theory of knowledge and action*. Technical Note 320, SRI International, AI Center, 1984.
- [MSH178] J. McCarthy, M. Sato, T. Hayashi, S. Igarashi, *On the model theory of knowledge*, Stanford AI Lab Memo AIM-312, Stanford University, 1978.
- [OC68] O'Connor, D.J.: Beliefs, dispositions and actions. *Proc. Aristotelian Society* 69(1968), pp. 1-16.
- [Pa84] Parikh, R.: *Monotonic and non-monotonic logics of knowledge*. Unpublished manuscript, 1984.
- [Re74] Rescher, N.: Alternatives in epistemic logic. In *Studies in Modality*, American Phil. Quarterly Monograph Series #8, 1974, pp. 99-114.
- [RP85] Rosenschein, S.J., Pereira, C.N.: *Knowledge and action in situated automata*. Forthcoming.
- [Sa76] Sato, M.: *A Study of Kripke-type Models for Some Modal Logics*, Research Institute for Mathematical Science, Kyoto University, Kyoto, Japan, 1976.
- [St81] Stark, W.R.: A logic of knowledge. *Zeit. fur Math. Logik und Grundlagen der Math.* 27(1981), pp. 371-374.