Stanford Artificial Intelligence Laboratory
Memo AIM-3 12

April 1978

Computer Science Department
Report No. STAN-CS-78-667

ON THE MODEL THEORY OF KNOWLEDGE

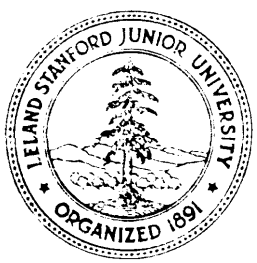
by

John McCarthy, Masahiko Sato, Takeshi Hayashi, Shigeru Igarashi

Research sponsored by

Advanced Research Projects Agency

COMPUTER SCIENCE DEPARTMENT
Stanford University



Stanford Artificial Intelligence Laboratory **April 1978**
Memo AIM-3 12

Computer Science Department
Report No. STAN-CS-78-667

ON THE MODEL THEORY OF KNOWLEDGE

by

John McCarthy, Masahiko Sato, Takeshi Hayashi, Shigeru Igarashi

Another language for expressing "knowing that" is given together with axioms and rules of inference and a Kripke type semantics. The formalism is extended to time-dependent knowledge. Completeness and decidability theorems are given. The problem of the wise men with spots on their foreheads and the problem of the unfaithful wives are expressed in the formalism and solved.

The authors' present addresses are as follows: John McCarthy, Stanford University; Masahiko Sato, University of Tokyo; Takashi Hayashi, Kyushu University; and Shigeru Igarashi, University of Tsukuba.

This research was supported by the Advanced Research Projects Agency of the Department of Defense under ARPA Order No. 2494, Contract MDA903-76-C-0206. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of Stanford University, or any agency of the U. S. Government.

Introduction

The need to represent information about who knows what in intelligence computer programs was the original motivation for this work. For example, a program that plans trips must know that travel agents know who knows the availability of rooms in hotels. An early problem is how to represent what people know about other people's knowledge of facts, and even the knowledge of propositions treated in this paper presented some problems that were not treated in previous literature.

We started with the following well known puzzle of the three wise men: *A king wishing to know which of his three wise men is the wisest, paints a white spot on each of their foreheads, tells them at least one spot is white, and asks each to determine the color of his spot. After a while the smartest announces that his spot is white reasoning as follows: "Suppose my spot were black. The second wisest of us would then see a black and a white and would reason that if his spot were black, the dumbest would see two black spots and would conclude that his spot is white on the basis of the king's assurance. He would have announced it by now, so my spot must be white."*

In formalizing the puzzle, we don't wish to try to formalize the reasoning about how fast other people reason. Therefore, we will imagine that either the king asks the wise men in sequence whether they know the colors of their spots or that he asks synchronously, "Do you know the color of your spot" getting a chorus of noes. He asks it again with the same result, but on the third asking, they answer that their spots are white. Needless to say, we are also not formalizing any notion of relative wisdom.

We start with a general set of axioms for knowledge based on the notation, axioms, and inference rules of propositional calculus supplemented by the notation $S*p$ standing for, "Person S knows proposition p." Thus $S_3*S_2*\neg(S_1*p_1)$ can stand for, "The third wise man knows that the second wise man knows that the first wise man does not know that the first wise man's spot is white".

We use axiom schemata with subscripted S's as person variables, subscripted p's and *q's as* propositional variables, and a special person constant called "any fool" and denoted by 0. It is convenient to introduce "any fool" because whatever he knows, everyone knows that everyone else knows. "Any fool" is especially useful when an event occurs in front of all the knowers, and we need sentences like, "S_1 knows that S_2 knows that S_3 knows etc.". Here are the schemata:

KO: $S*p\supset p$; What a person knows is true.

K 1: $0*(S*p\supset p)$; Any fool knows that what a person knows is true.

*K*2: $0*(0*p\supset 0*S*p)$; What any fool knows, any fool knows everyone knows, and any fool knows that.

K3: $0*(S*p\wedge S*(p\supset q)\supset S*q)$; Any fool knows everyone can do modus ponens.

There are two optional schemata *K4* and *K5:*

K4: $0*(S*p\supset S*S*p)$; Any fool knows that what someone knows, he knows he knows.

K5: $0*(\neg S*p\supset S*\neg S*p)$; Any fool knows that what some doesn't know he knows he doesn't know.

If there is only one person S, the system is equivalent to a system of modal logic. Axioms K l-K3 give a system equivalent to what Hughes and Cresswell [1] called T, and $K4$ and $K5$ give the modal systems S4 and S5 respectively. We call $K4$ and $K5$ the introspective schemata.

It is convenient to write $S\$p$ as an abbreviation for $S*p \vee S*\neg p$; it may be read "S knows whether p".

On the basis of these schemata we may axiomatize the wise man problem as follows:

C0: $p_1 \wedge p_2 \wedge p_3$

C1: $0*(p_1 \vee p_2 \vee p_3)$

C2: $0*(S_1\$p_2 \wedge S_1\$p_3 \wedge S_2\$p_1 \wedge S_2\$p_3 \wedge S_3\$p_1 \wedge S_3\$p_2)$

C2: $0*(S_2\$S_1*p_1)$

C3: $0*(S_3\$S_2*p_2)$

c4: $\neg S_1\$p_1$

c5: $\neg S_2\$p_2$

From KO-$K3$ and C l-C5 it is possible to prove S_3*p_3. CO is not used in the proof. In some sense C4 and C5 should not be required. Looking at the problem sequentially, it should follow that S 1 does not know p 1 initially, and that even knowing that, S_2 doesn't know p_2.

In order to proceed further with the problem, model theoretic semantics is necessary. In what follows, however, we will deal with the puzzle of unfaithful wives (cf. §4) rather than that of three wise men, because the latter may be considered as a simplified version of the former. To do so we must extend the system $K5$ to $KT5$ in which one can treat the notion of time as well. We will use slightly different notations in the following sections since they are convenient to denote time and have similarity to those used in ordiary modal logics.

We briefly describe the Hilbert-type formulation of the system $KT5$ in §2, and its model theory in §3. Finally, we will sketch the outline of the solution to the puzzle of unfaithful wives in this formalism in 54. The reader is referred to Sato [2] for details.

The Formal Systems

Basic Language

The basic language L is a triple (Pr, Sp, N^+), where

$Pr = p_1, p_2, \ldots;$
$Sp = S_0, S_1, \ldots.$
$N^+ = 1, 2, \ldots$

are denumerable sequence of distinct symbols. N^+ is the set of numerals denoting the corresponding positive integers. SO $\in S_p$ will be denoted by 0 and will called any fool.

Languages

A language L is a triple (Pr, Sp, T), where

$Pr \subseteq Pr$;
$Sp \subseteq Sp$;
$T \subseteq N^+$.

Elements in Pr, Sp and T denote propositional variables, persons and time, respectively. Our arguments henceforth will, unless stated otherwise, always be relative to a language L.

Well formed formulas

The set of well formed formulas is defined to be the least set Wff such that:

$(W1)$ $\perp \in Wff$,
$(W2)$ $Pr \subseteq Wff$,
(W3) a, $\beta \in Wff$ implies $\supset\alpha\beta \in Wff$,
$(W4)$ S $\in Sp$, t $\in T$, a $\in Wff$ implies $Sta \in Wff$.

The symbols $\perp$ and $\supset$ denote *false* and *implication*, respectively.

We will make use of the following abbreviations:

$\alpha\supset\beta = \supset\alpha\beta$	read "a implies β"
$\neg\alpha = \alpha\supset\perp$	read "not a"
$T = \neg\perp$	read "true"
$\alpha\vee\beta = \neg\alpha\supset\beta$	read "a or β"
$\alpha\wedge\beta = \neg(\alpha\supset\neg\beta)$	read "a and β"
$[St]\alpha$ = Sta	read "S knows a at time t" (This corresponds to $*$ in §1.)
$<St>\alpha = \neg[St]\neg\alpha$	read "a is possible for S at time t"
$\{St\}\alpha = [St]\alpha\vee[St]\neg\alpha$	read "S knows whether a at time t" (This corresponds to \$ in §1.)

For any a $\in Wff$, we define Sub(a) $\subseteq Wff$ inductively as follows:

$(S1)$ $a \in Pr$ u $\{\perp\}$ => Sub(a) = (a),
$(S2)$ $\alpha = \beta \supset \gamma$ => $Sub(a)$ = {a} u $Sub(\beta)$ u Sub(Y),
$(S3)$ a = $[St]\beta$ => Sub(a) u $Sub(\beta)$.

We say β is a *subformula* of a if $\beta \in$ Sub(a).

Hilbert-type system

We now define the modal system KT5. The axiom **schemata for KT5 are as follows:**

(A1) $7\neg\alpha\supset\alpha$
(A2) $\alpha\supset(\beta\supset\alpha)$
(A3) $(\alpha\supset(\beta\supset\gamma))\supset((\alpha\supset\beta)\supset(\alpha\supset\gamma))$
(A4) $[St]\alpha\supset\alpha$
(A5) $[0t]\alpha\supset[Ot][St]\alpha$
(A6) $[St](\alpha\supset\beta)\supset([Su]\alpha\supset[Su]\beta)$, where $t \leq u$
(A7) $\neg[St]\alpha\supset[St]\ 7\ [St]\alpha$

We have the following two inference rules:

(R1) $\dfrac{a \qquad \alpha\supset\beta}{\beta}$ (modus ponens)

(R2) $\dfrac{\alpha}{[St]\alpha}$ ([S&necessitation)

We write $\vdash$ **a** if **there exists a proof of a. For any** $\Gamma\subseteq Wff$ **we write** $\Gamma\vdash$ **a** if $\vdash\beta_1\supset(\beta_2\supset($ . . . $(\beta_m\supset\alpha)$ . . .)) for **some** $\beta 1, \ldots, \beta_m\in\Gamma$. Γ **is said to be** *consistent* **if** $\Gamma\not\vdash$ **1.**

Kripke-type Semantics

Definition of Kripke-type **models**

Let W **be any non-empty** set **(of possible worlds). A model** M **on** W is a **triple**

$$\langle W; r, v\rangle ,$$

where

$$r: Sp\ x\ T \longrightarrow 2^{W\times W}$$

and

$$v:\ Pr\ u\ \{\perp\}\longrightarrow 2^{W}.$$

Given any **model** M, **we define a relation** $\models\subseteq W\ x\ Wff$ **as follows:**
(E1) $a\in Pr$ u $\{\perp\}$ => $w\models$ a *iff* $w\in v(\alpha)$,
(E2) $\alpha=\beta\supset\gamma$ => $w\models$ *iff* not $w\models\beta$ *or* $w\models\gamma$,
(E3) $a=[St]\beta$ => $w\models$ *iff* **for all** $w'\in W$ **such that** (w, w') $\in r(S, t)$, w' f= β.

We will *write* **"**$w\models\alpha$ **(in** M**)" if we wish to make** M explicit. A formula a is said to **be** valid in

M, denoted by $M \vDash a$, if $w \vDash a$ for all $w \in M$. (By $w \in M$, we mean $w \in W$.) Furthermore, we will employ the following notation:

$w \vDash \Gamma$ (read "w realizes Γ") *iff* $w \vDash a$ for all $a \in \Gamma$

A model M is a KT5-model if

$(M1)$ $r(\perp) = \varnothing$,
$(M2)$ $r(O, t) \supseteq r(S, t)$ for any $S \in Sp$ and $t \in T$,
$(M3)$ $r(S, u) \supseteq r(S, t)$ for any $S \in Sp$ and $u, t \in T$ such that $u \leq t$,
$(M4)$ $r(S, t)$ is an equivalence relation for any $S \in Sp$ and $t \in T$.

A set Γ of well formed formulas is said to be *realizable* if there exists a KT-5 model M and $w \in M$ such that $w \vDash \Gamma$.

Soundness of KT5-models

We now wish to show that each formula provable in KT5 is valid in any KT5-model.

Theorem 1. (Soundness Theorem) If $\vdash a$ then $M \vDash a$ for any KT5-model M.

Corollary 2. (Consistency of KT5) $\perp$ is not provable in KT5.

Completeness of KT5-models

As for the completeness of KT5-models, we have the following theorems.

Theorem 3. (Generalized Completeness Theorem) Any consistent set of well formed formulas is realizable.

Theorem 4. (Completeness and Decidability Theorem) For any $a \in Wff$, a is a theorem of KT5 if and only if a is valid in all KT5-models whose cardinality $\leq 2^n$, where n is the cardinality of the finite set Sub (a) $\cup \{\perp\}$.

The Puzzle of Unfaithful Wives

We begin by explaining the notions of knowledge base and knowledge set, which are fundamental for our formalization of the puzzle of unfaithful wives.

Knowledge set and knowledge base

Let L be any language. We will make the notion of the totality of one's knowledge explicit by the following definitions:

Definition. $K \subseteq Wff$ is a *knowledge set* for St if K satisfies the following conditions:

$(KS1)$ K is consistent.

(KS2) $K = [St]\bar{K}$, where $\bar{K} = \{\alpha | K \vdash \alpha\}$.
(KS3) If $K \vdash [St]\alpha_1 \vee \ldots \vee [St]\alpha_n$ then $K \vdash \alpha_i$ for some i $(1 \leq i \leq n)$.

Definition. $B \subseteq Wff$ is a *knowledge base* for St if B satisfies the following conditions:

(KB1) B is consistent.
(KB2) $B \subseteq [St]\bar{B}$, where $\bar{B} = \{\alpha | B \vdash \alpha\}$,
(KB3) If $B \vdash [St]\alpha_1 \vee \ldots \vee [St]\alpha_n$ then $B \vdash \alpha_i$ for some i $(1 \leq i \leq n)$.

By (KS2) (or (KB2)) we see that any element in K (or B, resp.) has the form $[St]\alpha$. It is easy to see that if B is a knowledge base for St then $[St]\bar{B}$ is a knowledge set for St.

Let $\Gamma \subseteq Wff$ be consistent. We compare the following three conditions.

(1) If $\Gamma \nvdash \alpha$ then $\Gamma \vdash \neg[St]\alpha$.
(2) If $\Gamma \vdash [St]\alpha_1 \vee \ldots \vee [St]\alpha_n$ then $\Gamma \vdash \alpha_i$ for some i $(1 \leq i \leq n)$.
(3) If $\Gamma \vdash \{St\}\alpha$ then $\Gamma \vdash \alpha$ or $\Gamma \vdash \neg\alpha$.

Then we have the following

Lemma 5. (1), (2) and (3) are equivalent.

We now study the semantical characterization of knowledge sets. Let $M = \langle W; r, v \rangle$ be any KT5-model. For any $w \in W$ and $\langle S, t \rangle \in Sp \times T$, we define $K_w(St) \subseteq Wff$ by:

$$K_w(St) = \{[St]\alpha \mid w \models [St]\alpha\}.$$

Since, as we will see below, $K_w(St)$ is a knowledge set for St, we call it the knowledge set for st at w.

Lemma 6. $K_w(St)$ is a knowledge set for St.

Let K be a knowledge set for St. We say $w \in M$ *characterizes* K if $K = K_w(st)$.

Theorem 7. Any knowledge set is characterizable.

Informal presentation of the puzzle

The puzzle of unfaithful wives is usually stated as follows:

There was a country in which one million married couples inhabited. Among these one million wives, 40 wives were unfaithful. The situation was that each husband knew whether other men's wives were unfaithful but he did now know whether his wife was unfaithful. One day (call it the first day), the King of the country publicized the following decree:

(i) There is at least one unfaithful wife.

(ii) Each husband knows whether other men's wives are unfaithful or not.

(iii) Every night (from tonight) each man must do his deduction, based on his knowledge so far, and try to prove whether his wife is unfaithful or not.

(iv) Each man, who has succeeded in proving that his wife is unfaithful, must chop off his wife's head next morning.

(v) Every morning each man must see whether somebody chops off his wife's head.

(vi) Each man's knowledge before this decree is publicized consists only of the knowledge about other men's wive's unfaithfulness.

The problem is "what will happen under this situation?" The answer is that on the 41^{st} day 40 unfaithful wives will have their heads chopped off. We will treat this puzzle in a formal manner.

Formal treatment of the puzzle

We will treat this puzzle by assuming that there are $k(\geq 1)$ married couples in the country. Then the language $L = (Pr, Sp, T)$ adequate for this puzzle will be:

$$Pr = \{p_1, \ldots, p_k\},$$
$$Sp = \{O, S_1, \ldots, S_k\},$$
$$T = N^+.$$

where S_i denotes i^{th} husband, p_i means that S_i's wife is unfaithful and $t \in T$ denotes t^{th} day.

Let $\{\pm\}^k = \{+,-\}^k$ denote the k-fold Cartesian product of the vector space $GF(2) = \{+(=1), -(=0)\}$ with addition $\oplus$. We define

$$n : \{\pm\}^k \dashrightarrow Wff$$

by $\pi(\varepsilon_1 \ldots \varepsilon_k) = \overset{k}{\underset{i=1}{A}} p_i^{\varepsilon_i}$, where $\varepsilon_i \in \{\pm\}$ and $p_i^+ (p_i^-)$ denotes $p_i (\neg p_i$, resp.). We put $\Pi =$ Image(n) and $\Pi_0 = \Pi - \{\overset{k}{\underset{i=1}{A}} p_i^-\}$. We also use n to denote arbitrary element in Π. Now, let Γ denote what the King publicized on the first day, and $B_\pi(S_i n)\,(i = 1, \ldots, k)$ denote a knowledge base for $S_i n$ under the situation $n = \pi(\varepsilon_1 \ldots \varepsilon_k) \in \Pi_0$. Let us put:

$$\ulcorner B_\pi(S_i n) \vdash \alpha \urcorner = \begin{cases} \top & \text{if } B_\pi(S_i n) \vdash \alpha \\ \bot & \text{otherwise} \end{cases}$$

and

$$\ulcorner B_\pi(S_i n) \vdash \alpha \urcorner = \begin{cases} \top & \text{if } B_\pi(S_i n) \nvdash \alpha \\ \bot & \text{otherwise} \end{cases}$$

where $a \in Wff$. We also put $(\mathrm{It}) = \{1, \ldots, k\}$. Then, as a formalization of the puzzle, we postulate the following identities:

$$B_\pi(S_i 1) = [S_i 1]\Gamma \cup [S_i 1]p_j^{\varepsilon_j} \mid j \neq i, j \in (k)\}, \text{ where } \pi = \pi(\varepsilon_1, \ldots, \varepsilon_k) \qquad Eq(\pi, i, 1)$$

$$B_\pi(S_i n+1) = [S_i n+1]B_\pi(S_i n) \cup \{[S_i n+1][S_j n]p_j \mid B_\pi(S_j n) \vdash P_j, j \in (k)\}$$
$$\cup\{[S_i n+1] \neg [S_j n]p_j \mid B_\pi(S_j n) \nvdash p_j, j \in (k)\} \qquad Eq(\pi, i, n+1)$$

$$\Gamma = \{[01]\bigvee_{i=1}^{k} p_i\} \cup \{[01]\{S_i 1\}p_j \mid j \neq i, i \in (k), j \in (k)\}$$
$$\cup\{[01](\pi \supset (\ulcorner B_\pi(S_i n) \vdash p_i \urcorner \supset [0n+1][S_i n]p_i)) \mid \pi \in \Pi_0, i \in (k), n \in T\}$$
$$\cup\{[01](\pi \supset (\ulcorner B_\pi(S_i n) \nvdash p_i \urcorner \supset [0n+1]\neg[S_i n]p_i)) \mid \pi \in \Pi_0, i \in (k), n \in T\}$$
$$\cup\{[01](\ulcorner B_\pi(S_i n) \vdash \alpha \urcorner \supset [01](n \supset [S_i n]\alpha)) \mid n \in \Pi_0, i \in (k), \alpha \in Wff\} \qquad Eq(*)$$

Since the meta-notions such as knowledge base and provability ($\vdash$) cannot be expressed directly in our language, we were forced to interpret the King's decree into Γ in a somewhat indirect fashion.

Now, if we read $Eq(*)$ as the definition of Γ, then we find that the definition is circular, since in order that Γ may be definable by $Eq(*)$ it is necessary that $B_\pi(S_i n)$ are already defined, whereas $B_\pi(S_i n)$ are defined in terms of Γ in $Eqs(\pi, i, n)$. So, we will treat these equations as a system $\sum = \{Ea(\pi, i, n) \mid \pi \in \Pi_0, i \in (k), n \in T\} \cup \{Eq(*)\}$ of equations with the unknowns $\{B_\pi(S_i n) \mid \pi \in \Pi_0, i \in (k), n \in T\}$ and Γ. We will solve $\sum$ under the following conditions:

(*) For any $n \in \Pi_0$, $\Gamma \cup \{\pi\}$ is consistent.

(**) For any $\pi \in \Pi_0$ and $S_i n$, $B_\pi(S_i n)$ is a knowledge base for $S_i n$.

We think these conditions are natural in view of the intended meanings of Γ and $B_\pi(S_i n)$.

Let us define a *norm* on $E = \{\pm\}^k$ by $\|\varepsilon\| = |\{i \mid \varepsilon_i = +\}|$, where $\varepsilon = \varepsilon_1 . , \varepsilon_k$. For any $\varepsilon = \varepsilon_1 , . . \varepsilon_k \in E$ and $i = 1, \ldots, k$, we put

$$\varepsilon(+i) = \varepsilon_1 \ldots \varepsilon_{i-1} + \varepsilon_{i+1} \ldots \varepsilon_k ,$$
$$\varepsilon(-i) = \varepsilon_1 \ldots \varepsilon_{i-1} - \varepsilon_{i+1} \ldots \varepsilon_k ,$$

and for any $\pi = \pi(\varepsilon) \in \Pi$, we put

$$\pi(+i) = \pi(\varepsilon(+i)) \ ,$$
$$\pi(-i) = \pi(\varepsilon(-i)) \ .$$

We also put $E0 = E\text{-}(0) = E - \{-\ldots-\}$.

We define a KT5-model $M = \langle E_0; r, v \rangle$ as follows:

(i) $(\varepsilon, \delta) \in r(S_i, n)$ *iff*

(a) $\varepsilon = \delta$

or

(b) $\varepsilon \oplus \delta = \underset{***i}{t \ldots t - t \ldots t}$ and $n < \|\varepsilon(+i)\| = \|\delta(+i)\|$.

(ii) $(\varepsilon, \delta) \in r(0, n)$ *iff*

(c) $\varepsilon = \delta$

or

(d) $n < max\{\|\varepsilon(+i)\| \mid i \in (k)\}$ and $n < max\{\|\delta(+i)\| \mid i \in (k)\}$.

(iii) $\varepsilon \in v(p_i)$ *iff* $\varepsilon_i = +$

(iv) $v(1) = 0$.

Then we have the following theorem.

Theorem 8. Under the conditions $(*)$ and $(**)$, $\sum$ has the unique solution $\langle\langle \tilde{B}_\pi(S_i n)\rangle, \tilde{\Gamma}\rangle$, where the solution is characterized by the condition:

$$\tilde{B}_{\pi(\varepsilon)}(S_i n) \vdash \alpha \text{ if and only if } \varepsilon \models [S_i n]\alpha \text{ (in } M).$$

Thus we have seen that $\tilde{\Gamma}$ may be regarded as the formal counterpart of the King's decree in our formal system. The puzzle is then reduced to the problem of showing that:

(P_1) If $\|\varepsilon\| = n$ and $\varepsilon_i = t$, then $\tilde{B}_{\pi(\varepsilon)}(S_i n) \vdash p_i$ and $\tilde{B}_{\pi(\varepsilon)}(S_i n-1) \nvdash p_i$.

We note that we can moreover prove the following:

(P_2) If $\|\varepsilon\| = n$ and $\varepsilon_i = -$, then $\tilde{B}_{\pi(\varepsilon)}(S_i n+1) \vdash \bar{p_i}$ and $\tilde{B}_{\pi(\varepsilon)}(S_i n) \nvdash \bar{p_i}$.

It is clear that (P_1) and (P_2) follow at once from the condition stated in Theorem 8.

References

[1] Hughes, G.E., and Creswell, M. J., *An Introduction To Modal Logic*, London: Methuen and Co., Ltd., 1968.

[2] Sato, M., A study of Kripke-type models for some modal logics by Gentzen's sequential method, Publ. RIMS, Kyoto Univ., 13 (1977), 381-468.