On the Persistence of Knowledge and Ignorance: A Preliminary Report

Fangzhen Lin
Department of Computer Science
University of Toronto
Toronto, Canada M5S 1A4

Yoav Shoham
Department of Computer Science
Stanford University
Stanford, CA 94305

Abstract

We study a particular model of the way in which a single agent's knowledge evolves over time. The two fundamental properties of the model are that knowledge always persists (i.e., agents have perfect memory), and that ignorance persists by default (i.e., we have complete knowledge of exactly how much the agent learns at every time point). Despite its apparent simplicity, this model exhibits quite complex properties.

1 Introduction

Knowledge and time are two important concepts in AI, and have attracted much research in recent years. However, while they have each been heavily studied in isolation, relatively little research has been reported on the interaction between them; this is despite strong intuitions one has about such interactions. In this paper we study a particular model of how an agent's knowledge and ignorance persist or change over time, and show that while some aspects of these phenomena are quite intuitive and easy to formalize, others are not at all.

If nesting of knowledge operators with different time indices is not taken into account, then one need not represent time explicitly, but instead start with a knowledge base describing the agent's initial knowledge state, and define the knowledge base resulting from the agent's gaining or losing some knowledge. This approach was taken by Levesque [1984] and Chandy and Misra [1986].[1] When one allows arbitrary nesting of knowledge operators, one can represent more complex sentences such as "On Wednesday the agent knew that on Monday it didn't know whether $P = NP$." This approach, which involves explicit representation of time, was taken by Halpern and Vardi [1989] and Shoham [1989].[2] In this paper we adopt the latter approach, and treat both knowledge and time explicitly in our language.

Our problem does not lie in coming up with a general framework in which to represent the two; that is easy. Rather, our problem is to identify in that general framework restrictions that constitute a plausible theory of how knowledge evolves over time. We will adopt the standard $S5$ model of knowledge as the idealized theory of the *statics* of knowledge, and will seek a corresponding theory of the *dynamics* of knowledge. As a first step, in this article we consider the single agent case; as we shall see, it will already pose sufficiently challenging problems.

What are the dynamic aspects we wish to capture? Essentially, there are two. The first is that agents have memory. In fact, in this idealized model agents have *perfect* memory. This property is rather easy to capture, although already here there is a subtlety that is easy to miss. The other property we wish to capture is the default persistence of ignorance; if a fact is not known at some time point, and it is not learned later, then it is still not known later. This is similar in flavor to the persistence phenomena associated with the frame problem [McCarthy and Hayes 1969]. Indeed, were it not for the nesting of knowledge operators, capturing the default persistence of ignorance would not be hard. But knowledge operators may be nested, and, as we shall see, this considerably complicates the notion of persistence.

Before we proceed, it is worthwhile to point out an important difference between our setting and the literature on belief revision (cf. [Gardenfors 1988]) or update (cf. [Katsuno and Mendelzon 1991]). The difference hinges on the distinction between knowledge and belief; in the case of belief revision or update, the task is to minimize changes to the current belief set while accommodating new information. In our case, since knowledge is indefeasible, its persistence is uninterrupted: If I know now that $P = NP$, and I have perfect memory, I'll know it forever. (Notice the importance of the temporal indices: If I know now that the house is red today, I'll know forever that it *was* red today.) Subject to this condition, we aim to capture the intuition that the new facts explicitly mentioned are the *only* facts added to the knowledge base.

[1] Levesque investigated a single agent case. Chandy and Misra investigated a special multi-agents case where the knowledge an agent may gain or lose has some special forms.

[2] Halpern and Vardi studied a version of non-forgetting and non-learning in distributed computing. Shoham studied the issues in an informal setting.

2 The standard static theory of knowledge

We assume a discrete and linear time structure, represented by the set of positive natural numbers. We assume a propositional language augmented, for each time point t, by a modal operator K_t. Intuitively, for any formula φ, $K_t\varphi$ means that at the time t the agent knows that φ is true.

Our logic for K_t is the standard $S5$. Thus we have the following axioms:

1. All tautologies.
2. $K_t\varphi \supset \varphi$.
3. $K_t\varphi \supset K_tK_t\varphi$.
4. $\neg K_t\varphi \supset K_t\neg K_t\varphi$.
5. $K_t\varphi \wedge K_t(\varphi \supset \psi) \supset K_t\psi$.

The inference rules are modus ponens (from φ and $\varphi \supset \psi$ infer ψ) and knowledge generalization (from φ infer $K_t\varphi$ for any time point t). Let Φ be a set of formulas, and φ a formula. As conventional, we write $\Phi \vdash \varphi$ if there is a finite subset Φ' of Φ such that $\wedge\Phi' \supset \varphi$ is a theorem of the above axiom system.

The semantics for our logic is the conventional Kripke possible worlds semantics. A *Kripke structure* S is a triple (W, R, π), where W is a nonempty set of possible worlds, R a function such that for each time point t, $R(t)$ is an equivalence relation over W, and for each $w \in W$, π is a truth evaluation function on the primitive propositions. A *Kripke interpretation* is a pair (S, w), where $S = (W, R, \pi)$ is a Kripke structure, and $w \in W$ is the *actual world* of M. It is conventional to define the satisfaction relation "$\models$" between Kripke interpretations and formulas. Particularly, we have

$$(S, w) \models K_t\varphi \text{ if for any } w' \in W \text{ such that } (w, w') \in R(t), (S, w') \models \varphi.$$

A Kripke interpretation M is a *model* of a formula φ if $M \models \varphi$, that is, φ is true in the actual world of M. It is a model of a set of formulas if it is a model of every member of the set. We write $\Phi \models \varphi$ if every model of Φ is also a model of φ.

While intuitively we think of $K_1, K_2, ...$ as the knowledge of the same agent at different times, until we add further constraints between these modalities we might as well think of them as the knowledge of separate agents. Thus the system just presented is simply a propositional logic of knowledge with multiple agents, and results from, e.g., [Halpern and Moses 1985], can be readily adopted. In particular, we have the following soundness and completeness result:

Proposition 2.1 *Let Φ be a set of formulas, and φ a formula. Then $\Phi \vdash \varphi$ iff $\Phi \models \varphi$.*

We now begin to restrict the K_i modalities so that they indeed behave like the evolving knowledge of a single agent. Without loss of generality, in the following we consider the changes the agent's knowledge undergoes from time 1 to time 2, that is, we consider the operators K_1 and K_2. In the next section we capture the absolute persistence of knowledge; in the section following that we address the default persistence of ignorance.

3 Non-forgetting

An agent does not forget anything if its knowledge does not decrease over time. Thus if it knew that p is true (i.e., $K_1 p$ holds), then it will still know that (i.e., $K_2 p$ holds). Similarly, if it knew that it did not know that q is true (i.e., $K_1 \neg K_1 q$ holds), then it will know that it did not know that q is true (i.e., $K_2 \neg K_1 q$ holds). Formally, at 2, the agent *remembers everything it knew at* 1 if the following axiom holds for any formula φ:

$$K_1 \varphi \supset K_2 \varphi. \tag{1}$$

We use $Mem_{1,2}$ to denote the set of the axioms of the above form.

Semantically, the larger the agent's possible worlds are, the more ignorant the agent is. Formally, the agent remembers at 2 everything it knew at 1 in the Kripke interpretation $((W, R, \pi), w)$ if the following condition holds:

For any $w' \in W$, if $(w, w') \in R(2)$, then $(w, w') \in R(1)$.

It is easy to see that we can capture in a similar way notions such as: at t_2, the agent remembers everything it knew at t_1, or everything it knew before t_1, or everything it knew from t_0 to t_1. In the rest of this section, when there is no possibility of confusion, we shall call the agent *non-forgetting* if it remembers at 2 everything it knew at 1. We call a Kripke interpretation M non-forgetting if the agent modeled by it is non-forgetting.

The semantics completely captures (1) in the following sense:

Theorem 1 *For any formula φ, $Mem_{1,2} \vdash \varphi$ iff for any non-forgetting Kripke interpretation M, $M \models \varphi$.*

Proof:[3] First of all, if M is a non-forgetting one, then $M \models \Phi$. Thus if $\Phi \vdash \varphi$, then for all non-forgetting M, $M \models \varphi$.

[3] Adam Grove pointed out that the proof of this theorem was also implicitly implied in [Halpern and Moses 1985] in the proof of the completeness theorem there.

For the reverse, we prove that for any Kripke model M of Φ, there is a non-forgetting M' such that for any φ, $M \models \varphi$ iff $M' \models \varphi$.

Let $M \models \Phi$. We show that there is a non-forgetting M' such that for any φ, $M \models \varphi$ iff $M' \models \varphi$. Suppose $M = (S, w)$, where $S = (W, R, \pi)$. Define $M' = (S', w) = ((W, R', \pi), w)$ as follows. For any $w_1, w_2 \in W$, if $t \neq 1$, then $(w_1, w_2) \in R'(t)$ iff $(w_1, w_2) \in R(t)$. $R(1)$ is the smallest equivalent relation that satisfies the following conditions:

1. If $(w_1, w_2) \in R(1)$, then $(w_1, w_2) \in R'(1)$; and
2. If $(w, w_1) \in R(2)$, then $(w, w_1) \in R'(1)$.

In other words, M' is obtained from M by merging the accessibility relation of M at the actual word w at 2 into that at 1, and thus making M' a non-forgetting interpretation. Surprisingly it may seem, M and M' are equivalent: we show by the induction on the complexity of the formula φ that for any $w_1 \in W$, $(S, w_1) \models \varphi$ iff $(S', w_1) \models \varphi$. This is true if φ is a primitive propositions. For the inductive step, it is trivial if $\varphi = \neg\phi$, $A = \phi_1 \wedge \phi_2$, or $\varphi = K_t\phi$ for some $t \neq 1$. Let $\varphi = K_1\phi$. Suppose $(S', w_1) \models K_1\phi$. Then for any $w_2 \in W$ such that $(w_1, w_2) \in R'(1)$, $(S', w_2) \models \phi$, which implies $(S, w_2) \models \phi$ by the inductive assumption. By our definition, if (w_1, w_2) is in $R(1)$, then it is also in $R'(1)$. Therefore we have $(S, w_1) \models K_1\phi$. Now suppose $(S, w_1) \models K_1\phi$. If $(w, w_1) \notin R'(1)$, then for any w_2, $(w_1, w_2) \in R'(1)$ iff $(w_1, w_2) \in R(1)$. Therefore if $(w, w_1) \notin R'(1)$, then it easily follows from the inductive assumption that $(S', w_1) \models K_1\phi$. Now suppose that $(w, w_1) \in R'(1)$. We show that $(S, w) \models K_1\phi$. By our definition of $R'(1)$, there are two cases:

1. $(w, w_1) \in R(1)$: trivial.
2. There is a w_2 such that $(w, w_2) \in R(2)$ and $(w_2, w_1) \in R(1)$. In this case, firstly, we have $(S, w_2) \models K_1\phi$. Now if $(S, w) \models \neg K_1\phi$, then since $M = (S, w)$ satisfies Φ, we have that $(S, w) \models K_2\neg K_1\phi$. Thus $(S, w_2) \models \neg K_1\phi$, a contradiction.

Therefore, in either case, we have that $(S, w) \models K_1\phi$. We now show that this implies that for any w_2, if $(w_1, w_2) \in R'(1)$, then $(S, w_2) \models \phi$. Again we have two cases:

1. $(w, w_2) \in R(1)$: trivial.
2. There is a $w_3 \in W$ such that $(w, w_3) \in R(2)$ and $(w_3, w_2) \in R(1)$. In this case we have

   $$(S, w) \models K_1\phi \Rightarrow (S, w) \models K_2K_1\phi \Rightarrow (S, w_3) \models K_1\phi \Rightarrow (S, w_2) \models \phi$$

   Notice that the first "$\Rightarrow$" follows from the fact that M is a model of Φ.

Therefore by the inductive assumption we have that for any w_2, if $(w_1, w_2) \in R'(1)$, then $(S', w_2) \models \phi$. Thus $(S', w_1) \models K_1\phi$. ■

There is an interesting property about non-forgetting in S5. It says that in S5, if an agent does not forget, then it knows that:

Theorem 2 *For any formula φ we have*

$$Mem_{1,2} \vdash K_2(K_1\varphi \supset K_2\varphi).$$

Notice that a related property, which says that if an agent does not forget anything then it knows that it will not forget anything, is not true in general:

$$Mem_{1,2} \not\vdash K_1(K_1\varphi \supset K_2\varphi).$$

Semantically, if an agent knew that it will not forget, then it is captured by a Kripke interpretation $M = ((W, R, \pi), w)$ that satisfies the following condition:

For any $w_1, w_2 \in W$, if $(w, w_1) \in R(1)$ and $(w_1, w_2) \in R(2)$, then $(w_1, w_2) \in R(1)$.

As previously, we can prove that the set of axioms

$$K_1(K_1\varphi \supset K_2\varphi) \tag{2}$$

is sound and complete for the this class of Kripke interpretations. (Intuitively, axiom schema (1) corresponds to non-forgetting *in the actual world*, while (2) corresponds to non-forgetting *in every world* that is accessible from the actual world.)

In the following, we shall consider only non-forgetting as captured by (1). That is, we do not assume that the agent will always know that it will not forget.

4 Minimal learning

Non-forgetting captures the absolute persistence of knowledge. The formal dual of non-forgetting is non-learning, which captures the absolute persistence of an agent's ignorance. It can be captured by following axiom schema:

$$\neg K_1\varphi \supset \neg K_2\varphi. \tag{3}$$

We do not continue this line of development, however. The main reason is that, given the non-forgetting assumption, introducing non-learning would lead to a trivial system.[4]

[4] One could imagine an application which calls for non-learning but not for non-forgetting, say a model of senility, but this is not the model we investigate in this paper.

Instead, we will pursue a weaker form in which ignorance persists – default persistence. Specifically, we will consider the following general problem: given the agent's knowledge state at 1, and the assumption that between 1 and 2 the agent learns *only* that the proposition A is true, what is the agent's knowledge state at 2?

Given a formula φ, either $K_1\varphi$ is true or $\neg K_1\varphi$ is true. If $K_1\varphi$ holds, then from non-forgetting we get that $K_2\varphi$ holds, no matter what else was known at time 1 or learned subsequently. On the other hand, if $\neg K_1\varphi$ holds, then whether $\neg K_2\varphi$ still holds depends not only on how much the agent learns but also on the agent's other knowledge and ignorance. For example, suppose $K_1(p \supset q)$, $\neg K_1 p$, and $\neg K_1 q$ hold. If the agent is non-forgetting, and learns at 2 that p is true, then we will have that $K_2(p \supset q)$ and $K_2 p$ are true, and thus $K_2 q$ is true as well. This is the reason why we need to consider the agent's whole knowledge state at 1 for the concept of minimal learning to make sense, and as we shall see, the main reason why it is hard to capture minimal learning.

In this section, we shall provide a semantic characterization of minimal learning under the assumption of non-forgetting.

If we had not represented time explicitly, this would have been easy for the single agent case. Specifically, the agent's knowledge state at a time point would be represented by an ordinary propositional theory, and learning a new fact would amount to simply adding this new fact into the propositional theory, and take the logical closure of the resulting theory.

However, this procedure does not work for the framework we have presented. First of all, the agent's knowledge state at a time point can not be represented by an ordinary propositional theory. For example, saying that the agent's knowledge state at 3 is the logical closure of $\{p, q\}$ leaves it open whether it knows (at 3) that it knew at 2 that p is true, or that it will continue to know at 4 that p and q are true.

In our framework, the notion of knowledge states can be defined in terms of Kripke interpretations. Since we have assumed non-forgetting, in the following, unless otherwise stated, all Kripke interpretations will be assumed to be non-forgetting.

Definition 4.1 *If M is a Kripke interpretation, and t a time point, then we define $M(t)$ to be the following set:*

$$M(t) = \{\varphi \mid M \models K_t\varphi\}.$$

A set of sentences T is called a knowledge state at the time t if there is a Kripke interpretation M such that $T = M(t)$.

Notice that in terms of the new notation, if M is a Kripke interpretation, then the agent in M remembers at 2 everything it knew at 1 iff $M(1) \subseteq M(2)$. Now suppose that the agent learns only that A is true at 2. Motivated by the single agent without explicit time case, we would like to require that $M(2)$ be the logical closure of

$M(1) \cup \{K_2 A\}$. Unfortunately, this does not work. Most often, this logical closure is not even a knowledge state. For example, for some primitive proposition p, it is possible that neither $K_2 p$ nor $\neg K_2 p$ is a logical consequence of $M(1) \cup \{K_2 A\}$. But, of course, a knowledge state at 2 has to contain one of them.

Under the assumption that the agent learns *only* that A is true at 2, for any primitive proposition p, if neither $K_2 p$ nor $\neg K_2 p$ is a logical consequence of $M(1) \cup \{K_2 A\}$, then it should be the case that $\neg K_2 p$, for the agent's ignorance should persist as much as possible. Generalizing this reasoning, we might want to say that if a formula φ is not a logical consequence of $M(1) \cup \{K_2 A\}$, then $\neg K_2 \varphi$ should be in the new knowledge state. However, this will simply result in inconsistency. For example, it is possible that neither p nor $\neg K_2 p$ is a logical consequence of $M(1) \cup \{K_2 A\}$. But adding $\neg K_2 p$ and $\neg K_2 \neg K_2 p$ simultaneously would result in an inconsistency. In this case, our intuition is clear that we should add $\neg K_2 q$, not $\neg K_2 \neg K_2 p$. But the problem is how to have a uniform procedure that works for every case. Our intuition is that some kind of hierarchies based on degrees of nested modalities should work. But an ad hoc and complicated procedure would look suspicious. It would be best if we can "derive" this procedure in some semantics terms. To this we now turn.

As we said, if the agent learns only that A is true at 2, then $K_2 A$ should hold. Furthermore, the agent should be as ignorant as possible. Semantically, let $M = ((W, R, \pi), w)$ be a Kripke interpretation. One way to make M more ignorant at 2 is to expand the set of possible worlds W into W', and keep R the same except making the 2-accessible class at the actual world w larger, where for each t, the *t-accessible class* of M at w is the following set:

$$\{w' \mid (w, w') \in R(t)\}.$$

This motivates the following definition:

Definition 4.2 *Let $M_i = ((W_i, R_i, \pi_i), w_i)$, $i = 1, 2$, be two Kripke interpretations. We say that M_1 is* strictly as ignorant as *M_2 at 2, written $M_1 \leq_2 M_2$, if the following conditions hold:*

1. $W_2 \subseteq W_1$;
2. $w_1 = w_2$;
3. *For any* $w \in W_2$, $\pi_2(w) = \pi_1(w)$;
4. *For any* $w \in W_2$, *if* $(w_2, w) \in R_2(2)$, *then* $(w_1, w) \in R_1(2)$.

5. *For any time point t, and any $w \in W_2$, if $t \neq 2$ or $t = 2$ but $(w_2, w) \notin R_2(2)$, then for any $w' \in W_1$, $(w, w') \in R_1(t)$ iff $(w, w') \in R_2(t)$.*[5]

It turns out that this special way of making M more ignorant is quite enough if we consider two Kripke models to be equivalent whenever they determine the same knowledge states. Let us say that M_1 and M_2 are *equivalent at* 2, written $M_1 =_2 M_2$, if $M_1(2) = M_2(2)$. Formally we have the following definition:

Definition 4.3 *Let M_1 and M_2 be two Kripke interpretations. We say that M_1 is* as ignorant as *M_2 at 2, written $M_1 \prec_2 M_2$, if there are two Kripke interpretations M_3 and M_4 such that $M_1 =_2 M_3$, $M_2 =_2 M_4$, and $M_3 \leq M_4$. We say that M_1 is* more ignorant *than M_2 at 2, written $M_1 \preceq_2 M_2$, if $M_1 \prec_2 M_2$ and $M_1 \neq_2 M_2$.*

The following non-trivial property about $\prec_2$ is important, and reassuring:

Proposition 4.1 *If $M_1 \prec_2 M_2$ and $M_2 \prec_2 M_1$, then $M_1 =_2 M_2$.*

Proof: We first prove the following lemma:

Lemma 4.1 *Let $M_i = ((W_i, R_i, \pi_i), w_i)$, $i = 1, 2$, be two Kripke worlds, $M_1 \leq_2 M_2$, and A be a formula. If for any subformula (including A itself) of the form $K_2\varphi$ in A the condition that $M_1 \models K_2\varphi$ iff $M_2 \models K_2\varphi$ holds, then for any possible world $w \in W_2$, $((W_1, R_1, \pi_1), w) \models A$ iff $((W_2, R_2, \pi_2), w) \models A$.*

We prove by induction on the complexity of A. If A is a primitive proposition, then the result holds because $\pi_1(w) = \pi_2(w)$. The inductive step for $A = \neg B$ and $A = B_1 \vee B_2$ are trivial. Suppose $A = K_tB$, and $t \neq 2$. Then

$((W_1, R_1, \pi_1), w) \models K_2B$

iff for any $(w', w) \in R_1(t)$, $((W_1, R_1, \pi_1), w') \models B$

iff $w' \in W_2$ and (by inductive assumption) $((W_2, R_2, \pi_2), w') \models B$

iff $((W_2, R_2, \pi_2), w) \models K_tB$.

Suppose $A = K_2B$. There are two cases. The case where $(w, w_2) \notin R_2(2)$ is similar to the case for $A = K_tB$. If $(w, w_2) \in R_2(2)$, then $(w, w_1) \in R_1(2)$, and

[5]This is our formalization that R_1 and R_2 otherwise agree with each other. For example, it says that if $w \in W_2$ is not accessible from the actual world in M_2 at time 2, then for any possible world w' in M_1, if w' is accessible from w in M_1 at time 2, then w' has to be a possible world of M_2, and accessible from w in M_2 at time 2 as well.

$((W_2, R_2, \pi_2), w) \models K_2 B$

iff $((W_2, R_2, \pi_2), w_2) \models K_2 B$

iff $((W_1, R_1, \pi_1), w_1) \models K_2 B$

iff $((W_1, R_1, \pi_1), w) \models K_2 B$.

We now prove the proposition. We prove by induction on the degree of the nesting of $K_2\phi$ in φ that $M_1 \models K_2\varphi$ iff $M_2 \models K_2\varphi$.

By the assumption, there are four Kripke worlds M_i, $3 \leq i \leq 6$, such that $M_1 =_2 M_3 =_2 M_5$, $M_2 =_2 M_4 =_2 M_6$, $M_3 \leq_2 M_4$, and $M_6 \leq_2 M_5$.

Let φ be a formula that does not contain subformulas of the form $K_2\phi$. Then

$$M_1 \models K_2\varphi \Rightarrow M_3 \models K_2\varphi \Rightarrow M_4 \models K_2\varphi \Rightarrow M_2 \models K_2\varphi$$

Similarly, if $M_2 \models K_2\varphi$ then $M_1 \models K_2\varphi$.

Inductively, suppose that for any formula φ with the degree of nesting of $K_2\phi$ in it less than n, we have $M_1 \models K_t\varphi$ iff $M_2 \models K_2\varphi$. Let φ be a formula with the degree of nesting of $K_2\phi$ in it being n. We prove that $M_1 \models K_2\varphi$ iff $M_2 \models K_2\varphi$. By the inductive assumption, for any subformula of the form $K_2\phi$ in φ, $M_1 \models K_2\phi$ iff $M_2 \models K_2\phi$. Thus using the lemma it is easy to see that if $M_3 \models K_2\varphi$, then $M_4 \models K_2\varphi$; and if $M_6 \models K_2\varphi$, then $M_5 \models K_2\varphi$. ∎

Let T_1 be the agent's knowledge state at 1. A Kripke interpretation M models the assumption that the agent learns only that A is true at 2, written $M \models T_1 + A$, if $M(1) = T_1$, $M \models K_2 A$, and M is as ignorant as possible at 2:

Definition 4.4 *$M \models T_1 + A$ if $M(1) = T_1$, $M \models K_2 A$, and there is no M' that satisfies these two conditions and $M' \preceq_2 M$.*

We now have a semantic definition of minimal learning. The definition is fairly elaborated since arbitrary nesting of knowledge operators with different time indices are allowed. We now justify this definition by applying it to a class of situations about which we have clear intuitions, and show that our semantics gives intuitive results.

5 A class of situations: semantics

Recall that our main goal is to account for how an agent's knowledge evolves over time. The assumptions we have made are that the agent never forgets, and that we know exactly how much the agent learns at any time point. The simplest such situation is that initially the agent knew nothing, then at any later point, the agent learns only the truth value of a primitive proposition. However, in order to formalize this situation, we need to specify the agent's initial knowledge state where it knows nothing. This is an instance of "know only," a well-known difficult problem [Halpern and Moses 1984, Parikh 1984, 1991, Halpern 1987, and Levesque 1987].

It is easy to see that "know only" is the most convenient way to specify a knowledge state. Thus our problem, paraphrased in English, is "Given that at time 1 the agent *knew only* φ, and that in between times 1 and 2 it *learns only* A, what does it know at time 2?" It turns out that "know only" can be easily formalized using the notions we have introduced in last section.

We first notice that our definitions for $\prec_2$, $\preceq_2$, and $=_2$ can be extended straightforwardly to $\prec_t$ (as ignorant as at t), $\preceq_t$ (more ignorant at t), and $=_t$ (equivalent at t), for any time point.

Definition 5.1 *Let S be a set of sentences. M is a model of the agent knowing only that every member of S is true at t, written $M \models\!\!\approx_t S$, if M is a model of $K_t S$, and there is no other model M' of $K_t S$ such that $M' \preceq_t M$, where a Kripke interpretation M is a model of $K_t S$ iff for any $A \in S$, M is a model of $K_t A$.*

It turns out that we can define minimal learning in terms of "know only:"

Proposition 5.1 *For any Kripke interpretation M, any knowledge state at 1, T_1, and any formula A, $M \models T_1 + A$ iff $M \models\!\!\approx_2 S$, where $S = \{K_1\varphi \mid \varphi \in T_1\} \cup \{A\}$.*

We remark that so far we have found this proposition interesting only conceptually. In practice, knowing only S is interesting only when S is finite.

Our main goal in this section is to show that under certain conditions, there is a unique Kripke interpretation that captures the situations posed at the beginning of this section. First, we show that there is a unique knowledge state corresponding to "knows nothing (except tautologies)" at the starting time point. To this end, we show some properties about $\models\!\!\approx_t$. First, we show how to obtain a more ignorant model from a sequence of models. Our construction is similar to that in [Parikh 1984, 1991]. However our definition does not depend on "canonical models."

Let I be a set of natural numbers, and for each $i \in I$, $M_i = ((W_i, R_i, \pi_i), w_i)$ be a Kripke interpretation. Suppose that for any $i \neq j$ in I, $W_i \cap W_j = \emptyset$. We define $\sum_{i \in I}^{t} M_i = M_{i_1} +_t M_{i_2} +_t \ldots$ to be the Kripke interpretation $((W, R, \pi), w)$, where

1. $w = w_i$ for the least natural number $i \in I$.

2. $W = \cup_{i \in I} W_i$,

3. π is the function such that for any possible world $w \in W$, if $w \in W_i$ for some (unique) $i \in I$, then $\pi(w) = \pi_i(w)$, and

4. R is the smallest equivalence relation that satisfies the following two conditions:

   (a) $R_i(t') \subseteq R(t')$ for any time point t', $i \in I$.

   (b) $(w, w_i) \in R(t)$ for each $i \in I$.

It is easy to see that for any $k \in I$, $((W, R, \pi), w_k) \leq_t M_k$. Thus $\sum_{i \in I}^{t} M_i \prec_t M_k$ for any $k \in I$. Similar to Lemma 4.1, we have the following result:

Lemma 5.1 *Let F be any formula such that for any subformula (including F itself) of the form $K_t F_1$ in F, $M_i \models K_t F_1$ iff $M_j \models K_t F_1$, for any i, j in I. Let $\sum_{i \in I}^{t} M_i$ be $((W, R, \pi), w)$, and M_i be $((W_i, R_i, \pi_i), w_i)$ for any $i \in I$. Then*

- *For any $i \in I$, $((W, R, \pi), w_i) \models F$ iff $M_i \models F$.*
- *$\sum_{i \in I}^{t} M_i \models K_t F$ iff $M_i \models K_t F$ for every $i \in I$.*

Proof: The proof is similar to that of Lemma 4.1. ■

From this lemma, we can prove the following theorem:

Theorem 3 *Let A be a formula such that for any subformula of the form $K_t B$ in A, either $A \models K_t B$ or $A \models \neg K_t B$. If $K_t A$ is consistent, then there is a unique M such that $M \models\!\approx_t \{A\}$ in the sense that for any M', if $M' \models\!\approx_t \{A\}$, then $M =_t M'$.*

Proof: Suppose that $K_t A$ is consistent. Let $M_i = ((W_i, R_i, \pi_i), w_i)$, $i = 1, 2, ...$, be a sequence of models of $K_t A$ such that for any $i \neq j$, $W_i \cap W_j = \emptyset$, and for any model M of $K_t A$, there is a i such that $M =_t M_i$. Therefore, for any model M of $K_t A$, $\sum_{i \geq 1}^{t} M_i \prec_t M$. Thus for the proof of the theorem, we only have to show that $\sum_{i \geq 1}^{t} M_i \models K_t A$. But this follows from the above lemma. ■

Thus according to the theorem, there is a unique M such that $M \models\!\approx_1 \emptyset$. Let $T_\emptyset = M(1)$. Intuitively, $T_\emptyset$ captures the situation where the agent knows nothing (except tautologies) at 1. Let p_2 be a primitive proposition, then $T_\emptyset + p_2$ describes the situation where the agent goes from knowing nothing at 1 to knowing only that p_2 is true at 2. More generally, $T_\emptyset + p_2 + ... + p_n + ...$ will describe the situation where initially at 1, the agent knows nothing, then at any time point $t > 1$, the agent learns only that p_t is true. Formally, it is defined as follows.

Definition 5.2 *Let T_1 be a knowledge state at 1, and $A_2, ..., A_n$ be formulas. We say that an interpretation M models the situation where initially the agent's knowledge state is T_1, and at any later point $t \leq n$, the agent learns only A_t, written*

$$M \models T_1 + A_2 + ... + A_n,$$

if the following properties are satisfied:

1. *$M \models T_1 + A_2 + ... + A_{n-1}$.*
2. *$M \models K_n A_n$.*
3. *There is no M' such that $M'(n-1) = M(n-1)$, $M' \models K_n A_n$, and $M' \preceq_n M$.*

Notice that since $M \models T_1$ iff $M(1) = T_1$, Definition 5.2 generalizes Definition 4.4. The following theorem semantically captures the situation posed at the beginning of the section.

Theorem 4 *Let $\{K_i A_i \mid i = 2, 3, ...\}$ be a consistent set of formulas such that for any $i \geq 2$, A_i does not contain any modal operators. Let $T_\emptyset$ be the knowledge state at 1 where the agent knows nothing except tautologies. Then for any $n \geq 2$, there is a unique M such that $M \models T_\emptyset + A_2 + ... + A_n$ in the sense that if $M' \models T_\emptyset + A_2 + ... + A_n$, then $M(i) = M'(i)$ for $1 \leq i \leq n$.*

Proof: We prove by induction on n. We first prove the theorem for $n = 2$.

Let $M_1, M_2, ...$ be a sequence of interpretations such that

1. For each i, M_i is non-forgetting, $M_i(1) = T_1$, and $M_i \models K_2 A_2$.
2. For any M satisfying the above condition, there is a i such that $M =_2 M_i$.

We first show that such a sequence exists, i.e., there is a non-forgetting model M such that $M(1) = T_1$, and $M \models K_2 A_2$. By Theorem 3, there is a model M' such that $M'(1) = T_1$. Since $K_1 A_1 \wedge K_2 A_2$ is consistent, and none of A_1 and A_2 contain modal operators, it is easy to see that there is a non-forgetting model M'' such that $M'' \models K_1 A_1 \wedge K_2 A_2$. It is easy to see that $M = M'' +_1 M'$ satisfies the required conditions.

Let $M_i = ((W_i, R_i, \pi_i), w_i)$. Without the loss of generality, we suppose that for any $i \neq j$, $W_i \cap W_j = \emptyset$. For each i, let $M'_i = ((W'_i, R'_i, \pi'_i), w'_i)$ be a rename of M_i such that for any i, j, $W_i \cap W'_j = \emptyset$, and for any $i \neq j$, $W'_i \cap W'_j = \emptyset$.

Now let $M = \sum^2_{i \geq 1}(M_i +_1 M'_i)$. We claim that (A) for each i, $M \prec_2 M_i$; (B) $M \models K_2 A_2$; (C) $M(1) = T_1$; and (D) M satisfies the non-forgetting axiom schema (1).

For each i, let $M_{ii} = M_i +_1 M'_i$. Then it is easy to see that M_{ii} is non-forgetting for each i. By Lemma 5.1 we notice that $M_{ii} = M_i$ in the sense that for any formula φ, $M_{ii} \models \varphi$ iff $M_i \models \varphi$. Thus (A) follows from $M \prec_2 M_{ii}$. (B) is also an easy consequence of Lemma 5.1. Suppose $M = ((W, R, \pi), w_1) = (K, w_1)$. Then $M =_1 (K, w'_1)$. But $(K, w'_1) \leq_1 M'_1$, and $(K, w'_1) \models K_1 A_1$ according to Lemma 5.1. Therefore $M(1) = (K, w'_1)(1) = T_1$. Thus (C) is proved. Similarly, we can prove that for each i,

$$(K, w_i)(1) = (K, w'_i)(1) = T_1 \tag{4}$$

For the proof of (D), define $M' = ((W, R', \pi), w_1) = (K', w_1)$, where R' is the smallest equivalence relation such that $R \subseteq R'$, and for each i, $(w_1, w_i) \in R'(1)$. Since each M_{ii} is non-forgetting, therefore M' is also non-forgetting. We now show that $M = M'$ in the sense that for any φ, $M \models \varphi$ iff $M' \models \varphi$. For any φ, we prove by using induction on the complexity of φ that for any $w \in W$, $(K, w) \models \varphi$ iff $(K', w) \models \varphi$. The base case that φ is a primitive proposition is trivial. Inductively, the cases for $\varphi = \neg\phi$, $\varphi = \varphi_1 \vee \varphi_2$, and $\varphi = K_t\phi$ for $t \neq 1$ are easy to see. Now suppose that $\varphi = K_1\phi$. If $(w, w_1) \notin R'(1)$, then $(K, w) \models K_1\phi$ iff $(K, w') \models \phi$ for every $(w', w) \in R(1)$ iff $(K', w') \models \phi$ for every $(w', w) \in R'(1)$ iff $(K', w) \models K_1\phi$. Now suppose $(w, w_1) \in R'(1)$. We show that $(K, w_1) \models K_1\phi$ iff $(K', w_1) \models K_1\phi$. Other cases are similar. $(K', w_1) \models K_1\phi$ iff $(K', w) \models \phi$ for every $(w, w_1) \in R'(1)$. But $(w, w_1) \in R'(1)$ iff there is a i such that $(w, w_i) \in R(1)$. Therefore by the inductive assumption, $(K', w_1) \models K_1\phi$ iff $(K, w_i) \models K_1\phi$ for every i. Thus by (4), $(K', w_1) \models K_1\phi$ iff $(K, w_1) \models K_1\phi$.

Therefore we have proved (A) to (D). Thus $M \models T_1 + A_2$, and for any M', if $M \models T_1 + p$, then $M =_2 M'$.

Suppose we have proved the theorem for n, we show that it is also true for $n+1$. Let $M \models T_1 + A_2 + ... + A_n$, and $T_n = M(n)$. Similar to the case of $n = 2$, let $M_1, M_2, ...$ be a sequence of interpretations such that

1. For each i, M_i is non-forgetting at n, $M_i(n) = T_n$, and $M_i \models K_{n+1}A_{n+1}$.

2. For any M satisfying the above condition, there is a i such that $M =_{n+1} M_i$.

Let $M_i = ((W_i, R_i, \pi_i), w_i)$. Without the loss of generality, we suppose that for any $i \neq j$, $W_i \cap W_j = \emptyset$. For each i, let $M'_i = ((W'_i, R'_i, \pi'_i), w'_i)$ be a rename of M_i such that for any i, j, $W_i \cap W'_j = \emptyset$, and for any $i \neq j$, $W'_i \cap W'_j = \emptyset$.

Now let $M = \sum_{i \geq 1}^{2}(M_i +_1 M'_i)$. Again we claim that (A) for each i, $M \prec_{n+1} M_i$; (B) $M \models K_{n+1}A_{n+1}$; (C) $M(n) = T_n$; and (D) M satisfies the non-forgetting axiom schema at n, which is the schema (1) with the time point 1 being replaced by n, and the time point 2 by $n+1$.

The proof is similar to the case for 2 by noticing that for model M', if $M'(n) = T_n$, then for any $t \leq n$, $M'(t) = T_t$. ■

Theorem 3 and 4 are the main technical results of the paper. They showed that for the situations posed at the beginning of this section, our definitions of "know only" and "learn only" correspond to our intuitions.

6 A class of situations: decision procedures

As we said in section 4, we would like to somehow derive a procedure for deciding whether a proposition is known after the agent learns only something. Fortunately, as corollaries to the proofs of Theorem 3 and 4, we have two inductive decision procedures for the class of the theories studied in last section.

The proof of Theorem 3 shows the following decision procedure for "know only," which to our knowledge is the first one about "know only" in a language with multiple modal operators:

Proposition 6.1 *Let A be a formula satisfying the condition in Theorem 3. Let $K_t A$ be consistent, and $M \approx_t \{A\}$. For any formula φ, $\varphi \in M(t)$ iff $K_t A \wedge sub_t(\varphi) \models \varphi$, where $sub_t(\varphi)$ is the conjunction such that for any subformula of φ (including φ) having the form $K_t \phi$, if $\phi \in M(t)$, then $K_t \phi$ is a conjunct, otherwise $\neg K_t \phi$ is a conjunct.*

Thus $\neg K_1 p \in T_\emptyset$ since $sub_1(\neg K_1 p)$ is $\neg K_1 p$, and $\neg K_1 p \models K_1 \neg K_1 p$. Similarly $\neg K_1 K_2 \neg K_1 p \in T_\emptyset$ since $\neg K_1 p \not\models K_1 K_2 \neg K_1 p$.

The proof of Theorem 4 also gives the following result:

Proposition 6.2 *Let $K_2 A_2, ..., K_n A_n$ be as those in Theorem 4, and $M \models T_\emptyset + A_2 + ... + A_n$. For any formula φ, $\varphi \in M(n)$ iff*

$$K_n(M(n-1)) \cup \{K_n ... K_2 A_2 \wedge ... \wedge K_n A_n \wedge SUB_n(\varphi)\} \vdash K_n \varphi, \quad (5)$$

where $K_n(M(n-1)) = \{K_n \phi \mid \phi \in M(n-1)\}$, and $SUB_n(\varphi)$ is the conjunction such that for any subformula of φ of the form $K_t \phi$, $1 \leq t \leq n$, if $\phi \in M(n)$, then $K_t \phi$ is a conjunct of $SUB_n(\varphi)$, otherwise $\neg K_t \phi$ is a conjunct of $SUB_n(\varphi)$.

However, this procedure is hard to use since $K_n(M(n-1))$ is an infinite set. We conjecture that (5) holds iff

$$K_n ... K_2 A_2 \wedge ... \wedge K_n A_n \wedge SUB_n(\varphi) \vdash K_n \varphi$$

holds. In general, this is not true if A_i, $2 \leq i \leq n$, contains modal operators. It seems plausible to us if A_i's do not contain modal operators, which is the case in the proposition.

7 Restricting our language?

As we have seen throughout the paper, most of our difficulties are with formulas with nested modalities. For example, we see that for any formula A that does not include any modal operators, and is not a tautology, we have $\neg K_1 A \in T_\emptyset$. By S5 axioms, this in turn implies that $K_1\neg K_1 A \in T_\emptyset$, $K_1\neg K_2 K_1 A \in T_\emptyset$, $K_1\neg K_3\neg K_1\neg K_2 K_1 A \in T_\emptyset$, etcetera. However, it is doubtful that formulas such as $K_1\neg K_3\neg K_1\neg K_2 K_1 A$ would be needed in practice. It makes sense to focus our attention on formulas of simpler forms. For example, if we focus only on formulas without nested modal operators, then we have the following result:

Proposition 7.1 *Let $K_2 A_2, ..., K_n A_n$ be as those in Theorem 4, and $M \models T_\emptyset + A_2 + ... + A_n$. Let $T_n = M(n)$. For any formula A not containing modal operators, $K_t A \in T_n$ iff either $t > n$ and A is a tautology, or $t \leq n$, and $A_2 \wedge ... \wedge A_t \models A$.*

Notice that this proposition essentially captures minimal learning for the single agent case without explicit times. However, this time it is "justified" under a more general definition. It could also be seen as an evidence for the correctness of our more general definition.

We can also restrict our language to the set of formulas that are constructed from primitive propositions and formulas of the form $K_t A$ by using logical connectives $\neg$ and $\supset$, where A does not contain modal operators. This is still a considerable restriction of the full modal language. However, we'll be able to express in the language facts such as:

$$K_5 p \vee K_5 \neg p$$

(the agent will know the truth value of p at 5),

$$\neg K_4 p \supset (K_7 q \vee K_7 \neg q)$$

(if at 4 the agent does not know that p is true, then it will know the truth value of q at 7), and

$$\neg K_9 p \wedge \neg K_9 \neg p$$

(the agent will never know the truth value of p until time 9). We hope we shall have some computationally positive results to report about minimal learning in this restricted language in the near future.

8 Concluding remarks

We have studied a particular model of the way in which a single agent's knowledge evolves over time. The two fundamental properties of the model are that knowledge always persists (i.e., agents have perfect memory), and that ignorance persists by default (i.e., we have complete knowledge of exactly how much the agent learns at every time point). Despite the apparent simplicity of this model, it turns out to exhibit quite complex properties. For some people this may sound surprising, since the model is quite similar to other, well understood frameworks, involving various combinations of time, knowledge and nonmonotonicity. The complexities become clear once it is realized that nested knowledge operators interact with each other in a non-trivial way. We have showed that, for a class of situations about which we have clear intuitions, our formalization gives intuitive results. For more complicated cases, we are forced to restrict the language in order to retain intuition.

Acknowledgements

We would like to thank Adam Grove, Joe Halpern, Hector Levesque, Becky Thomas, and the other members of the 1991 Stanford Nobotics group for their ideas, comments, and suggestions. Most of the research was done while the first author was with the Department of Computer Science at Stanford University, where both authors' research was supported in part by a grant from the Air Force Office of Scientific Research. The work of the first author was also supported in part by the Government of Canada Institute for Robotics and Intelligent Systems.

References

[1] Chandy, K. M. and J. Misra (1986), How processes learn, *Distributed Computing* 1:1, 1986.

[2] Halpern, J. (1987), Review of Parikh [1984], *Mathematical Review*, 87h:68110.

[3] Halpern, J. Y. and Y. O. Moses (1984), Towards a theory of knowledge and ignorance: preliminary report, IBM Technical Report RJ 4448 (48136), 1984.

[4] Halpern, J. Y. and Y. O. Moses (1985), A guide to the modal logics of knowledge and beliefs: preliminary draft, IBM Technical Report RJ 4753 (50521), 1985.

[5] Halpern, J. Y. and M. Y. Vardi (1989), The complexity of reasoning about knowledge and time, I: lower bounds, IBM Technical Report RJ 5764 (58103), 1989.

[6] Gärdenfors, P. (1987), *Knowledge in Flux: modeling the dynamics of epistemic states*, MIT Press, Cambridge, MA, 1987.

[7] Katsuno, H and A. O. Mendelzon (1991), On the difference between updating a knowledge base and revising it, in *Proc. Second Conference on Knowledge Representation and Reasoning*, Boston, MA, 1991.

[8] Levesque, H. (1984), Foundations of a functional approach to knowledge representation, *Artificial Intelligence*, 23(2), 1984.

[9] Levesque, H. (1987), All I know: an abridged report, *Conference Proceedings of AAAI–87*, 426–431.

[10] Levesque, H. (1990), All I know: A study in autoepistemic logic, *Artificial Intelligence* 42 (1990) 263 – 309.

[11] McCarthy, J. and P. Hayes (1969), Some philosophical problems from the standpoint of artificial intelligence. In *Machine Intelligence* 4, Meltzer, B and Michie, D. (eds), Edinburgh University Press.

[12] Parikh, R. (1984), Logics of knowledge, games and dynamic logic, FST-TCS 1984, Springer LNCS 181.

[13] Parikh, R. (1991), Monotonic and non-monotonic logics of knowledge, Manuscript, 1991.

[14] Shoham, Y. (1989), Time for action, in *Proceedings of IJCAI-89*.