FORMALIZATION OF TWO PUZZLES INVOLVING KNOWLEDGE

John McCarthy
Computer Science Department
Stanford University
Stanford, CA 94305
jmc@cs.stanford.edu
http://www-formal.stanford.edu/jmc/

1978-1981

This paper describes a formal system and uses it to express the puzzle of the three wise men and the puzzle of Mr. S and Mr. P. Four innovations in the axiomatization of knowledge were required: the ability to express joint knowledge of several people, the ability to express the initial non-knowledge, the ability to describe *knowing what* rather than merely *knowing that*, and the ability to express the change which occurs when someone learns something. Our axioms are written in first order logic and use Kripke-style possible worlds directly rather than modal operators or imitations thereof. We intend to use functions imitating modal operators and taking "propositions" and "individual concepts" as operands, but we haven't yet solved the problem of how to treat learning in such a formalism.[1]

[1] 1997: The puzzle of the three wise men is old and well known. I have not been able to trace *Mr. S and Mr. P* back beyond its alleged appearance on a bulletin board at Xerox PARC. 2005 note: The puzzle has been recently traced to the Dutch mathematician Hans Freudenthal. Freudenthal didn't publish it and didn't give a reference. The consensus opinion is that Freudenthal invented it.

1 THE PUZZLES

The *three wise men puzzle* is as follows:

> *A certain king wishes to test his three wise men. He arranges them in a circle so that they can see and hear each other and tells them that he will put a white or black spot on each of their foreheads but that at least one spot will be white. In fact all three spots are white. He then repeatedly asks them, "Do you know the color of your spot?" What do they answer?*

The solution is that they answer, *"No,"* the first two times the question is asked and answer *"Yes"* thereafter.

This is a variant form of the puzzle. The traditional form is:

> *A certain king wishes to determine which of his three wise men is the wisest. He arranges them in a circle so that they can see and hear each other and tells them that he will put a white or black spot on each of their foreheads but that at least one spot will be white. In fact all three spots are white. He then offers his favor to the one who will first tell him the color of his spot. After a while, the wisest announces that his spot his white. How does he know?*

The intended solution is that the wisest reasons that if his spot were black, the second would see a black and a white and would reason that if his spot were black, the third would have seen two black spots and reasoned from the king's announcement that his spot was white. This traditional version requires the wise men to reason about how fast their colleagues reason, and we don't wish to try to formalize this.

Here is the *Mr. S and Mr. P* puzzle:

> *Two numbers m and n are chosen such that $2 \leq m \leq n \leq 99$. Mr. S is told their sum and Mr. P is told their product. The following dialogue ensues: Mr. P: I don't know the numbers.*
>
> *Mr. S: I knew you didn't know. I don't know either.*
> *Mr. P: Now I know the numbers.*
> *Mr S: Now I know them too.*

In view of the above dialogue, what are the numbers?

2007 note: At the time I wrote this article, I was unable to discover the author. It was Hans Freudenthal, Nieuw Archief Voor Wiskunde, Series 3, Volume 17, 1969, page 152).

2 AXIOMATIZATION OF THE WISE MEN

The axioms are given in a form acceptable to FOL, the proof checker computer program for an extended first order logic developed by Richard Weyhrauch at the Stanford Artificial Intelligence Laboratory (Weyhrauch, 1977). FOL uses a sorted logic. Constants and variables are declared to have given sorts, and quantifiers on these variables are interpreted as ranging over the sorts corresponding to the variables.

The axiomatization has the following features:

1. It is entirely in first order logic rather than in a modal logic.

2. The Kripke accessibility relation is axiomatized. No knowledge operator or function is used. We hope to present a second axiomatization using a knowledge function, but we haven't yet decided how to handle time and learning in such an axiomatization.

3. We are essentially treating "knowing what" rather than "knowing that". We say that p knows the color of his spot in world w by saying that in all worlds accessible from w, the color of the spot is the same as in w.

4. We treat learning by giving the accessibility relation a time parameter. To say that someone learns something is done by saying that the worlds accessible to him at time $n+1$ are the subset of those accessible at time n in which the something is true.

5. The problems treated are complicated by the need to treat joint knowledge and joint learning. This is done by introducing fictitious persons who know what a group of people know jointly. (When people know something jointly, not only do they all know it, but they jointly know that they jointly know it).

This isn't the place for a description of the FOL interactive theorem prover. However a few remarks will make it easier to read the axioms.

Since FOL uses a sorted logic, it must be told the sorts of the variables and constants, so it can determine whether a substitution is legitimate. This is done by *declare* statements resembling declarations in programming languages. The notation for formulas in as is usual in logic, so there shouldn't be difficulty reading it. Writing it so that the computer will accept it is a more finicky task.

declare $INDCONST\ RW \in WORLD$;

declare $INDVAR\ w\ w1\ w2\ w3\ w4\ w5 \in WORLD$;

RW denotes the real world, and $w, w1, \ldots, w5$ are variables ranging over worlds.

$declare\ INDVAR\ m\ n\ m1\ m2\ m3\ n1\ n2\ n3 \in NATNUM;$

We use natural numbers for times.

$declare\ INDCONST\ S1\ S2\ S3\ S123 \in PERSON;$

$declare\ INDVAR\ p\ p0\ p1\ p2 \in PERSON;$

$S1$, $S2$ and $S3$ are the three wisemen. $S123$ is a fictitious person who knows whatever $S1$, $S2$ and $S3$ know jointly. The joint knowledge of several people is typified by events that occur in their joint presence. Not only do they all know it, but $S1$ knows that $S2$ knows that $S1$ knows that $S3$ knows etc. Instead of introducing $S123$, we could introduce prefixes of like "$S1$ knows that $S2$ knows" as objects and quantify over prefixes.

$declare\ PREDCONST\ A(WORLD, WORLD, PERSON, NATNUM);$

This Kripke-style accessibility relation has two more arguments than is usual in modal logic — a person and a time.

$declare\ INDVAR\ c\ c1\ c2\ c3\ c4 \in COLORS;$

$declare\ INDCONST\ W\ B\ \in COLORS;$

There are two colors - white and black.

$declare\ OPCONST\ color(PERSON, WORLD) = COLORS;$

A person has a color in a world. A previous axiomatization was simpler. We merely had three propositions WISE1, WISE2 and WISE3 asserting that the respective wise men had white spots. We now need the colors, because we want to quantify over colors.

$axiom\ reflex: \quad \forall\ w\ p\ m.A(w, w, p, m);;$

The accessibility relation is reflexive as is usual in the Kripke semantics of M. It is equivalent to asserting that what is known is true.

$axiom\ transitive:$

$$\forall w1\, w2\, w3\, p\, m.(A(w1, w2, p, m) \wedge A(w2, w3, p, m) \supset A(w1, w3, p, m));;$$

Making the accessibility relation transitive gives an S4 like system. We use transitivity in the proof, but we aren't sure it is necessary.

$axiom\ who: \forall p.(p = S1 \vee p = S2 \vee p = S3 \vee p = S123);;$

We need to delimit the set of wise men.

$axiom\ w123:$

$$\forall w1\, w2\, m.(A(w1, w2, S1, m) \vee A(w1, w2, S2, m) \vee A(w1, w2, S3, m) \\ \supset A(w1, w2, S123, m));;$$

This says that anything the wise men know jointly, they know individually.

$axiom\ foolspot: \forall w.(color(S123, w) = W);;$

This ad hoc axiom is the penalty for introducing $S123$ as an ordinary individual whose spot must therefore have a color. It would have been better to distinguish between real persons with spots and the fictitious person(s) who only know things. Anyway, we give $S123$ a white spot and make it generally known, e.g. true in all possible worlds. I must confess that we do it this way here in order to repair a proof that the computer didn't accept on account of people not knowing the color of $S123$'s spot.

$\mathit{axiom\ color} : \neg(W = B); \quad \forall c.(c = W \vee c = B);;$

Both of these axioms about the colors are used in the proof.

$\mathit{axiom\ rw} : \mathit{color}(S1, RW) = W \wedge \mathit{color}(S2, RW) = W \wedge \mathit{color}(S3, RW) = W;;$

In fact all spots are white.

$\mathit{axiom\ king} : \forall w.(A(RW, w, S123, 0) \supset \mathit{color}(S1, w) = W \vee \mathit{color}(S2, w) = W \vee \mathit{color}(S3, w) = W);;$

They jointly know that at least one spot is white, since the king stated it in their mutual presence. We use the consequence that $S3$ knows that $S2$ knows that $S1$ knows this fact.

$\mathit{axiom\ initial} :$

$$\begin{array}{l}
\forall c\ w.(A(RW, w, S123, 0) \supset \\
\qquad\qquad (c = W \vee \mathit{color}(S2, w) = W \vee \mathit{color}(S3, w) = W \\
\qquad\qquad\qquad\qquad \supset \exists w1.(A(w, w1, S1, 0) \wedge \mathit{color}(S1, w1) = c)) \wedge \\
\qquad\qquad (c = W \vee \mathit{color}(S1, w) = W \vee \mathit{color}(S3, w) = W \\
\qquad\qquad\qquad\qquad \supset \exists w1.(A(w, w1, S2, 0) \wedge \mathit{color}(S2, w1) = c)) \wedge \\
\qquad\qquad (c = W \vee \mathit{color}(S1, w) = W \vee \mathit{color}(S2, w) = W \\
\qquad\qquad\qquad\qquad \supset \exists w1.(A(w, w1, S3, 0) \wedge \mathit{color}(S3, w1) = c))); \\
\forall w\ w1.(A(RW, w, S123, 0) \wedge A(w, w1, S1, 0) \\
\qquad\qquad \supset \mathit{color}(S2, w1) = \mathit{color}(S2, w) \wedge \mathit{color}(S3, w1) = \mathit{color}(S3, w)); \\
\forall ww1.(A(RW, w, S123, 0) \wedge A(w, w1, S2, 0) \\
\qquad\qquad \supset \mathit{color}(S1, w1) = \mathit{color}(S1, w) \wedge \mathit{color}(S3, w1) = \mathit{color}(S3, w)); \\
\forall w\ w1.(A(RW, w, S123, 0) \wedge A(w, w1, S3, 0) \\
\qquad\qquad \supset \mathit{color}(S1, w1) = \mathit{color}(S1, w) \wedge \mathit{color}(S2, w1) = \mathit{color}(S2, w));;
\end{array}$$

These are actually four axioms. The last three say that every one knows that each knows the colors of the other men's spots. The first part says that they all know that no-one knows anything more than what he can see and what the king told them. We establish non-knowledge by asserting the existence of enough possible worlds. The ability to quantify over colors is convenient for expressing this axiom in a natural way. In the S and P problem it is essential, because we would otherwise need a conjunction of 4753 terms.

$\mathit{axiom\ elwek1} :$

$\forall w.(A(RW, w, S123, 1) \equiv A(RW, w, S123, 0)$

$\wedge\forall p.(\forall w1.(A(w,w1,p,0) \supset color(p,w1) = color(p,w))$
$\equiv \forall w1.(A(RW,w1,p,0) \supset color(p,w1) = color(p,RW))));$
$\forall w1w2.(A(w1,w2,S1,1) \equiv A(w1,w2,S1,0) \wedge A(w1,w2,S123,1));$
$\forall w1w2.(A(w1,w2,S2,1) \equiv A(w1,w2,S2,0) \wedge A(w1,w2,S123,1));$
$\forall w1w2.(A(w1,w2,S3,1) \equiv A(w1,w2,S3,0) \wedge A(w1,w2,S123,1));;$

This axiom and the next one are the same except that one deals with the transition from time 0 to time 1 and the other deals with the transition from time 1 to time 2. Each says that they jointly learn who (if anyone) knows the color of his spot. The quantifier $\forall p$ in this axiom covers $S123$ also and forced us to say that they jointly know the color of $S123$'s spot.

*axiom elwek*2 :
$\forall w.(A(RW,w,S123,2) \equiv A(RW,w,S123,1)$
$\wedge\forall p.(\forall w1.(A(w,w1,p,1) \supset color(p,w1) = color(p,w))$
$\equiv \forall w1.(A(RW,w1,p,1) \supset color(p,w1) = color(p,RW))));$
$\forall w1w2.(A(w1,w2,S1,2) \equiv A(w1,w2,S1,1) \wedge A(w1,w2,S123,1));$
$\forall w1w2.(A(w1,w2,S2,2) \equiv A(w1,w2,S2,1) \wedge A(w1,w2,S123,1));$
$\forall w1w2.(A(w1,w2,S3,2) \equiv A(w1,w2,S3,1) \wedge A(w1,w2,S123,1));;$

The file WISEMA.PRF[S78,JMC] at the Stanford AI Lab contains a computer checked proof from these axioms of

$$\forall w.(A(RW,w,S3,2) \supset color(S3,w) = color(S3,RW))$$

which is the assertion that at time 2, the third wise man knows the color of his spot. As intermediate results we had to prove that previous to time 2, the other wise men did not know the colors of their spots. In this symmetrical axiomatization, we could have proved the theorem with a variable wise man instead of the constant $S3$.

3 AXIOMATIZATION OF MR. S AND MR. P

These axioms involve the same ideas as the wise man axiomatization. They are a debugged version of the axioms by Ma Xiwen of Peking University, which, in turn, were a variant of my earlier axiomatization.

This formalization separates the knowledge part of the problem from the arithmetic part in a neat way. Ma Xiwen verified, using FOL, that his axioms imply a certain purely arithmetic condition on the pair of numbers. It can be then shown that the only pair satisfying that condition is $4,13$.

$declare\ indvar\ t \in natnum;$
$declare\ indconst\ k0 \in pair;$
$declare\ indvar\ k$
$k1\ k2\ k3 \in pair;$
$declare\ indconst\ RW \in world;$
$declare\ indvar\ w\ w1\ w2\ w3 \in world;$
$declare\ indconst\ S\ P\ SP \in person;$
$declare\ indvar\ r \in person;$
$declare\ opconst\ K(world) = pair;$
$declare\ opconst\ s(pair) = natnum;$
$declare\ opconst\ p(pair) = natnum;$
$declare\ predconst\ A(world, world, person, natnum);$

The predicates Qs, Qp, $Q1$, $Q2$, $R1$, $R2$ and $R3$ are represent arithmetic conditions on the pair of numbers that are used to express the arithmetic conditions on the pair that replace the knowledge conditions given in the problem statement.

$declare\ predconst\ Qs(pair)\ Qp(pair)\ Q1(pair)\ Q2(pair)\ Q3(pair);$
$declare\ predconst\ Bs(world)\ Bp(world)\ B1(world)\ B2(world);$
$declare\ predconst\ R1(pair)\ R2(pair)\ R3(pair);$
$declare\ predconst\ C1(world)\ C2(world);$

The first two axioms state that the accessibility relation is reflexive and transitive. They assert that what is known is true and one knows that one knows what one knows. Got that?

$axiom\ ar : \forall w\ r\ t.A(w, w, r, t);;$
$axiom\ at : \forall w1\ w2\ w3\ r\ t.(A(w1, w2, r, t) \wedge A(w2, w3, r, t) \supset A(w1, w3, r, t));;$

This axiom says that the joint person knows what Mr. S and Mr. P both know. At first sight it seems too weak, but transitivity tells us that what they jointly know, they jointly know they jointly know.

$axiom\ sp : \forall w1\ w2\ t.(A(w1, w2, S, t) \vee A(w1, w2, P, t) \supset A(w1, w2, SP, t));;$

This next axiom is just a definition for the purposes of abbreviation. RW is the real world, so $k0$ is just the real pair.

$axiom\ rw : k0 = K(RW);;$

In the initial situation they jointly know that Mr. S knows the sum and Mr. P the product. They also jointly know that this is all that Mr. S and Mr. P know about the numbers. This is asserted by saying that given any pair of numbers with the right sum, there is a world possible for Mr. S in which this is the pair of numbers.

$axiom\ init :$

$\forall w\ w1.(A(RW, w, SP, 0) \wedge A(w, w1, S, 0) \supset s(K(w)) = s(K(w1)));$

$\forall w\ w1.(A(RW, w, SP, 0) \wedge A(w, w1, P, 0) \supset p(K(w)) = p(K(w1)));$

$\forall w\ k.(A(RW, w, SP, 0) \wedge s(K(w)) = s(k) \supset \exists w1.(A(w, w1, S, 0) \wedge k = K(w1)));$

$\forall w\ k.(A(RW, w, SP, 0) \wedge p(K(w)) = p(k) \supset \exists w1.(A(w, w1, P, 0) \wedge k = K(w1)));;$

These axioms on the pair will be used to translate knowledge assertions. For example, $Qs(k)$ asserts that there is another pair that has the same sum as k and is used in the assertion that Mr. S, knowing the sum, does not know the pair, i.e. does not know the numbers. As we proceed through the dialog the arithmetic conditions become more complex.

$axiom\ qs : \forall k.(Qs(k) \equiv \exists k1.(s(k) = s(k1) \wedge \neg(k = k1)));;$

$axiom\ qp : \forall k.(Qp(k) \equiv \exists k1.(p(k) = p(k1) \wedge \neg(k = k1)));;$

$axiom\ q :$

$\forall k.(Q1(k) \equiv \forall k1.(s(k) = s(k1) \supset Qp(k1)));$

$\forall k.(Q2(k) \equiv \forall k1.(R1(k1) \wedge p(k) = p(k1) \supset k = k1));$

$\forall k.(Q3(k) \equiv \forall k1.(R2(k1) \wedge s(k) = s(k1) \supset k = k1));;$

$axiom\ r :$

$\forall k.(R1(k) \equiv Qs(k) \wedge Q1(k));$

$\forall k.(R2(k) \equiv R1(k) \wedge Q2(k));$

$\forall k.(R3(k) \equiv R2(k) \wedge Q3(k));;$

$Bs(w)$ asserts that in the possible world w, Mr. S doesn't know the numbers.

$axiom\ bs : \forall w.(Bs(w) \equiv \exists w1.(A(w, w1, S, 0) \wedge \neg(K(w) = K(w1))));;$

$axiom\ bp : \forall w.(Bp(w) \equiv \exists w1.(A(w, w1, P, 0) \wedge \neg(K(w) = K(w1))));;$

$axiom\ b :$

$$\forall w.(B1(w) \equiv \forall w1.(A(w, w1, S, 0) \supset Bp(w1)));$$
$$\forall w.(B2(w) \equiv \forall w1.(A(w, w1, P, 1) \supset K(w) = K(w1)));;$$

$axiom\ c :$

$$\forall w.(C1(w) \equiv Bs(w) \wedge B1(w));$$
$$\forall w.(C2(w) \equiv C1(w) \wedge B2(w));;$$

The previous axioms were just definitions. Now we have the information coming from the dialog. SKNPK stands for "S Knows that P does Not Know" and the axiom asserts this about the real world at time 0.

$axiom\ sknpk : B1(RW);;$

NSK stands for "S doesn't know", and the axiom asserts this about the real world.

$axiom\ nsk : Bs(RW);;$

PK stands for "P knows the numbers", and the axiom asserts this about the real world at time 1.

$axiom\ pk : B2(RW);;$

SK tells us that at time 2, Mr. S knows the numbers.

$axiom\ sk : \forall w.(A(RW, w, S, 2) \supset K(RW) = K(w));;$

The last two axioms are the learning axioms. LS tells us that everyone learns by time 1 that Mr. S knew at time 0 that Mr. P didn't know the numbers, and he didn't know them either. LP tells us that everyone learns by time 2 that Mr. P knew the numbers by time 1.

$axiom\ lp :$

$\forall ww1.(A(RW, w, SP, 1) \supset (A(w, w1, P, 1) \equiv A(w, w1, P, 0) \wedge C1(w1)));;$

$axiom\ ls :$

$\forall ww1.(A(RW, w, SP, 2) \supset (A(w, w1, S, 2) \equiv A(w, w1, S, 1) \wedge C2(w1)));;$

Ma Xiwen's proof was carried out using a sequence of lemmas. The first lemma shows the essential equivalence of a condition on possible worlds with a condition on pairs.

Lemma 1. $\forall wk.(A(RW, w, SP, 0) \wedge k = K(w) \supset (Qs(k) \equiv Bs(w)))$.

Lemma 2. $\forall wk.(A(RW, w, SP, 0) \wedge k = K(w) \supset (Qp(k) \equiv Bp(w)))$.

Lemma 3. $\forall wk.(A(RW.w.SP, 0) \wedge k = K(w) \supset (Q1(k) \equiv B1(w)))$.

Lemma 4. $\forall wk.(A(RW, w, SP, 0) \wedge k = K(w) \supset (R1(k) \equiv C1(w)))$.

Lemma 5. $R2(k0)$.

Lemma 6. $\forall wk.(A(RW, w, SP, 0) \wedge k = K(w) \supset (R2(k) \supset C2(w)))$.

Lemma 7. $Q3(k0)$.

Main Theorem. $R3(k0)$.

$R3(k0)$ is a purely arithmetic condition on the pair of numbers. From an axiom asserting that $k0$ is a pair of numbers in the interval $2 \leq x \leq 99$ and Peano's axioms for arithmetic, the numbers can be proved to be 4 and 13. Alternatively, $R3(k0)$ can be translated into a computer program for computing the numbers. Most people who solve the problem write such a program.

4 References

Weyhrauch, Richard W. (1977). *FOL: A proof checker for first-order logic* (Stanford Artificial Intelligence Laboratory Memo AIM–235.1). Stanford University, Stanford.

/@steam.stanford.edu:/u/ftp/jmc/puzzles.tex: begun 1996 May 14, latexed 2007 Nov 19 at 12:45 p.m.