Memo 29 -- Artificial Intelligence Project
RLE and COMPUTATION CENTER
Massachusetts Institute of Technology
Cambridge 39, Massachusetts

November 10, 1961

INTRODUCTION TO THE CALCULUS OF KNOWLEDGE

by Bertram Raphael

ABSTRACT

This paper deals with the "Calculus of Knowledge", an extension of the propositional calculus in which one may reason about what other people know. Semantic and Syntactic systems are developed, certain theorems are proven, and a formal solution in the system of a well-known reasoning problem is presented.

I. INTRODUCTION

This paper deals with a calculus which enables one to formalize reasonings about what people know, as proposed by Dr. John McCarthy in his IR Memo, "A Calculus of the Knowledge of Propositions". We treat the Calculus of Knowledge as an extension of the propositional calculus which permits propositions of a certain form to be interpreted, "person S knows that proposition A is true". Following McCarthy, the precise formulation of this new system is restricted by requiring that the system have the following basic properties (motivated by our intuitive notion of the nature of such reasoning, modified by an attempt to keep this first version of a new logic reasonably simple):

1. What anyone knows is true;
2. For any proposition one knows whether one knows it;
3. One knows any logical consequences of the other things one knows. By "logical consequences" we mean all propositions deducible in the Calculus of Knowledge, since this Calculus is an attempt to formalize just such reasoning. In other words, the set of propositions which any person knows is logically closed in the system. Therefore any person knows all tautologies in the system (which must include all propositional-calculus tautologies).

McCarthy proposed a model and an axiom schema for this calculus, and he suggested certain problems and properties of his system which should be studied. However, the present author has shown the model to be inconsistent and would prefer to work with a somewhat different form of syntactic system. Therefore this paper deals with the following new work (based on McCarthy's basic formulations and notation):

1. The construction of a more direct (and, hopefully, correct) model will be described.
2. Viewing the Calculus of Knowledge as an extention of the Propositional Calculus, we have carried over concepts of implication and rules of deduction from a system of mathematical logic studied under Dr. Hartley Rogers, Jr. Some important theorems of the resulting system will be proven.
3. As an example of the use of the decision procedure, a formal solution to "the three wise men problem" will be presented.

II. FORMULAS:

The basic symbols to be used in formulas of the Calculus of Knowledge are the connectives

$$*,\ \sim,\ \wedge,\ \vee,\ \Rightarrow,\ \Leftrightarrow$$

and the variables

p, q, r, ... (standing for propositions)
S, S', S", ... (standing for persons).

Formulas are defined, in the calculus, as follows:

If α is a propositional variable, then α is a formula
If α is a formula, then $\sim\alpha$ is a formula
If α, β are formulas, then $[\alpha\wedge\beta]$ is a formula
If α, β are formulas, then $[\alpha\vee\beta]$ is a formula
If α, β are formulas, then $[\alpha\Rightarrow\beta]$ is a formula
If α, β are formulas, then $[\alpha\Leftrightarrow\beta]$ is a formula
If α is a formula and S is a people-variable, then $[S*\alpha]$ is a formula.

(Motivation: "$[S * \alpha]$" is to be interpreted as the propositional expressions, "Person S knows that formula α is true". The remainder is the usual structure of the Propositional Calculus.)

III. SEMANTICS:

A. Motivation for structure of models

In the Propositional Calculus, a model consists of any mapping π of all propositional variables into the set $\{0,1\}$. Once such an assignment is made, the truth-values of all formulas are uniquely determined by a straight-forward truth-table analysis. Our problem here is more difficult since we interpret $[S*\alpha]$ as, "person S knows that formula α is true". Thus even after propositional variables have been assigned values certain additional arbitrary decisions must be made to allow different people to "know" different "facts" (or, equivalently, the same person to know different facts in different models). However, the values of many formulas of the form $[S*\alpha]$ are uniquely determined if they are to be consistent with the three basic properties of the calculus (as listed informally in the Introduction). The third property, in particular, indicates that what one knows may be determined by other things he knows, and therefore the assignment to $[S*\alpha]$ depends on all assignments previously made to formulas of the form $[S*\beta]$.

B. Construction of satisfying models

We shall now describe how to construct a satisfying model for an arbitrary formula F. We shall then point out how that model could be extended to simultaneously satisfy additional formulas, and thus to be a full model for the calculus. First some definitions:

Def: A base element = df any formula which is either a propositional variable or a formula of the form [S*A].

Def: A model = df an assignment π which maps base elements into the set $\{0,1\}$. This assignment must be made in a particular manner to be described below.

Def: τ_π = df that mapping of all formulas into $\{0,1\}$ which is determined in a natural way by the assignment π;
i.e., for α any base element, $\tau_\pi(\alpha) = \pi(\alpha)$;
for α,β any formulas, $\tau_\pi([\alpha \wedge \beta]) = \begin{cases} 1 \text{ if } \tau_\pi(\alpha) = \tau_\pi(\beta) = 1, \\ 0 \text{ otherwise} ; \end{cases}$
and similarly for the other Boolian connectives.

Def: A formula A is a tautology = df in all possible models (i.e. for all permissible assignments π to the base elements which are subformulas of A), $\tau_\pi(A) = 1$.

Def: B is said to be a consequence of A = df in all possible models, if $\tau_\pi(A) = 1$, then $\tau_\pi(B) = 1$. (We also say A implies B, or B is deducible from A). (It follows immediately that A implies B if and only if $[A \Rightarrow B]$ is a tautology.)

Def: B is said to be a consequence of a set of formulas $\mathcal{A}$ = df in all possible models, if $\tau_\pi(A_i) = 1$ for every $A_i \in \mathcal{A}$, then $\tau_\pi(B) = 1$.

Def: A set of formulas $\mathcal{A}$ is logically closed = df for any formula A, if $\mathcal{A}$ implies A, then $A \in \mathcal{A}$.

Let K be a formula which, at any stage during the construction of our model, has the following property: For any formula A, S "knows" A (i.e., the assignment $\pi([S*A]) = 1$ is to be made), if and only if A is a consequence of K. Then, by adding conjuncts to K, we can increase the scope of S's knowledge while satisfying the property that the set of propositions which S knows is logically closed (since it is just the set of consequences of K).

Note 1: We must insure that K remains consistent in order to obtain an interesting model.

Note 2: Since any formula has only a finite number of subformulas, thus only a finite number of base elements and a finite number of models, we can effectively test whether a formula is a tautology by simply enumerating the possible models. A systematic proceddre for doing this will be described below.
Let $\mathcal{L}$ be the set of formulas $\{B_i\}$ which are true but which S must not know (i.e. for which $\tau_\pi(B_i) = 1$, but the assignment $\pi([S*B_i]) = 0$ has been made).

Note 3: We cannot permit the formula $[K \Rightarrow B_i]$ to be a tautology, for any $B_i \in \mathcal{L}$, for if it were the model would be inconsistent since $[S*B_i]$ would have to be mapped into both 0 and 1.

First we assign to each formula a level. Formulas not involving any S*'s are assigned level 0. A combination of formulas by propositional connectives is assigned a level equal to the highest of the levels of its components, and a formula S*A is assigned a level one greater than that assigned to A. Now with K initially the null formula and $\mathcal{L}$ the empty set, we construct an assignment π to base elements of our formula F as follows: First make assignments to all base elements of the lowest level, then to those of the next higher level, etc., according to the following restrictions:

1) Assign all propositional variables (base elements of level zero) 0 or 1 arbitrarily.

2) Assign formulas of the form [S*A] (higher level base elements) as follows:

a) If A is [S*B] or ~[S*B] for any formula B, set $\pi([S*A]) = 1$ and replace K by [KΛA].

b) If [K⇒A] is a tautology, set $\pi([S*A]) = 1$.

c) If [[KΛA]⇒B_i] is a tautology for any $B_i \in \mathcal{L}$, set $\pi([S*A]) = 0$ and replace $\mathcal{L}$ by $\mathcal{L} \cup A$.

d) If $\tau(A) = 0$, set $\pi([S*A]) = 0$

e) Otherwise, assign $\pi([S*A])$ arbitrarily to 0 or 1. However, if $\pi([S*A]) = 1$, replace K by [KΛA and if $\pi([S*A]) = 0$, replace $\mathcal{L}$ by $\mathcal{L} \cup A$.

Now to extend this model to cover additional formulas, simply order the base elements of the next formula according to level. Then, keeping the K and $\mathcal{L}$ which resulted from the assignments thus far, assign all new propositional variables arbitrarily and proceed to assign other base elements as in 2) above.

C. Semantic Decision Procedure

We can test whether a formula is a tautology in the calculus by means of a truth-table format, by following rules a) thru e) above and addigg one line for each possible model. The details of the procedure are presented by means of an example (Table 1), explained below, in which it is established that

[[[S*[pVq]] Λ [S*~p]] ⇒ [S*q]]

is a tautology. The table is set up with one column for each base element (I, III, V) and Boolian connective (II, IV) at the top level of the formula being tested, and additional columns for propositional variables (VI, VII), other relevant sub-expressions (VIII, IX), and the formula K and set $\mathcal{L}$ for each model. Columns VI thru IX, lines 1 thru 4, are immediately generated by making all possible assignments to propositional variables. Then column V was chosen to be filled in (I, III, or V could be filled in in any order). The assignment "0" in lines 1 and 2 is determined by rule d). Lines 3 and 4 are to be assigned arbitrarily, by rule e). Thus they were assigned "0" and "q" placed under $\mathcal{L}$, and also copies of lines 3 and 4 thus far, with a "1" in column V and "q" added to K, were added to the table (lines 5 and 6), so that all possible models are represented. (We shall speak of this as the "splitting" of models.)

Now column III is assigned values. Lines 1,3, and 5 each split, generating lines 7, 8, 9. All previous assignments, including K and $\mathcal{L}$, are always carried along in the split and added to if necessary. In column I: Lines 1 and 7 are determined by d). Line 2 splits (generating 10). Lines 3 and 4 split, after we establish that [[pVq]⇒q] and [[pVq]⇒~p] are not tautologies so that c) does not apply. Lines 5, 6, and 9 do not split, since [q⇒[pVq]] and [[pΛ~p]⇒[pVq] are tautologies (which can be established in similar, smaller truth-tables), so the assignment is determined by b). In line 8, since [[~pΛ[pVq]]⇒q] is a tautology, the assignment is determined by c).

Now the remaining columns may be filled in in the usual way, and, since column IV turns out to have all "1"s, the tautology is established.

IV. PROPERTIES OF THE FORMAL SYSTEM

A. Required properties

The models described above all satisfy the conditions required in the introduction to this paper, for the following reasons:

1) "What anyone knows is true." By rule d), [S*A] cannot be mapped into 1 unless A is mapped into 1.

2) } "For any proposition one knows whether one knows it."
Rule a } precisely covers this case.

3) "One knows any logical consequences of the other things one knows." This is satisfied since the set of things S knows is logically closed (being the closure of a single formula K), in any one model. Formally, this condition is equivalent to establishing that the formula

$$[[[[S*A] \wedge [S*B]] \wedge [[A \wedge B] \Rightarrow C]] \Rightarrow [S*C]]$$

is a tautology. This could be done as in the example above in III C; or one could use precisely the statement which was proven a tautology in the example if one established that

$$[[S*A] \wedge [S*B]] \Leftrightarrow [S * [A \wedge B]]$$

is a tautology, and that "Substitutivity of Equivalence" is a legitimate derived rule of the system (both of which are probably very easy tasks).

B. Further Theorems

Thm 1: The rules of deduction in this system are sound and (semantically) complete.

Proof: Immediate from the usual definitions of soundness and completeness, since we have defined implication in terms of the satisfying models.

Thm 2: If $\mathcal{A}$ implies B, then there is a finite subset $\mathcal{D} \subseteq \mathcal{A}$ such that $\mathcal{D}$ implies B.

PROOF: Assume $\mathcal{A}$ is infinite, $A_0, A_1, \ldots \in \mathcal{A}$. Define $\mathcal{S}_n = \{A_0, A_1, \ldots, A_n\}$. Suppose the theorem is false. Then for every finite $\mathcal{D} \subset \mathcal{A}$, it is not true that $\mathcal{D}$ implies B:
i.e., there exists a model π such that

$$\tau_\pi(\mathcal{D}) = 1 \quad \text{and} \quad \tau_\pi(B) = 0;$$

i.e., there exists a model π such that

$$\tau_\pi(\mathcal{D}, \sim B) = 1, \quad \text{for any finite } \mathcal{D};$$

i.e., for every n, there exists a π such that

$$\tau_\pi(\mathcal{S}_n, \sim B) = 1.$$

We shall now show that this implies that there exists a particular model λ such that for every n, $\tau_\lambda(\mathcal{S}_n, \sim B) = 1$; therefore $\tau_\lambda(\mathcal{A}, \sim B) = 1$, which contradicts the assumption that $\mathcal{A}$ implies B.

Let $C_0, C_1, C_2, \ldots$ be any enumeration of all base elements appearing in $\{\sim B, A_0, A_1, A_2, \ldots\}$, listing first all base elements in $\sim B$, then all those in A_0, etc., subject to the restriction that if C_i is a subexpression of C_j, then $i < j$. (This requires propositional variables to appear before other base elements which contain them.)

Let $\lambda_0 = 0$ if, for every n (no matter how large), there exists a model $\lambda^{(n)}$ such that $\tau_{\lambda^{(n)}}(\sim B, \mathcal{S}_n) = 1$ and $\lambda^{(n)}(C_0) = 0$.

Let $\lambda_0 = 1$ otherwise.

(Note 1: If $\tau_\lambda(\sim B, \mathcal{S}_n) = 1$, then $\tau_\lambda(\sim B, \mathcal{S}_m) = 1$ for all $m \leq n$.)

Therefore for every n, there is a model $\lambda^{(n)}$ such that $\tau_{\lambda^{(n)}}(\sim B, \mathscr{L}_n) = 1$ and $\lambda^{(n)}(C_0) = \lambda_0$.

Now assume $\lambda_0, \lambda_1, \ldots, \lambda_i$ have already been defined.

$$\lambda_{i+1} = \begin{cases} 0 \text{ if, for every n, there exists a model } \lambda^{(n)}, \text{ constructed according to the rules a) through e), such that the following conditions all hold:} \\ \qquad \alpha) \ \tau_{\lambda^{(n)}}(\sim B, \mathscr{L}_n) = 1 \\ \qquad \beta) \ \lambda^{(n)}(C_j) = \lambda_j \text{ for } 0 \leq j \leq i \\ \qquad \gamma) \ \lambda^{(n)}(C_{i+1}) = 0. \\ 1 \text{ otherwise} \end{cases}$$

(Note 2: By hypothesis, there always exists at least one model such that α) holds. By the construction thus far, at least that model also satisfies β). Because of Note 1, for every n there exists at least one model $\lambda^{(n)}$ satisfying α) and β) for which either γ) is also satisfied, or for every n, $\lambda^{(n)}(C_{i+1}) = 1$. This choice determines λ_{i+1}.)

Define λ by $\lambda(C_i) = \lambda_i$ for all i.
Then $\tau_\lambda(\sim B, \mathscr{L}_n) = 1$ for all n.
Then $\tau_\lambda(\sim B, \alpha) = 1$, contradiction.

qed

V. THE WISE MEN PROBLEM

As an example of the use of our decision procedure, a formal solution will now be presented to a classical problem in reasoning.

Problem: Three wise men are told by their king that he has marked their foreheads with paint (white or black but not both), and that at least one of them has a white spot. The man are placed so that each can see the color painted on the foreheads of each of the others, but not the color of his own. The king asks the first and then the second wise man whether he knows the color of hisowwn spot, and received negative answers. The third wise man sees white spots on the foreheads of the other two. Prove that he (the third wise man) now has enough information to know that his own spot must be white.

Solution: Call the wise men W_1, W_2, W_3. Let P_i denote the proposition that W_i has a white spot.

We have not thus far discussed formulas involving more than one person-variable, but the extention is straightforward. In testing for tautologies, we must keep track of separate book-keeping functions K and $\mathscr{L}$ for each person.

The problem will be solved if we can show that a certain formal statement is a tautology; namely, a statement to be interpreted,

"If W_3 knows certain facts described in the hypothesis of the Wise Men Problem, then W_3 knows that his own spot is white".

Note that if S_1 knows that S_2 knows A, then S_1 must also know A (since $[S_2*A]$ implies A). Therefore it will be sufficient to hypothesize that W_3 knows just that W_2 has a white spot, plus that which W_2 knows; namely, that W_2 doesn't realize that he has a white spot, but he knows whether W_3 has a white spot, that W_1 has a white spot, that W_1 doesn't know that he (W_1) has a white spot, that W_1 knows whether W_2 and W_3 have white spots, and that W_1 knows there is at least one white spot.
Formally, we consider the following statement:

$$[W_3*\{p_2\wedge\sim[W_2*p_2]\wedge[[W_2*p_3]\vee[W_2*\sim p_3]]\wedge$$
$$[W_2*\langle p_1\wedge\sim[W_1*p_1]\wedge[[W_1*p_2]\vee[W_1*\sim p_2]]\wedge[[W_1*p_3\vee[W_1*\sim p_3]]\wedge$$
$$[W_1*[p_1\vee p_2\vee p_3]]\rangle]\}] \Rightarrow [W_3*p_3].$$

(Brackets have been omitted where it is unambiguous to do so). To test whether this is a tautology by our decision procedure, we need not consider all models, but rather only all those which map the left side into 1, therefore only those which map each conjunct in the $\{\}$ into 1, therefore only those which also map each conjunct in the $\langle\rangle$ into 1 (since for all other models the implication is automatically satisfied). The work is carried out in Table 2.

By the above remarks and columns b and j, we need only consider models which map p_1 and p_2 into 1. All other columns except a,h, and i, are also quickly determined by similar considerations. Since the expression in Table II b is a tautology, 1i must be zero and row 1 is eliminated from consideration. Since the only model which maps $\{\}$ into 1 also maps p_3 into 1, $[\{\}\Rightarrow p_3]$ is a tautology, so that $[W_3*p_3]$ must be mapped into 1 in the only model in which the hypotheses are satisfied. Therefore we have shown that the third wise man knows his spot is white.

VI. CONCLUSIONS

The semantic decision procedure, involving the construction of all possible models, is extremely long and awkward. Perhaps some general rules for decreasing the work required can be found. At least in certain special cases, e.g. the wise men problem, much less than the full construction is really necessary.

The precise formulation and solution to the wise men problem seemed unnecessarily difficult largely because it was "unnatural", i.e. not the way people usually attack the problem. I believe this was necessary because we are working with the bare skeleton of a formal system. If some derived rules of inference are added, such as reducto ad absurdum, the cut rule, subordinate proof derivations, etc., proofs will be possible in fewer, more "natural", steps.

This calculus is a first attempt to formalize an entire branch of common language usage -- namely reasoning about reasoning. It may be of interest to try to axiomatize the formalism into a Hilbert system, and investigate completeness of the resulting system, and how well it matches the intended model under standard interpretation. However, it is probably worthwhile to first generalize the system by adding quantifiers (to permit statements like, "There is an x such that for all S, S knows p(x).")

Relationships betweentthis system and various model logics shou should also be studied.

Eventually a system such as the one described above may be useful in enabling a computer to carry out the solution of reasoning problems, such as for example in the "advise taker" project.

Line	Generated by	I	II	III	IV	V	VI	VII	VIII	IX			
		[[[S*[pVq]]Λ[S*~p]] ⇒ [S*q]					p	q	~p	pVq	K	$\mathcal{L}$	Test expressions
1		0	0	0	1	0	0	0	1	0		~p	
2		0	0	0	1	0	1	0	0	1		[pVq]	
3		0	0	0	1	0	0	1	1	1		q, ~p, [pVq]	pVq ⇒ q, pVq ⇒ ~p
4		0	0	0	1	0	1	1	0	1		q, [pVq]	
5	3V	1	0	0	1	1	0	1	1	1	q	~p	
6	4V	1	0	0	1	1	1	1	0	1	q		
7	1III	0	0	1	1	0	0	0	1	0	~p		
8	3III	0	0	1	1	0	0	1	1	1	~p	q	~pΛ[pVq] ⇒ q
9	5III	1	1	1	1	1	0	1	1	1	[qΛ~p]		
10	2I	1	0	0	1	0	1	0	0	1	[pVq]		
11	3I	1	0	0	1	0	0	1	1	1	[pVq]		
12	4I	1	0	0	1	0	1	1	0	1	[pVq]		

TABLE 1

Mémo 29 -- 10

	$[W_3*\{p_2\wedge\sim[W_2*p_2]\wedge[[W_2*p_3]\vee[W_2*\sim p_3]]\wedge[W_2*\langle p_1\wedge\sim[W_1*p_1]\wedge[[W_1*p_2]\vee[W_1*\sim p_2]]\wedge$																	
1	0	1	1	0		0	1	1		0	0	1	1	0		1	1	0		1
2	1	1	1	0		1	1	0		1	1	1	1	0		1	1	0		1
column	a	b	c	d		e	f	g		h	i	j	k	l		m	n	o		p

contd next line

	$[[W_1*p_3]\vee[W_1*\sim p_3]]\wedge[W_1*[p_1\vee p_2\vee p_3]]\rangle]\}]\Rightarrow[W_3*p_3]$						K_1	K_2	K_3	$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_3$
1	0	1	1		1		1	00	$p_2\wedge\sim p_3$	$\sim p_3$		p_1	p_2	
2	1	1	0		1		1	11	$p_2\wedge p_3$	p_3	{}	p_1	p_2	
column	q	r	s		t		u	vw						

T A B L E I I a

$[\sim p_3\wedge\langle p_1\wedge\sim[W_1*p_1]\wedge[[W_1*p_2]\vee[W_1*\sim p_2]]\wedge[[W_1*p_3]\vee[W_1*\sim p_3]]\wedge[W_1*[p_1\vee p_2\vee p_3]]\rangle]\Rightarrow p_2$													K_1K_1	K_2	K_3	$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_3$
10	1	1	0		1	1	0		1	0	1	1	1	11	$\sim p_3\wedge p_2$			p_1		
10	1	1	0		0	1	1		0	0	1	1	0	10	$\sim p_3\wedge\sim p_2$			p_1		

↑ because

$[p_2\Rightarrow[p_1\vee p_2\vee p_3]]$

and $[\sim p_3\wedge\sim p_2\wedge[p_1\vee p_2\vee p_3]\Rightarrow p_1]$ are tautologies

T A B L E I I b

This empty page was substituted for a blank page in the original document.

CS-TR Scanning Project
Document Control Form

Date : 11/30/95

Report # AIM-29

Each of the following should be identified by a checkmark:
Originating Department:

☒ Artificial Intellegence Laboratory (AI)
☐ Laboratory for Computer Science (LCS)

Document Type:

☐ Technical Report (TR) ☒ Technical Memo (TM)
☐ Other:_____

Document Information

Number of pages: 12(16-IMAGES)
Not to include DOD forms, printer intstructions, etc... original pages only.

Originals are:
☐ Single-sided or
☒ Double-sided

Intended to be printed as :
☐ Single-sided or
☒ Double-sided

Print type:
☐ Typewriter ☐ Offset Press ☐ Laser Print
☐ InkJet Printer ☐ Unknown ☒ Other: COPY OF MIMEOGRAPH

Check each if included with document:

☐ DOD Form ☐ Funding Agent Form ☐ Cover Page
☐ Spine ☐ Printers Notes ☐ Photo negatives
☐ Other: _____

Page Data:

Blank Pages(by page number): _____

Photographs/Tonal Material (by page number): _____

Other (note description/page number):

Description :	Page Number:
IMAGE MAP: (1-12) UNNUMBERED TITLE PAGE, 2-9, UNNUMBERED FIGS (2), UNNUMBERED BLANK	
(13-16) SCANCONTROL, TARGETS (3)	

Scanning Agent Signoff:
Date Received: 11/30/95 Date Scanned: 12/12/95 Date Returned: 12/14/95

Scanning Agent Signature: Michael W. Cook

Scanning Agent Identification Target

Scanning of this document was supported in part by the **Corporation for National Research Initiatives,** using funds from the **Advanced Research Projects Agency** of the **United states Government** under Grant: **MDA972-92-J1029.**

The scanning agent for this project was the **Document Services** department of the **M.I.T Libraries.** Technical support for this project was also provided by the **M.I.T. Laboratory for Computer Sciences.**

Scanned

Date: 12/12/1995

M.I.T. Libraries
Document Services