Towards a Theory of Knowledge and Ignorance: Preliminary Report*

Joseph Y. Halpern
IBM Research, San Jose, CA 95193

Yoram Moses
Computer Science Department,
Stanford University, Stanford, CA 94305
and
IBM Research, San Jose, CA 95193

Abstract:

Situations in which the information about a given domain is partial are common in many AI applications. In planning and analysis of scenarios involving partial information, the state of knowledge of an intelligent agent in such circumstances becomes important. This paper addresses the problem of characterizing this state of knowledge, with the emphasis on the single-agent case. We give a number of equivalent ways to characterize this state of knowledge, as well as an algorithm for computing the formulas that are true in this state. The relationship between this work and related works by Stark, Konolige, and Moore is discussed.

1. Introduction

This research was originally motivated by the question of how communication in a distributed system changes the state of knowledge of the processors in the system (cf. [HM1]). Answering such a question clearly requires characterizing a processor's state of knowledge at a given point in time. To see some of the difficulties here, consider a processor that has only one bit of information, namely, that the propositional fact P is true. We assume that processors can do perfect propositional reasoning, so that our processor also knows all the logical consequences of P, but this is far from all it knows. Suppose Q is another propositional fact. By introspection it can discover that it doesn't know Q, and by further introspection it discovers that it *knows* that it doesn't know Q. (Note that we are assuming here that an ideal processor has perfect introspective knowledge about its knowledge and lack of knowledge.) But not knowing Q is *not* a logical consequence of knowing P.

The situation is further complicated by the presence of a second processor. Since the first processor does not know Q, it knows that the second processor cannot know that the first processor knows Q (we assume that only true facts can be known). And since the first processor also knows that the second processor can do perfect introspection, the first processor knows that the second processor knows that it does not know that the first processor knows Q. Thus a processor can make inferences about another processor's knowledge through its own ignorance! (See [FHV] for further discussion on this point.)

* This paper appears in *Logics and Models of Concurrent Systems* (ed. K. Apt), Springer-Verlag, 1985, pp. 459–476 and in *Proceedings of the Workshop on Non-Monotonic Reasoning*, 1984, pp. 125–143.

In order to focus in on these issues, we concentrate in this paper on the one-processor case. Many of the general problems are already present here. This discussion is not limited to the domain of distributed systems of processors. It applies just as well to intelligent robots functioning in the real world or to knowledge bases. Our assumptions that a processor or knowledge base can do perfect propositional reasoning and has complete introspection regarding its own knowledge and ignorance make our knowledge bases like those of Levesque [Le] and like Konolige's introspection machines [Ko1].

Intuitively, given a complete description of the information on which the knowledge is based, it seems that there should be a unique state of knowledge characterizing what the knowledge base knows. Every query of the form "Do you know q?" should have a unique answer. But, as the discussion above suggests, describing this state of knowledge is nontrivial. It is also easy to see that the state of knowledge changes non-monotonically as more information is acquired. If the knowledge base "knows only P", then it knows that it does not know Q. But if it later discovers Q, then of course it does know Q.

Further problems arise because some formulas do not uniquely characterize a knowledge state. For example, it cannot be the case that all a knowledge base knows is that it either knows P or it knows Q. (Note that this is quite different from knowing that one of P or Q holds.) If the only information the knowledge base has is that it knows P or it knows Q, then it doesn't know P (since all it knows is that it knows one of P and Q), and similarly, it doesn't know Q. But this state of affairs is inconsistent! A knowledge base cannot know one of P and Q without knowing either one of them.

In the next section, we introduce various approaches to the characterization of the state of an agent's knowledge corresponding to "knowing only α". In each of the approaches there are formulas that do not uniquely characterize an agent's state of knowledge. We call formulas that do uniquely characterize an agent's state of knowledge *honest*, while we call formulas that do not uniquely characterize a state of knowlege *dishonest*.* Intuitively, an agent is being dishonest if it claims to "know only α" for a dishonest formula α. All of the approaches are shown to lead to the same notions of honesty, and for honest formulas, they are shown to specify the same state of knowledge. One of the approaches also gives an algorithm that, given an honest formula α and a formula p, decides whether an agent whose sole information is α knows p.

As suggested in the discussion above, the multi-agent case is even harder to analyze than the single agent case, since nontrivial inferences about the knowledge and ignorance of other agents can now be made. In section 3, we briefly discuss how the results of section 2 can be extended to the multi-agent case.

Using the theory developed in sections 2 and 3, we can define a nonmonotonic provability relation $\mathrel{|\!\!\sim}_A$, where $\alpha \mathrel{|\!\!\sim}_A p$ exactly if agent A knows p, when his knowledge

* Mike Fischer and Neil Immerman suggested the use of the word honest for this notion. Neil Immerman first convinced us of the existence of dishonest formulas.

is based solely on α. This nonmonotonic provability relation is much in the spirit of Stark's nonmonotonic *model theorist's deduction rule* [St], but has wider applicability.

Two other works in a spirit similar to ours are those of Konolige [Ko2] and Moore [Mo2]. Konolige is also concerned with a situation where agents have only a limited amount of information, but in his formalism, agents cannot use introspection to acquire knowledge of their ignorance. Moore is concerned with describing what an ideally rational agent should *believe* (rather than know) given some information about the world. This leads to some interesting differences between our results and those of [Mo2]. The relationship between our work and that of Stark, Konolige, and Moore is discussed in section 4. We conclude with remarks about the relationship this work has with default rules in non-monotonic reasoning.

2. The knowledge theory

Let us first consider the case of a single knower or knowledge base. Imagine an ideal agent A with very powerful computing capabilities and perfect introspection. A knows precisely what facts he knows, and what facts he doesn't know. A lives in a propositional real world, and his conceptual world consists of formulas regarding the real world and his knowledge. The class $\mathcal{L}$ of formulas of the propositional logic of knowledge, within which A reasons, is defined as follows:

(L1) All primitive propositions P, Q, ... are formulas.
(L2) If p and q are formulas, then $\neg p$, $p \wedge q$, $p \vee q$, $p \supset q$ are formulas.
(L3) If p is a formula then $K_A p$ is a formula (denoting "A knows p").
(L4) The only formulas of $\mathcal{L}$ are those required by (L1)–(L3).

We call the set of formulas A knows at a given point in time A's *knowledge state* (cf. [MSHI]). Thus a formula p will be in A's knowledge state iff $K_A p$ is true. What properties should this set have? Let T be a knowledge state. Since we assume that A can do perfect propositional reasoning, we have:

(St1) All instances of propositional tautologies are in T.

A knows about modus ponens, therefore

(St2) If $p \in T$ and $p \supset q \in T$, then $q \in T$.

We also assume that A is capable of introspection with regards to his own knowledge, so:

(St3) $p \in T$ iff $K_A p \in T$.

(St4) $p \notin T$ iff $\neg K_A p \in T$.

Finally, we demand that a knowledge state be consistent

(St5) T is (propositionally) consistent.

Following Stalnaker [S], we call such a set a *stable* set of formulas. (Actually, Stalnaker does not require that a stable set satisfy (St5); we add this requirement for convenience. Note that the only set of formulas satisfying (St1)–(St4) and not satisfying

(St5) is the inconsistent set $\mathcal{L}$ of all the formulas in the language. Property (St5) simply says that $\mathcal{L}$ is not an admissible state of knowledge.) Properties (St3) and (St4) imply that lower depth formulas in a stable set determine those of higher depth. In fact we have:

Proposition 1 ([Mo2]): A stable set is uniquely determined by the propositional formulas it contains.

This result is also proved in [Mo2]; we reprove it here for completeness. We first need to prove the following:

Lemma 1: Let S be a stable set. For any formulas $p, q \in \mathcal{L}$, (a) $K_A p \vee q \in S$ iff $p \in S$ or $q \in S$, and (b) $\neg K_A p \vee q \in S$ iff $p \notin S$ or $q \in S$.

Proof: The 'if' direction in both (a) and (b) is immediate from (St1)–(St4). We now prove the 'only if' direction. Assume $K_A p \vee q \in S$. If $p \in S$ then we are done. Otherwise, $p \notin S$ and by (St4), $\neg K_A p \in S$. But since S is closed under propositional reasoning by (St1) and (St2), $q \in S$ must hold. To show (b), assume that $\neg K_A p \vee q \in S$. If $p \in S$ then $K_A p \in S$, and again by propositional reasoning $q \in S$ must hold. Otherwise, $p \notin S$ and we are done. □

Proof of Proposition 1: Assume that S and S' are two stable sets containing exactly the same propositional formulas. We will prove by induction on the depth of nesting of the K_A operators in a formula p that $p \in S$ iff $p \in S'$. For propositional formulas this is given. Assume that for all formulas of depth less than n the claim holds, and that p is a formula of depth n. By propositional reasoning, p is equivalent to a formula p' that is in "conjunctive normal form", i.e., p' is of the form $\bigwedge_i d_i$, where each d_i is a disjunction of the form $K_A q_1 \vee \cdots \vee K_A q_l \vee \neg K_A q_{l+1} \vee \cdots \vee \neg K_A q_m \vee g$, with the q_j's all formulas of degree less than n, and g a propositional formula. By (St1) and (St2), $p \in S$ iff $p' \in S$, and by propositional reasoning $(\bigwedge_i d_i) \in S$ iff $d_i \in S$ for all i. By Lemma 1, $d_i \in S$ iff either one of $g \in S$, $q_1 \in S$, ..., $q_l \in S, q_{l+1} \notin S, \ldots, q_m \notin S$ holds. An analogous property holds for S'. Since S and S' agree on all formulas of depth less than n we have $d_i \in S$ iff $d_i \in S'$. Therefore $p' \in S$ iff $p' \in S'$ and $p \in S$ iff $p \in S'$. □

Suppose α is a formula that describes all the facts that A has learned or observed. What is A's knowledge state when he "knows only α"? Clearly, this knowledge state should contain α, and since "only α" is known, it seems that it should be in some sense minimal among knowledge states containing α. However, the obvious notion of "minimal" – set inclusion – will not work. As the following proposition shows, no two knowledge states are comparable with respect to inclusion:

Proposition 2: No stable set properly includes another stable set.

Proof: Assume that a stable set S' properly includes a stable set S. There is some formula p such that $p \in S'$ and $p \notin S$. By properties (St3) and (St4) it follows that $K_A p \in S'$ and $\neg K_A p \in S$. Now since S' properly includes S, it must be the case that $\neg K_A p \in S'$, but then S' is inconsistent, a contradiction. □

By Proposition 1, a stable set is uniquely determined by the purely propositional formulas it contains. We denote by $Prop(S)$ the subset of S consisting of its purely propositional formulas. A possible candidate for the "minimal" knowledge state containing α is the stable set containing α whose propositional subset is minimum (w.r.t. inclusion). Not all formulas α have such a minimal set. For example, consider the formula $\alpha = K_A P \vee K_A Q$. Any stable set containing α must contain either P or Q. Furthermore, there is a stable set S_P that contains α and P but does not contain Q, and a stable set S_Q containing α and Q and not containing P. However, the intersection of $Prop(S_P)$ and $Prop(S_Q)$ contains neither P nor Q. Thus, there is no stable set T containing α with $Prop(T) \subset Prop(S_P)$ and $Prop(T) \subset Prop(S_Q)$. This leads us to the following definition: A formula α is *honest*$_S$ iff there exists a stable set containing α whose propositional subset is minimum. For an honest$_S$ formula, we denote this stable set by S^α.

Our intention is that S^α denote the stable set that describes A's state of knowledge if he "knows only α" (at least if α is an honest$_S$ formula). This definition may seem somewhat *ad hoc*, so we now consider a number of other ways of characterizing this state of knowledge.

Possible-world or *Kripke* semantics have been frequently used as a means of giving semantics to logics of knowledge (cf. [Hi,MSHI,Mo1]). Given our assumptions about an agent's power of introspection, the appropriate logic of knowledge is one that satisfies the axioms of the modal logic S5 (cf. [HC]), namely:

A1. All substitution instances of propositional tautologies.
A2. $K_A(p \supset q) \supset (K_A p \supset K_A q)$
A3. $K_A p \supset p$
A4. $K_A p \supset K_A K_A p$
A5. $\neg K_A p \supset K_A \neg K_A p$
A6. $K_A \psi$, if ψ is an instance of axiom A1 – A6.

The inference rule for S5 is *modus ponens*: from p and $p \supset q$ infer q.

In the case of one agent, the possible world semantics for S5 have a particularly simple structure: a (Kripke) model is just a nonempty set of *states*, where a state is an assignment of truth values to the primitive propositions of $\mathcal{L}$. We can think of these states as the worlds that the agent thinks are possible. Taken another way, they are the propositional assignments that do not contradict A's knowledge.

Given a model M, we now define what it means for $p \in \mathcal{L}$ to be true at a state $s \in M$, written $M, s \models p$, inductively as follows:

$M, s \models P$ iff $s(P) = \mathbf{true}$ (i.e., s assigns P the value **true**).

$M, s \models p \wedge q$ iff $M, s \models p$ and $M, s \models q$.

$M, s \models \neg p$ iff $M, s \not\models p$.

$M, s \models K_A p$ iff $M, t \models p$ for all $t \in M$.

Thus $K_A p$ is true at a state in a Kripke model if p is true in all the worlds that A thinks are possible, and $K_A p$ is false exactly if there is a possible world where p is false.

This semantics precisely captures the intuition captured in the axioms above. In fact, as Kripke showed,

Theorem 1 ([Kr]): Axioms A1—A6, together with the inference rule *modus ponens*, form a sound and complete axiomatization for (S5) Kripke models.

We now relate Kripke models for knowledge and stable sets. Given a Kripke model M, we define $K(M)$, the set of facts that are known in M, to be the set $\{p : M, t \models p$ for all $t \in M\}$. Note that $p \in K(M)$ iff $M, t \models K_A p$ for all states $t \in M$ and $p \notin K(M)$ iff $M, t \models \neg K_A p$ for all states $t \in M$.

Lemma 2: If M is a Kripke model then $K(M)$ is a stable set.

Proof: Axiom A6 implies that $K(M)$ satisfies (St1) by A1, (St2) by A2, (St3) by A3 and A4, and (St4) by A5. By Theorem 1, $K(M)$ is consistent and therefore satisfies (St5). □

Proposition 3: Every stable set S determines a Kripke model M_S for which $S = K(M_S)$. Furthermore, if $\mathcal{L}$ has only a finite number of primitive propositions, then M_S is the unique Kripke model with this property.

Proof: Given S, let M_S consist of all the states consistent with S; i.e.,

$$M_S = \{s : s \text{ is a state that satisfies all the propositional formulas of } S\}.$$

By Lemma 2, $K(M_S)$ is a stable set. By the definition of M_S, the propositional formulas of $K(M_S)$ are exactly those of S. By Propostition 1, this implies that $S = K(M_S)$. Now assume that $\mathcal{L}$ has only finitely many primitive propositions, say $P_1, \ldots, P_N$. To show that the model M_S is the unique Kripke model such that $S = K(M_S)$, it suffices to show that if M and M' are two Kripke models and $M \neq M'$, then $K(M) \neq K(M')$.

So suppose $M \neq M'$. Without loss of generality, there is some state s such that $s \in M$ and $s \notin M'$. Let g be the propositional formula that completely describes the assignment s, namely $g = q_1 \wedge \ldots \wedge q_n$, where $q_i = P_i$ if $s(P_i) = \textbf{true}$ and $q_i = \neg P_i$ if $s(P_i) = \textbf{false}$. It is now easy to see that $\neg g \in K(M')$, since $s \notin M'$, but $\neg g \notin K(M)$ since $M, s \models g$. So $K(M) \neq K(M')$, and we are done. □

Lemma 2 and Proposition 3 were also observed independently by R. Moore, M. Fitting, and J. van Benthem. As a corollary to Proposition 3, we get:

Corollary 1: Stable sets are closed under S5 consequence. □

We remark that Corollary 1 shows that we could have replaced (St1) by (St1$'$):

(St1$'$) T contains all instances of S5 tautologies.

Which Kripke model is the one where A knows "only α"? Recall that a Kripke model consists of states which may be viewed as the worlds that A thinks are possible. Thus if $M \supset M'$, then the intuition is that A "knows less" in M than he does in M', since there are more worlds that he thinks the real world could be. Since the model in which A knows "only α" is intuitively the model in which he knows the least among all the models in which he knows α, this suggests that the appropriate model is one which is a superset of any other model in which A knows α. Denote by M_α the model that is the union of all models in which $K_A\alpha$ holds. By the above analysis, M_α should be the model that corresponds to A's state of knowledge when he "knows only α". However, it turns out that in some cases $\alpha \notin K(M_\alpha)$, meaning that $K_A\alpha$ does not hold in M_α. In those cases it seems that there is no good candidate for A's knowledge state when he knows only α. We call a formula α *honest*$_M$ if $\alpha \in K(M_\alpha)$.* An example of a formula that is not honest$_M$ is the familiar $\alpha = K_AP \vee K_AQ$. The models $M_P = \{s : s(P) = \mathbf{true}\}$ and $M_Q = \{s : s(Q) = \mathbf{true}\}$ both satisfy $K_A\alpha$. Let $M' = M_P \bigcup M_Q$. M' is a submodel of M_α, and therefore $K(M') \supseteq K(M_\alpha)$. But $\alpha \notin K(M')$ (check!) and therefore $\alpha \notin K(M_\alpha)$, and thus our chosen α is not honest$_M$.

As we show below, the notions of honest$_M$ and honest$_S$ coincide, as do $K(M_\alpha)$ and S^α for an honest formula α. But before we do this we present yet another approach to the problem, this one motivated by the intuition that given formulas p and α, there ought to be an algorithm for deciding if A knows p given that A knows only α. We now present such an algorithm. Our algorithm constructs a set D^α which is intended to consist of the facts that A knows, if A "knows only α".

What formulas belong in D^α? Since our perfect reasoner knows that his knowledge satisfies S5, any formula p for which $K_A\alpha \supset p$ is S5-valid should surely be in D^α. (We remark that S5 validity is decidable. Ladner [L] shows that, in the case of one knower, S5 validity is in Co-NP, which makes it no harder than the validity problem for propositional logic.) However, our previous remarks show that more than just the logical consequences of α should be in D^α. For example, if $\alpha = P$, then $\neg K_AQ$ is in D^α, as is $P \wedge \neg K_AQ$.

The algorithm is simply:

$p \in D^\alpha$ iff $[K_A\alpha \wedge \psi_\alpha(p)] \supset p$ is S5-valid,

where $\psi_\alpha(p)$ is the conjunction of K_Aq for all subformulas K_Aq of p for which $q \in D^\alpha$, and $\neg K_Aq$ for all subformulas K_Aq of p for which $q \notin D^\alpha$ (where p is considered a subformula of itself).

It is easy to see that the above algorithm decides for any formula in the language whether or not it is a member of D^α. In order to decide if $p \in D^\alpha$, we must only invoke the algorithm on strict subformulas of p and then use the decision procedure for S5. Note that if a formula p is n characters long then it has no more than n subformulas, so that $\psi_\alpha(p)$ is finite and not much larger than p.

The intuition behind the algorithm is that a formula p is in D^α exactly if it is a logical consequence of knowing α and the K_A-subformulas of p that have already been decided. To understand what the algorithm does a bit better, the reader should note

* M. Vardi first suggested this definition of honesty to us. H. Levesque suggested it independently.

that a propositional formula g is in D^α exactly if $K_A\alpha \supset g$ is S5-valid. If $q \in D^\alpha$, then by definition K_Aq is one of the conjuncts of $\psi_\alpha(K_Aq)$, so $K_Aq \in D^\alpha$. Similarly, if $q \notin D^\alpha$, then $\neg K_Aq$ is one of the conjuncts of $\psi_\alpha(\neg K_Aq)$ and $\neg K_Aq \in D^\alpha$.

This discussion suggests that D^α is a stable set, but this is not necessarily true. It turns out that there are formulas α for which D^α is not consistent. For example, consider $\alpha = P \wedge \neg K_AP$. Clearly α is consistent, but $K_A\alpha$ implies both K_AP and $\neg K_AP$, and therefore is not consistent. Obviously, if $K_A\alpha$ is inconsistent then **false** $\in D^\alpha$, hence D^α is also inconsistent.

Even for formulas α for which $K_A\alpha$ is consistent, the set D^α generated by the algorithm might not be consistent. Consider again the formula $\alpha = K_AP \vee K_AQ$. The reader can easily verify that $\neg K_AP \in D^\alpha$, $\neg K_AQ \in D^\alpha$, and therefore $\neg K_AP \wedge \neg K_AQ \in D^\alpha$. But $\alpha = K_AP \vee K_AQ \in D^\alpha$. So, for this α, D^α is inconsistent. We call a formula α *honest*$_D$ if the set D^α is S5-consistent. At this point, the reader will not be surprised to learn that all the notions of honesty that we have defined coincide. We prove this in Theorem 2 below, but first we need:

Proposition 4: If α is honest$_D$ then D^α is a stable set.

Proof: Suppose α is honest$_D$. For any propositional tautology g, $K_A\alpha \supset g$ is S5-valid (by A1 and some straightforward propositional reasoning), so $g \in D^\alpha$ and (St1) is satisfied. By the discussion above, (St3) and (St4) are satisfied. By assumption, α is honest$_D$, so D^α is consistent and (St5) holds. Finally, for (St2), suppose that $p \supset q \in D^\alpha$, $p \in D^\alpha$, and $q \notin D^\alpha$. By (St3) and (St4), it follows that $K_A(p \supset q) \in D^\alpha$, $K_Ap \in D^\alpha$, and $\neg K_Aq \in D^\alpha$. Thus D^α is not S5-consistent, which contradicts the assumption that α is honest$_D$. □

Our standard example of a "troublesome" formula is $K_AP \vee K_AQ$. As we have argued earlier, it is inconsistent for A to know this formula while knowing neither P nor Q. More generally, we say that a formula α is honest$_K$ if it satisfies the *propositional disjunction property*: whenever $K_A\alpha \supset K_Ag_0 \vee K_Ag_1 \vee \cdots \vee K_Ag_m$ is an S5-valid formula, where $g_0, \ldots, g_m$ are propositional formulas, it is the case that $K_A\alpha \supset g_j$ is S5-valid, for some $0 \le j \le m$. A formula α that satisfies the propositional disjunction property also satisfies a slightly stronger property, namely: Whenever $K_A\alpha \supset g_0 \vee K_Ag_1 \vee \cdots \vee K_Ag_m$ is an S5-valid formula, where $g_0, \ldots, g_m$ are propositional formulas, it is the case that $K_A\alpha \supset g_j$ is S5-valid, for some $0 \le j \le m$. This follows because in S5 $K_A\alpha \supset K_Ag_0 \vee K_Ag_1 \vee \cdots \vee K_Ag_m$ and $K_A\alpha \supset g_0 \vee K_Ag_1 \vee \cdots \vee K_Ag_m$ are equivalent.

We are finally ready to prove

Theorem 2:

(a) A formula α is honest$_M$ iff it is honest$_D$ iff it is honest$_S$ iff it is honest$_K$.

(b) For an honest α, $K(M_\alpha) = D^\alpha = S^\alpha$.

Proof: We will prove (a) by showing a cycle of implications involving the different notions of honesty. (b) will follow from the proof of (a).

honest$_M \Rightarrow$ honest$_D$: If α is honest$_M$ then M_α is the maximum model that satisfies $K_A\alpha$. We claim that $K(M_\alpha) = D^\alpha$, and show this by proving, by induction on the structure of p, that $p \in D^\alpha$ iff $p \in M_\alpha$. Thus assume that for any strict subformula q of p, we have $q \in D^\alpha$ iff $q \in K(M_\alpha)$. Suppose $p \in D^\alpha$. Thus we must have $\models_{S5} K_A\alpha \wedge \psi_\alpha(p) \supset p$. For every conjunct of the form K_Aq in $\psi_\alpha(p)$, we must have $q \in D^\alpha$, and thus by inductive hypothesis $q \in K(M_\alpha)$, so that $M_\alpha, t \models K_Aq$ for every state $t \in M$. Similarly, for every conjunct of the form $\neg K_Aq$ in $\psi_\alpha(p)$, we have $q \notin D^\alpha$, so $q \notin K(M_\alpha)$ and $M_\alpha, t \models \neg K_Aq$ for all $t \in M_\alpha$. Thus $M_\alpha, t \models K_A\alpha \wedge \psi_\alpha(p)$ for all $t \in M_\alpha$. Since $\models_{S5} K_A\alpha \wedge \psi_\alpha(p) \supset p$, we must have $M_\alpha, t \models p$ for all $t \in M_\alpha$, and thus $p \in K(M_\alpha)$. For the converse, suppose $p \in K(M_\alpha)$, but $p \notin D^\alpha$. Thus it follows that $K_A\alpha \wedge \psi_\alpha(p) \wedge \neg p$ is S5-consistent, so there must be a model M' and a state $s' \in M'$, such that $M', s' \models K_A\alpha \wedge \psi_\alpha(p) \wedge \neg p$. Since $M' \subset M_\alpha$, we have $s' \in M_\alpha$. We claim that necessarily $M_\alpha, s' \models \neg p$. This contradicts the assumption that $p \in K(M_\alpha)$, so once we prove the claim we will be done. We prove the claim by showing, by induction on the structure of subformulas q of p, that if both $M_\alpha, s' \models \psi_\alpha(q)$ and $M', s' \models \psi_\alpha(q)$, then $M_\alpha, s' \models q$ iff $M', s' \models q$. The cases where q is a primitive proposition, a conjunction, or a negation are all straightforward and left to the reader. If q is of the form K_Aq', then $M_\alpha, s' \models K_Aq'$ iff K_Aq' is one of the conjuncts of $\psi_\alpha(K_Aq')$ (since $M_\alpha, s' \models \psi_\alpha(K_Aq')$ by hypothesis, and one of K_Aq' and $\neg K_Aq'$ must be a conjunct of $\psi_\alpha(K_Aq')$) iff $M', s' \models K_Aq'$. Since $K(M_\alpha) = D^\alpha$, D^α must be consistent, so α is honest$_D$.

honest$_D \Rightarrow$ honest$_S$: By Proposition 4, if α is honest$_D$ then D^α is stable. By construction, $\alpha \in D^\alpha$, and for any propositional formula g, we have $g \in D^\alpha$ iff $K_A\alpha \supset g$ is S5-valid. Thus, D^α must be the stable set containing α with the smallest propositional subset. Therefore α is honest$_S$ and $D^\alpha = S^\alpha$.

honest$_S \Rightarrow$ honest$_K$: Since stable sets are closed under S5 consequence (Corollary 1), if $K_A\alpha \supset K_Ag_0 \vee K_Ag_1 \vee \cdots \vee K_Ag_m$ is S5-valid, then every stable set containing α (and thus also $K_A\alpha$ by (St3)) must also contain $K_Ag_0 \vee K_Ag_1 \vee \cdots K_Ag_n$. By repeated applications of Lemma 1, it follows that every stable set containing α must contain one of the g_i's. Given that α is honest$_S$, S^α is a stable set containing α. Let g_{i_0} be one of the g_i's such that $g_{i_0} \in S^\alpha$. g_{i_0} appears in all the stable sets that contain α. It follows that $K_A\alpha \supset g_{i_0}$ must be S5-valid. Otherwise, $K_A\alpha \wedge \neg g_{i_0}$ is S5-consistent, so for some model M and state $s \in M$, we have $M, s \models K_A\alpha \wedge \neg g_{i_0}$. Thus $K(M)$ is a stable set containing α but not containing g_{i_0}, a contradiction.

honest$_K \Rightarrow$ honest$_M$: Let $\mathcal{L}'$ be the sublanguage of $\mathcal{L}$ whose only primitive propositions are those that appear in α. It is straightforward to check that α has a maximum model with respect to $\mathcal{L}'$ (i.e. where the states only give truth assignments to the propositions in $\mathcal{L}'$) iff it has a maximum model with respect to $\mathcal{L}$. Thus without loss of generality we can assume that we are dealing with a language with only finitely many primitive propositions. Assume that α is honest$_K$ and not honest$_M$. Thus there are maximal models (w.r.t. inclusion) of $K_A\alpha$, but their union – M_α – is not a model of $K_A\alpha$. Since we have assumed that we are dealing with a language with only finitely many primitive propositions, there can only be finitely many maximal models of $K_A\alpha$, say $M_0, \ldots, M_{m-1}$. Each model M_i must have a state, say s_i which is not in $M_{i+1(mod\ m)}$. Let g_i be the formula that completely describes s_i, constructed just as in the proof of Proposition 3. The proof of Proposition 3 also shows that $\neg g_i \in K(M_{i+1(mod\ m)})$. It follows immediately from the Kripke semantics for S5 that $\neg g_i \in K(M')$ for any model $M' \subset M_{i+1(mod\ m)}$. Thus the formula $K_A\neg g_0 \vee \cdots \vee K_A\neg g_{m-1}$ is true at every state in every model of $K_A\alpha$. By Theorem 1, it follows that $K_A\alpha \supset K_A\neg g_0 \vee \cdots \vee K_A\neg g_{m-1}$ is an S5-valid formula. From (g), it follows that $K_A\alpha \supset \neg g_i$ must also be S5-valid for some i. But by construction, $M_i, s_i \models g_i \wedge K_A\alpha$, a contradiction. □

Theorem 2 indicates that the notion of what an agent knows if it "knows only α" is quite robust, as is the notion of honesty. Since we have proved that all our notions of honesty coincide, we will henceforth drop the subscript. The proof of the theorem also shows

Theorem 3: Honesty is decidable.

Proof: To check whether α is honest, it suffices to consider Kripke models for the subset of $\mathcal{L}$ containing only the primitive propositions that appear in α. There are only finitely many such Kripke models, and by enumerating them and checking in which ones α is known, it is simple to check whether $\alpha \in K(M_\alpha)$. □

Of course, the decison procedure for honesty described above is computationally ineffecient, taking both exponential time and space. In the full paper we show that α is honest iff $\psi_\alpha(K_A\alpha)$ is S5-consistent. This gives us a method of deciding honesty which takes space linear in the size of α.

3. Extending to many knowers

In section 2, the agent A is the only knower. A can reason only about propositional facts in the world and about his own knowledge. In a society of intelligent agents, A can also gather information about other agents' knowledge. As mentioned in the introduction, the many-knower case is much more complex than the one-knower case. We briefly discuss some of the issues here.

We extend the language $\mathcal{L}$ to allow formulas of the form K_Bp, for all agents B. We can now define a notion of stable set that describes the properties of A's knowledge state in the many knowers case. Clearly properties (St1) – (St5) still hold, but (St1) is not quite strong enough. For example, A might know that B doesn't know P, i.e., $\neg K_BP \in T$. But since A knows that B can do perfect introspection, A also knows that B knows that B doesn't know P, so $K_B\neg K_BP \in T$. We could therefore expect $K_B\neg K_BP \in T$ whenever $\neg K_BP \in T$. Although $\neg K_BP \supset K_B\neg K_BP$ is a tautology of S5, it is not a propositional tautology. Thus we must strengthen (St1) to

(St1′) T contains all instances of S5 tautologies.

Recall that in the one-knower case, replacing (St1) by (St1′) yielded the same notion; this is not true in the many-knower case. We call a set satisfying (St1′), (St2) – (St5) *stable with respect to A*.

We remark that the validity problem for S5 with many knowers is complete for polynomial space (see [HM2] for a proof), while for a single knower it is in Co-NP. This increase in complexity supports our experience that the many-knower case is more complicated than the single knower case.

Characterizing honesty and the state of knowledge described by an honest α in the many-knower case is quite subtle. We outline some of the difficulties here, leaving details to the full paper.

For example, it is easy to see that the analogue of Proposition 1 no longer holds in the many-knower case, i.e., the propositional formulas no longer determine a stable set. Indeed, even all of the K_A-free formulas do not uniquely determine a set that is stable w.r.t. A. It is easy to construct two sets, both stable w.r.t. A, that agree on all K_A-free formulas, but differ on what agent B knows about what agent A knows. Thus the obvious approach to defining minimality for stable sets (in order to define the stable set S^α where A "knows only α") will not work. At this point, finding such a definition remains an open problem.*

Since propositional formulas no longer play the same essential role in the many-knower case, it should come as no surprise that honesty$_K$ (satisfying the propositional disjunction property) does not correspond to our intuitive notion of honesty in the many-knower case. For example, one can show that $K_A K_B P \vee K_A \neg K_B P$ is honest$_K$, although it should be considered dishonest for the same reasons $K_A P \vee K_A Q$ was dishonest in the single-knower case.

A more promising approach is via Kripke models. As is well-known, Kripke models for multi-agent S5 can be constructed (see, for example, [MSHI,Mo1,HM2]), and given a Kripke model M, it is straightforward to define a stable set $K(M)$ in a manner completely analogous to the one-knower case, and show that every stable set is of the form $K(M)$ for some Kripke model. We can also define a notion of a *canonical* Kripke model, for which a notion of a maximum model M_α makes sense, and thus define a notion of honest$_M$. The algorithm of the previous section can also be extended to the many-knower case, leading to a definition of honest$_D$ that is provably equivalent to

---

* A stable set w.r.t. A *is* uniquely determined by its non-K_A formulas (i.e., those in which no K_A operator appears at top level). However, no formula α has a stable set with a minimum non-K_A subset. To see this, pick your favorite α. Let Q be a primitive proposition that does not appear in α, and is therefore independent of $K_A\alpha$. Any stable set that contains α but does not contain Q contains $\neg K_A Q$ and $\neg K_B K_A Q$. But there is a stable set containing α and Q that does not contain $\neg K_B K_A Q$. Therefore, there is no stable set containing α with a minimum non-K_A subset.

honest$_M$. We leave further details of the multi-agent case to the full paper.

4. Comparison to related work

As we remarked in the introduction, our work is closely related to [St], [Ko2] and [Mo2]. We discuss the relationship in this section.

Suppose we define a nonmonotonic provability relation $\mid\!\sim_A$ via

$$\alpha \mid\!\sim_A p \ \text{ iff } \ p \in D^\alpha.$$

Of course, if α is dishonest, then $\mid\!\sim_A$ "proves" inconsistent statements, so we restrict our attention to honest α's.

In [St], Stark introduces a nonmonotonic logic MK, where a nonmonotonic inference rule, the *model theorist's deduction rule* (m.t.d.r.), that allows individuals to reason about their ignorance, is added to the S5 axioms. The rule is:

$$\text{From } T \not\vdash_{S5} K_A p \ \text{ deduce } T \vdash_{MK} \neg K_A p.$$

The intuition behind this rule is that if T characterizes A's knowledge, and from T it is not possible to conclude (using S5) that A knows p, then indeed A does not know p.

As Stark observes, we quickly run into inconsistencies if we allow unrestricted use of this rule. For example, if $T = \{\mathbf{true}\}$, the empty theory, we get $T \not\vdash_{S5} K_A p$ and $T \not\vdash_{S5} K_A \neg K_A p$, so by the m.t.d.r. we get $T \vdash_{MK} \neg K_A p$ and $T \vdash_{MK} \neg K_A \neg K_A p$. But $\vdash_{S5} \neg K_A \neg K_A p \equiv K_A p$, which gives us an inconsistency.*

Stark's solution to this problems is to restrict the p in the conclusion of the m.t.d.r. to be K_A-free. However, this restriction limits the usefulness of the rule. In a multi-agent scenario, A might certainly want to reason about what B knows about A's knowledge. This restriction also has the effect that the m.t.d.r. cannot be used repeatedly, since the result of an application of m.t.d.r. is a K_A formula.

Our $\mid\!\sim_A$ relation satisfies a cleaner version of the m.t.d.r. rule, namely:

$$\text{If } \ \alpha \not\mid\!\sim_A K_A p \ \text{ then } \ \alpha \mid\!\sim_A \neg K_A p.$$

In this rule the application of the model theorist's deduction is done within the theory. Furthermore, for honest α's its usage is unrestricted. Thus $\mid\!\sim_A$ captures the intuitive intent of the m.t.d.r. without imposing any unnatural restrictions. (See also [Pa] for a treatment of this issue.)

In [Ko2], Konolige develops a theory of "Circumscriptive Ignorance". He treats knowledge for which he assumes the axioms of S4; i.e., S5 without the axiom $\neg K_A p \supset$

---

* Stark also notices that if T is a dishonest formula, such as $K_A P \vee K_A Q$, we again get an inconsistency (since $T \vdash_{MK} \neg K_A P$ and $T \vdash_{MK} \neg K_A Q$). He shows that no problems arise if T is *natural* in that whenever $T \vdash_{S5} K_{A_{i_1}} q_1 \vee \ldots \vee K_{A_{i_m}} q_m$, then $T \vdash_{S5} K_{A_{i_j}} q_j$ for some $1 \leq j \leq m$. Compare this with our notion of honest$_K$ and Theorem 2.

$K_A\neg K_A p$. In his formalism, $[K_A p]q$ denotes that A knows q if all he knows is p. Roughly speaking, $[K_A p]q$ if $\models_{S4} K_A p \supset q$. This roughly corresponds to $p \mid\!\sim_A q$. Indeed, $[K_A p]q$ implies $[K_A p]K_A q$, just as $p \mid\!\sim_A q$ implies $p \mid\!\sim_A K_A q$. However, while $p \not\mid\!\sim_A q$ implies $p \mid\!\sim_A \neg K_A q$, it is *not* the case that $\neg[K_A p]q$ implies $[K_A p]\neg K_A q$. (The fact that q is not an S4 consequence of $K_A p$ does not imply that $\neg K_A q$ is an S4 consequence of $K_A p$.) Thus, in Konolige's formalism, although given α you can deduce your ignorance, there is no way to incorporate this information into the set of things you know.

In [Mo2], Moore presents a non-monotonic logic of belief, where he assumes that belief satisfies the axioms of K5 (S5 without the axiom $Kp \supset p$). Moore defines a stable set T to be "grounded in a set of premises α" if, roughly speaking, a rational agent is justified in believing T given that the agent knows only α. More formally, Moore denotes belief by L, and defines T to be a *stable expansion of* α if T is equal to the set of *propositional* consequences of

$$\{\alpha\} \bigcup \{Lp : p \in T\} \bigcup \{\neg Lp : p \notin T\}.$$

Not surprisingly, it turns out that for many formulas α, T is a stable expansion of α exactly if $T = S^\alpha$. In fact, it can be shown that if an honest formula has a stable expansion, then that set is unique and equals S^α.

However, it turns out that there are formulas, both honest and dishonest, that have no stable expansion, and dishonest formulas that have a stable expansion. The reason hinges on the difference between knowledge and belief. For example, the formula $\alpha = LP$ has no stable expansion. Technically, this happens because any stable set containing LP must contain P, but P is not a propositional consequence of any set of L and $\neg L$ formulas. More informally, this is true because believing P does not give any grounds for concluding that P is true in the world. On the other hand, $K_A P$ is honest because an ideally rational agent that claims to know only $K_A P$ is completely describing its state of knowledge (and incidentally also saying that P is true in the world). Conversely, $\neg K_A P \supset Q$ (or equivalently $K_A P \vee Q$) is dishonest, while $\neg LP \supset Q$ has a unique stable expansion. We return to this point in the next section.

Roughly speaking then, if we consider knowledge rather than belief, an ideally rational agent knowing α also knows all the facts about the world that are S5 consequences of $K_A\alpha$. This suggests the following alternative definition:

A set R is *rooted in* α if R equals the set of propositional consequences of

$$\{\text{propositional } g : K_A\alpha \supset g \text{ is S5} - \text{valid}\} \bigcup \{K_A p : p \in R\} \bigcup \{\neg K_A p : p \notin R\}.$$

This notion is clearly closely related to the Moore's notion of a stable expansion. The following theorem relates it to our other notions.

Theorem 4:

(a) For all α, there is a unique set R^α rooted in α.

(b) If $K_A\alpha$ is consistent then R^α is stable; otherwise, R^α is the inconsistent set $\mathcal{L}$ of all formulas.

(c) α is honest iff R^α is consistent and $\alpha \in R^\alpha$.

(d) For an honest α, $R^\alpha = S^\alpha = D^\alpha = K(M_\alpha)$.

Proof: Let R be a set that is rooted in α. If $K_A\alpha$ is inconsistent, then **false** $\in R$, and because R contains all the propositional tautologies and is propositionally closed, $R = \mathcal{L}$. Otherwise it is easy to check that R must be the stable set S such that $Prop(S) = \{\text{propositional } g : K_A\alpha \supset g \text{ is S5} - \text{valid}\}$. The theorem follows from these observations; we leave details to the reader. □

5. Conclusions

The main purpose of this paper was to investigate an agent's state of knowledge when her knowledge is based on a formula α alone. One scenario where this arises is when α is the formula that completely describes the agent's database of known facts. We made a number of attempts at characterising this state of knowledge, motivated by semantic considerations and heuristic guidelines. For a certain class of formulas α that we call "honest", the state of knowledge corresponding to knowing "only α" turned out to be the same in all our approaches. For α's that are not honest, none of the approaches specifies a state of knowledge corresponding to knowing "only α". This suggests that the notions involved are in some precise sense robust, and do not depend in an essential way on the specific definition of what "knowing only α" means.

Let us briefly consider how this work relates to default rules in non-monotonic reasoning (cf. [Re]). Roughly speaking, a standard default rule has the form $\neg K_A p \supset q$, meaning that q (the default) is true unless p is known to be true. If p and q are propositional formulas, then the formula $\neg K_A p \supset q$ by itself is dishonest. In fact, for an honest α, this formula is a consequence of "knowing only α" exactly if one of p or q is. It follows that formulas of the form $\neg K_A p \supset q$ do not behave as default rules in our formalism. It might seem that we must now make a grave decision: either give up on default rules, viewing the results of this paper as sound technical testimony to the inexistence of consistent non-monotonic default rules, or give up on the modal logic S5 as an appropriate way to model our knowledge, and resort either to a logic that exludes one or both of the introspective axioms, or to a logic of belief such as K5. In a sense, Moore [Mo2] chooses to go to belief.

It is our opinion that for certain applications the axioms of S5 are indeed a good and useful way of modelling an agent's knowledge (cf. [HM1]). Nevertheless, we believe that in such circumstances it is often desirable to have and use default rules in order to compensate for the agent's lack of complete information. It seems that we can have our cake and eat it too if we extend our language to talk about knowledge *and* belief. A default rule is in some sense a "rule of conjecture", because for the propositional p and q above it is not our knowledge or ignorance of p that makes q true, but it is our information regarding our knowledge-gathering capabilities that leads us to believe q in the absence of our knowledge of p. It follows that in many cases of interest, the default rules can be taken to be "rules of conjecture", and therefore may be believed rather than known. A default rule may therefore be written as

$$B_A(\neg K_A p \supset q),$$

where B_A stands for A's belief. We may assume that the belief operator is fully introspective and in fact satisfies the axioms of K5 (cf. [Le]), and the axiom linking knowledge to belief is

$$K_A p \supset B_A K_A p.$$

We conjecture that such an approach can be successfully carried out, and that it may turn out combine knowledge and belief in a framework that retains the best of both worlds.

Acknowledgements: We would like to thank Ron Fagin, Mike Fischer, Neil Immerman, Hector Levesque, Bob Moore, Johan van Benthem, and Moshe Vardi for many stimulating discussions on knowledge which helped decrease our ignorance. We also thank Bob Moore for pointing out an error in an early version of Theorem 4. The work of Yoram Moses was supported in part by DARPA contract N00039-82-C-0250.

References

[FHV] R. Fagin, J. Y. Halpern, and M. Y. Vardi, A model-theoretic analysis of knowledge: preliminary report, IBM RJ 4373; to appear in *Proceedings of the 25th IEEE Symposium on Foundations of Computer Science*, 1984.

[HM1] J. Y. Halpern and Y. Moses, Knowledge and common knowledge in a distributed environment, *Proceedings of the 3rd ACM Symposium on Principles of Distributed Computing*, 1984, pp. 50–61.

[HM2] J. Y. Halpern and Y. Moses, A guide to modal logics of knowledge, to appear as an IBM RJ, 1984.

[Hi] J. Hintikka, *Knowledge and Belief*, Cornell University Press, 1962.

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

[Kr] S. Kripke, Semantical considerations of modal logic, *Zeitschrift fur Mathematische Logik und Grundlagen der Mathematik* **9**, pp. 67–96, 1963.

[Ko1] K. Konolige, A deduction model of belief, unpublished manuscript, 1984.

[Ko2] K. Konolige, Circumscriptive ignorance, *Conference Proceedings of AAAI–82*, pp. 202–204.

[La] R. E. Ladner, The computational complexity of provability in systems of modal propositional logic, *Siam J. Computing*, **6**:3, 1977, pp. 467–480.

[Le] H. J. Levesque, A formal treatment of incomplete knowledge bases, Fairchild Technical Report No. 614, FLAIR Technical Report No. 3, 1982.

[Mc] J. McCarthy, Circumscription – a form of non-monotonic reasoning, *Artificial Intelligence* **13**, 1980, pp. 27–39.

[MSHI] J. McCarthy, M. Sato, T. Hayashi, S. Igarashi, On the model theory of knowledge, *Computer Science Technical Report* STAN-CS-78-657, Stanford University, April 1978.

[Mo1] R. C. Moore, Reasoning about knowledge and action, *Artificial Intelligence Center Technical Note* 191, SRI International, 1980.

[Mo2] R. C. Moore, Semantical considerations on nonmonotonic logic, *SRI International Technical Note 284*, 1983.

[Pa] R. Parikh, Monotonic and non-monotonic logics of knowledge, unpublished manuscript, 1984.

[Re] R. Reiter, A logic for default reasoning, *Artificial Intelligence* **13**, 1980, pp. 81–132.

[S] R. Stalnaker, A note on non-monotonic modal logic, unpublished manuscript, Department of Philosophy, Cornell University.

[St] W. R. Stark, A logic of knowledge, *Zeitschrift fur Mathematische Logik und Grundlagen der Mathematik* **27**, pp. 371 – 374, 1981.