PROCEEDINGS OF THE INTERNATIONAL CONFERENCE
ON FIFTH GENERATION COMPUTER SYSTEMS 1988,
edited by ICOT.

REASONING ABOUT KNOWLEDGE AND IGNORANCE

Luigia Carlucci Aiello, Daniele Nardi, Marco Schaerf

Dipartimento di Informatica e Sistemistica
Università di Roma "La Sapienza"
Via Buonarroti 12, I-00185 Roma,

Abstract

In this paper we discuss reasoning about reasoning in a multiple agent situation. We consider agents that are perfect reasoners, loyal, and that can take advantage of both the knowledge and ignorance of other agents. The knowledge representation formalism we use is (full) first order predicate calculus, where different agents are different theories and reasoning about reasoning is realized via a meta-level representation of knowledge and reasoning. The framework we provide is pretty general: we illustrate it by showing a solution to the three wisemen puzzle.

The solution we present relies on an appropriate organization of each wiseman knowledge into units: his own knowledge about the world and his knowledge about other wisemen are units containing object-level knowledge; a unit containing meta-level knowledge embodies the reasoning about reasoning features and realizes the link among the units.

1 INTRODUCTION

One of the goals of Artificial Intelligence is the construction of an artificial agent that autonomously behaves in the real world. To this end, a major issue, among the many relevant ones, is the understanding of what it means for an intelligent agent, to gather knowledge via observation of the world and perception of utterances from other companion agents. Of course, these forms of observation entail an acquisition of knowledge on the side of the agent, that henceforth updates its knowledge base and is enabled to use the newly acquired knowledge in its own reasoning activity.

This process implies the existence of a knowledge representation formalism that the agent uses to build its own symbolic representation of the world, including the fellow agents, and a deductive apparatus. An intelligent agent has to be capable of reasoning using its own knowledge, reasoning about its own knowledge and reasoning activity, and about other agents' knowledge and reasoning.

The main issue we address in the paper is the interaction process that happens among agents: each agent has its own view of the surrounding world, the other agents being part of such a world, and can reason about it. Agents "listen" to one another and are able to use conclusions drawn by others in a non trivial way; that is to say, an agent, by knowing that another agent has a viewpoint on the world different from its own, gathers information about the world seen from this other viewpoint.

The situation we have depicted so far is not one of cooperative agents that, for instance, aim to achieve a common goal, because in that case one should imagine that the communication among the agents is perfect and total, so as to make almost pointless to take into consideration the existence of various viewpoints. Conversely, this situation can be described as one of loyal agents that are perfect reasoners: when asked a question, an agent tells the truth, to the best of its knowledge. An agent does not cheat both on the nature of its conclusion, and on its ability to draw one (i.e. it does not say "I don't know" if it does know). It is of particular interest that an agent can reason on another agent's conclusions, hence coming to know even things that the other agent has not explicitly communicated, and to reason about "I don't know" conclusions of another agent, hence gathering knowledge even from another agent's ignorance.

The problem described so far is very well illustrated by a simple puzzle, known in the AI community as the three wisemen problem: a king wishing to know which of three wisemen is the wisest, puts a white hat on each of their heads, and tells them that the hats are black or white and that at least one of them is white. Each wiseman can see the other wisemen's hats but not his own. The king asks the first wiseman to tell the color of his hat. The man answers that he does not know. The second man gives the same answer to the same question. The third one instead answers that the color of his hat is white.

This puzzle exemplifies the problem of loyal agents that are not cooperative: they simply provide an answer to the question they have been asked, without providing explanations on their reasoning process that could help the other agents.

In the paper, we propose an architecture for representing situations with multiple agents, each one being able to reason about the world, to take advantage of what it "hears" from other agents, to acquire knowledge relevant to its own goals and reason about it.

The knowledge representation formalism chosen for our solution is (full) first order predicate calculus, where different agents are different theories and reasoning about reasoning is realized via a meta-level representation of knowledge and reasoning. The framework we provide is pretty general: we illustrate it in Section 2 by showing a solution to the three wisemen puzzle.

The solution we present, which improves on the one we illustrated in [4], relies on an appropriate organization of each wiseman knowledge into units: his own knowledge about the world and his knowledge about other wisemen are units containing object-level knowledge; a unit containing meta-level knowledge embodies the reasoning about reasoning features and realizes the link among the units.

The proof we present for the three wisemen puzzle has been machine checked using a version of Weyhrauch's FOL system [28]. The relevant features of FOL are sketched in Section 3, while the description of the proof is reported in Section 4.

The three wisemen puzzle has often been a vehicle for discussing reasoning about reasoning. In Section 5 we relate our solution to the other ones proposed in the literature, in particular those realized in Prolog [8,21], in OMEGA [5] and with modal logics [10,17], pointing out the advantages of our approach.

2 FORMALIZATION IN FIRST ORDER LOGIC

In this section we provide a formalization of the three wisemen puzzle in first order logic, as a set of theories and meta-theories (i. e. theories about theories), which are interrelated by means of linking rules.

More precisely, we describe the knowledge of each wiseman as a structure composed by three object-level theories, and a meta-level theory. We call this structure, which is shown in fig. 1, an *agent*. The first object-level theory, called $ownT_i$, contains the agent's own knowledge, and the remaining two, called T_i^j, contain the knowledge that the agent A_i knows to be owned by the other agents (specified by the superscript j). In addition, each agent has a meta-theory, called MT_i, where the (meta-) knowledge for reasoning about the other agents' knowledge and reasoning, is represented.

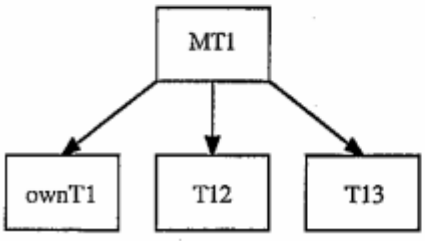

Fig.1 The architecture of the agent A1

The relationship between meta and object-level theories is established by the following reflection principle [11]: if $THEOREM('w')$ holds in the meta theory, w can be derived in the theory, and viceversa. Conventionally, 'n' is the name in the meta-theory for the object n of the theory. In the direct formulation such principle can be interpreted as an auxiliary inference rule, which allows to derive a logical consequence of the theory by applying a specialized deduction procedure defined in the meta-theory.

The agents' meta-theory is connected with each of the object-level theories, and the reflection principle holds between each pair $< MT_i, T_i >$. In fact, the meaning of the connection is very different in the case of the $ownT_i$ and T_i^j: in the first one, it allows to derive the agent's own beliefs, while in the case of T_i^j, it allows to draw the conclusions that other agents could possibly derive. Such conclusions are not necessarily believed by the agent, and may even be inconsistent with the agent's beliefs.

2.1 The object-level axioms

The object-level theories contain axioms representing the knowledge available (or known to be available) to the wisemen. Each of them contains the axiom expressing the constraint that at least one of the wisemen has a white hat:

$atleast : white1 \vee white2 \vee white3$

where the constant $white1$, for example, is interpreted as: the color of the hat of the first wiseman is white. We provide names to axioms for later reference.

Furthermore, each $ownT_i$ represents what the wisemen can see. For example, $ownT_1$ also contains the axioms:

$white2$
$white3$

The knowledge about the other wisemen, represented by T_i^j, is initially limited to the axiom $atleast$; in fact this is the only fact known by all the agents and that all the agents know to be common knowledge.

2.2 The meta-level axioms

In the meta-theory we formalize the reasoning about knowledge and reasoning.

The predicate $THEOREM(T,'w')$ is used to denote that a formula w is a logical consequence of the axioms of the theory T. In particular, according to the reflection principle, if the formula w is asserted in the theory T, the formula $THEOREM(T,'w')$ holds in the meta-theory.

The predicate $KNOWS(A,'w')$ describes the agent's knowledge relative to the other agents; $KNOWS$ never refers to the agent's own knowledge,which is explicitly represented in its $ownT_i$.

In the following, we introduce the meta-level axioms, with

the convention that all the variables refer to meta-theory symbols, and that the subscript i means that the axiom is in the meta-theory of each agent, with the occurrences of i replaced by the agent's number, 1, 2 and 3 respectively.

$$knows : \forall jw.\ KNOWS(A_j, w) \wedge \neg THEOREM(T_i^j, mknot(w)) \Rightarrow THEOREM(T_i^j, w)$$

This axiom links the knowledge about other agents, expressed by the predicate $KNOWS$, to the corresponding theory T_i^j; it is used when the reasoning about other agents' knowledge leads to conclusions that can be explicitly represented in the T_i^j, and used for future derivations.

$$confidence : \forall jw.KNOWS(A_j, w) \wedge \neg THEOREM(ownT_i, mknot(w)) \Rightarrow THEOREM(ownT_i, w)$$

This axiom is used when an agent comes to believe the consequences that it has drawn reasoning about another agent's knowledge. Both the axiom *knows* and *confidence* have the form of the axiom for necessity in modal systems, $Ka \Rightarrow a$. In fact, they state more specific properties of reasoning about reasoning: the former is needed to maintain the representation of other agents' knowledge; the latter is used to infer facts known from other agents' reasoning.

$$reason : \forall jw1w2.\ \neg KNOWS(A_j, w1) \wedge CANPROVEIF(w2, T_i^j, w1) \wedge ACCEPTVIEW(A_j, w2) \Rightarrow KNOWS(A_j, mknot(w2))$$

The formula $\neg KNOWS(A_j, w1)$ is used to represent the "I don't know" answer, given by the agent A_j, when asked about the color of his hat. $CANPROVEIF(w2, T_i^j, w1)$ is interpreted as: $w1$ can be deduced in the theory T_i^j extended by the formula $w2$. In this case the notion of provability refers to a theory, which is not explicitly represented within the system, and therefore it is not directly expressed via the predicate $THEOREM$.

The axiom $reason$[1] defines a way of reasoning by contradiction upon the deductions that can be performed by other agents. For instance the reasoning of the second wiseman about the first one, is the following: "If the first wiseman does not know $white1$ and I know that he should be able to prove $white1$ if he sees two black hats then, I can deduce that he knows they aren't both black".

$$\neg KNOWS(A_j, w1) \wedge CANPROVEIF(w2, T_i^j, w1)$$

[1] This fact can actually be derived from more primitive statements about the completeness of particular subtheories, as it is shown in [4]. Due to the lack of space the details are not presented here, hence we introduce it as an axiom.

Is the condition that implies the non provability of $w2$ in T_i^j, because T_i^j is consistent and faithfully, even though not completely, represents the knowledge of A_j. The definition of $reason$ requires the completeness of the theory T_i^j, or close world assumption, since the non provability of $w2$ leads to the assumption $\neg w2$.

In fact, the theories representing the knowledge of the wisemen are not complete because, for example, in $ownT_1$ we can prove neither $white1$ nor $\neg white1$; but the condition on the completeness of the theories can, in this case, be weakened by verifying that $w2$ expresses an acceptable point of view of the agent A_i. This allows to avoid circularities: the formula w should not refer to the agent whose reasoning is being considered. For example, the reasoning done by A_2 about the hypothetical reasoning of A_1 can not be based on the hypothesis that A_1 knows the color of his own hat.

$$same : \forall w.\ KNOWS(A_1, w) \wedge SAMEVIEW(A_2, A_1) \Rightarrow THEOREM(T_3^2, w)$$

The axiom *same* is necessary in order for A_3 to reason about the reasoning that A_2 has performed upon the "I don't know" answer given by A_1. In fact A_3 reaches the same conclusions as A_2, when reasoning about A_1; therefore, A_3 can assume that A_2 has reasoned upon A_1. The condition $SAMEVIEW(A_2, A_1)$ reads as: I have the same view as A_2 upon A_1. This is one of the hypotheses of the puzzle. Actually the axiom *same* can be made available also in A_1 and in A_2, relatively to their views of the other agents, we do not introduce them since they are not needed in the solution of this puzzle.

Any axiomatization based on a language allowing self-referenciality poses the problem of consistency. For a full treatment of this issue see [24,25], and the approach discussed in [9]. In the present formalization meta-level axioms always refer to formulas of the object-level theory. Therefore the problems of inconsistency caused by the self-references do not arise.

2.3 The proof

In this section we present the proof, which provides the solution to the puzzle. The style of the presentation is informal; a complete formalization in the FOL system is described in the next sections. The answer to the question "which is the color of your hat?", posed to the first wiseman, is simply obtained by trying to prove the predicate $THEOREM(ownT_1, 'white1')$. Since the proof fails, A_2 and A_3 get to know $\neg KNOWS(A_1, 'white1')$.

Note that, in order to increase the readability of the proof, we use object-level expressions within quotes to stand for meta-level names of the corresponding object-level formulas. Hence, for instance, $'\neg(\neg white2 \wedge \neg white3)'$ stands

for $mknot(mkand(mknot(white2), mknot(white3)))$, which is the explicit meta-level denotation for the object-level expression $\neg(\neg white2 \wedge \neg white3)$.

The second wiseman reasons as follows.

1 - $\neg KNOWS(A_1, 'white1')$

2 - $KNOWS(A_1, '\neg(\neg white2 \wedge \neg white3)')$
by the axiom *reason* applied to A_1 and the wffs $'white1'$, and $'\neg white2 \wedge \neg white3'$. The condition $CANPROVEIF('\neg white2 \wedge \neg white3', T_2^1, 'white1')$ can be easily verified, as well as the condition $ACCEPTVIEW(A_1, '\neg white2 \wedge \neg white3')$.

3 - $THEOREM(T_2^1, '\neg(\neg white2 \wedge \neg white3)')$
by the axiom *knows*.

4 - $THEOREM(ownT_2, '\neg(\neg white2 \wedge \neg white3)')$
by the axiom *confidence*.

At this point the proof of $THEOREM(ownT_2, 'white2')$ is attempted, and A_2 answers "I don't know". A_3 notes that $\neg KNOWS(A_2, 'white2')$, and its proof is as follows:

1 - $\neg KNOWS(A_1, 'white1')$

2 - $THEOREM(ownT_3, '\neg(\neg white2 \wedge \neg white3)')$
obtained by the same reasoning done by A_2.

3 - $THEOREM(T_3^2, '\neg(\neg white2 \wedge \neg white3)')$
by the axiom *same*, where the condition $SAMEVIEW(A_2, A_1)$ is verified by hypothesis.

4 - $\neg KNOWS(A_2, 'white2')$

5 - $KNOWS(A_2, '\neg(\neg white3)')$
by the axiom *reason* applied to A_2, and the wffs $'white2'$, and $'\neg white3'$. The condition $CANPROVEIF('\neg white3', T_3^2, 'white2')$ can be easily verified, as well as $ACCEPTVIEW(A_2, '\neg white3')$.

6 - $THEOREM(T_3^2, '\neg(\neg white3)')$
by the axiom *knows*.

7 - $THEOREM(ownT_3, '\neg(\neg white3)')$
by the axiom *confidence*.

At this point $THEOREM(ownT_3, 'white3')$ can be proved by A_3.

3 THE META-LEVEL ARCHITECTURE

Several meta-level architectures have been proposed so far, the reader can find references in [2,3,7,13,20]. Our solution to the three wisemen puzzle has been carried out within the FOL system [28]. Although much of the discussion about the use of a meta-level architecture for reasoning about reasoning and knowledge applies independently of the particular system, the FOL meta-level architecture embodies several distinguished features, which are presented in this section.

In the three wisemen puzzle we need to represent the knowledge of the three wisemen as well as their reasoning. In FOL the knowledge base of each wiseman is represented by a base level or *object-level context*. The reasoning of the wisemen is represented at the meta-level by means of *meta-contexts*. Contexts and meta-contexts have the same structure: in fact the attribute meta simply emphasizes that a context contains a description of another context. It is therefore possible to build hierarchies of contexts where meta-contexts are described by meta-meta-contexts and so on.

A context provides a finite representation of a first order theory. In this theory the knowledge can be expressed in the form of axioms in a first order logic language, and by defining a partial model of the theory, called *simulation structure*. The simulation structure associates an interpretation, expressed as functions and data structures of a LISP-like applicative language, with some of the symbols of the theory .

Given this twofold form for expressing knowledge, the deduction must take into account both specifications, that in FOL are termed *syntactic* and *semantic*. In fact FOL, provides different types of reasoners, called *evaluators*, which allow to draw conclusions from the syntactic and semantic knowledge, separately. Syntactic evaluators implement standard deduction procedures for first order logic, such as equation rewriting and tautology checking. The semantic evaluator checks for satisfiability in the partial model defined by the interpretations associated with the symbols of the theory.

A very powerful reasoning tool is the *FOL evaluator*, which makes use of both syntactic and semantic knowledge, by combining the semantic evaluation with syntactic rewriting. A detailed description of the FOL evaluator can be found in [22].

A meta-context, as any other context, has both a syntactic and a semantic component, but since it refers to parts of the system itself, the data structures used in the simulation structure can be those actually implementing the system. For instance, in the specification of a meta-context , a constant symbol of sort wff can be given an interpretation in the simulation structure by means of the data structure internally created by the system to represent a context. Other meta-level architectures define different naming relations between the symbols of the meta-level and the objects representing the base level (see for example [7,14]). FOL is unique in that it characterizes the data structures manipulated at the object-level as the partial model of the meta-theory represented by the meta-context. Therefore the relation between the object-level and a meta-level is not just a naming relation, but the object-level data structure are given a cognitive account in terms of the interpretations of the symbols in the meta-context.

FOL contexts can be linked through the simulation structure and the reflection mechanism. This allows to generate a proof step in the object context by making a derivation in the meta-context. The consistency of the result of such an inference is ensured whenever the relationship between the meta-context and the object context is well defined (see [22] and [28] for a deeper discussion).

4 THE FOL AXIOMATIZATION

In this section we present the solution implemented in the FOL system, it closely follows the one just described. Each agent A_i is realized within FOL as a system S_i consisting of a meta-level context and three object-level contexts, corresponding to $ownT_i$ and the two T_i^j, respectively. In the FOL axiomatization we represent the three object-contexts in the simulation structure of the corresponding meta-context. We name the three meta-contexts as $meta1, meta2$ and $meta3$.

4.1 Building the meta-contexts

We start by defining the first meta-context, $meta1$.

NAMECONTEXT $meta1$;

At the meta-level we have the definitions of the data structures for manipulating contexts and well formed formulas (wffs). The definition of a data structure is given by declaring a symbol to denote the sort of the elements of its domain, symbols of constants representing domain elements, function symbols operating on them, and variable symbols ranging over the defined sort.

DECLARE SORT $AWFF\ WFF$;
MOREGENERAL $WFF < AWFF >$;
DECLARE INDVAR $w\ w1\ w2[WFF]$;
DECLARE INDCONST $white1\ white2\ white3$
$[AWFF]$;

For example here we define the sort for wffs and atomic wffs (awffs); the command MOREGENERAL builds a lattice of sort definitions, in this case every awff is a wff as well; w is a variable symbol ranging over WFF.

DECLARE FUNCONST $mkand\ mkor\ mkimp$
$(WFF, WFF) = WFF[INF]$;
DECLARE FUNCONST $mknot(WFF) = WFF$;

The functions $mkand, mkor, mkimp$ and $mknot$ are the wff constructors we need, INF means we use these functions in infix mode. These definitions are given in the simulation structure by associating with each symbol the LISP-function implementing it. The names of the functions used in the simulation structure have the prefix I standing for interpretation.

ATTACH $mkand$ TO I-$mkand$;
ATTACH $mkor$ TO I-$mkor$;
ATTACH $mkimp$ TO I-$mkimp$;
ATTACH $mknot$ TO I-$mknot$;

DECLARE INDCONST $atleast\ black23[WFF]$;
LET $atleast = (white1\ mkor\ white2)\ mkor\ white3$;
LET $black23 = mknot(white2)\ mkand\ mknot(white3)$;

The command LET links the constant symbol on the left side of equality with the evaluation, in the simulation structure, of the expression on the right side. In this case, the constant $atleast$ is defined in the simulation structure in terms of a formula representing the fact that at least one hat is white and $black23$ by the formula representing the fact that the second and third wisemen have black (not white) hats.

Below we give the definition of the data structures for systems and contexts.

DECLARE SORT $CONTEXT\ SYSTEM$;
DECLARE INDCONST $S2\ S3[SYSTEM]$;
DECLARE INDCONST $C0\ emptyC\ ownC1\ C12\ C13$
$[CONTEXT]$;
DECLARE INDVAR $xC[CONTEXT]$;
DECLARE FUNCONST $declsentconst$
$(AWFF, CONTEXT) = CONTEXT$;
DECLARE FUNCONST $addfact$
$(WFF, CONTEXT) = CONTEXT$;

The function $declsentconst$ takes an atomic wff and a context and returns a new context with the declaration of a sentential constant. This is the meta-level description of the effect of the command DECLARE SENTCONST issued at the object-level. The function $addfact$ takes a wff and a context and returns a new context with a new fact in it. Both the functions and the constant $emptyC$ are defined in terms of the corresponding element of the simulation structure. The constant $emptyC$ is associated with the LISP-structure representing an empty context.

ATTACH $emptyC$ TO $emptyC$;
ATTACH $declsentconst$ TO I-$declsentconst$;
ATTACH $addfact$ TO I-$addfact$;

The objects defined in the simulation structure for constant symbols have the same name as the symbol. In the following, the commands for the definition of the simulation structure will be omitted.

In the axiomatization we need to use some predicates to formalize the hypotheses of the puzzle and the relations between contexts and systems.

DECLARE PREDCONST $CANPROVE$
$(CONTEXT, WFF)$;
DECLARE PREDCONST $ACCEPTVIEW$

$(SYSTEM, WFF)$;
DECLARE PREDCONST $KNOWS$
$(SYSTEM, WFF)$;
DECLARE PREDCONST $SAMEVIEW$
$(SYSTEM, SYSTEM)$;
DECLARE PREDCONST $HOLDS$
$(CONTEXT, WFF)$;

The first two predicates are defined in the simulation structure, $CANPROVE$ checks if the wff can be derived in the context; it is implemented via the system procedure for checking tautologies. $ACCEPTVIEW$ checks if a wff is an acceptable view for a system; a wff is an acceptable view for a system if it does not contain references to the color of the hat of the wiseman represented by the system, this because a wiseman cannot see his own hat. For example, $ACCEPTVIEW(S1, w)$ is true if $white1$ does not occur in w. The predicates $KNOWS$ and $SAMEVIEW$ are exactly equivalent to the ones defined in Section 2. In order to express the linking rule between the meta-context and the object-contexts we define the predicate $HOLDS$ and the function $updatectx$. To avoid inference rules that, when trying to prove w, look from the object-level context at the meta-level for a proof of $HOLDS(C,' w')$, we explicitly assert at the object-level any fact w which corresponds to the argument of $HOLDS(C,' w')$. Therefore whenever $HOLDS(C,' w')$ is derived at the meta-level, the context C is updated through the command LET and the function $updatectx$.

DECLARE FUNCONST $updatectx$
$(CONTEXT, WFF)$;
AXIOM $update : \forall w\, xC.\ updatectx(xC, w) =$
$IF\ HOLDS(xC, w)$
$THEN addfact(w, xC)$
$ELSE\ xC$;

The declarations in $meta2$ and $meta3$ are identical to those just shown, but for the subscripts that decorate systems and contexts.

4.2 Building the object-contexts

The object-level contexts are built by means of the command LET, which constructs a data structure and associates it with the symbol of the meta-context. To shorten the construction, we first define a context $C0$, and then build all the other contexts on it.

LET $C0 = declsentconst(white3,$
$declsentconst(white2,$
$declsentconst(white1, emptyC)))$;

After the execution of this command, the constant symbol $C0$ of the meta-context has an interpretation in the simulation structure as the data structure representing a context with three sentential constants (corresponding to $white1$, $white2$, $white3$).

LET $ownC1 = addfact(atleast,$
$addfact(white3,$
$addfact(white2, C0)))$;

LET $C12 = addfact(atleast, C0)$;

LET $C13 = addfact(atleast, C0)$;

The contexts $ownC1, C12, C13$ describe the fact that the first wiseman knows that the hats of the two other wisemen are white and that at least one of the three hats is white. He also knows that the other two wisemen know that at least one hat is white. Analogous constructions are performed to set up the object-contexts for the systems $S2$ and $S3$.

4.3 The meta-level axioms

The meta-level axioms are the same used in the solution described in Section 2.

AXIOM $knows2 : \forall w.KNOWS(S2, w) \wedge$
$\neg CANPROVE(C12, mknot(w))$
$\Rightarrow HOLDS(C12, w)$;
AXIOM $knows3 : \forall w.KNOWS(S3, w) \wedge$
$\neg CANPROVE(C13, mknot(w))$
$\Rightarrow HOLDS(C13, w)$;
AXIOM $confidence2 : \forall w.KNOWS(S2, w) \wedge$
$\neg CANPROVE(ownC1, mknot(w))$
$\Rightarrow HOLDS(ownC1, w)$;
AXIOM $confidence3 : \forall w.KNOWS(S3, w) \wedge$
$\neg CANPROVE(ownC1, mknot(w))$
$\Rightarrow HOLDS(ownC1, w)$;
AXIOM $reason2 : \forall w1\, w2.$
$\neg KNOWS(S2, w1) \wedge$
$CANPROVE(addfact(w2, C12), w1) \wedge$
$ACCEPTVIEW(S2, w2)$
$\Rightarrow KNOWS(S2, mknot(w2))$;
AXIOM $reason3 : \forall w1\, w2.$
$\neg KNOWS(S3, w1) \wedge$
$CANPROVE(addfact(w2, C13), w1) \wedge$
$ACCEPTVIEW(S3, w2)$
$\Rightarrow KNOWS(S3, mknot(w2))$;

This completes the axiomatization for $meta1$. In the other meta-contexts the axioms are the same except for the names of contexts and systems. In addition in $meta3$ we have the axiom $same$.

AXIOM $same : \forall w.KNOWS(S1, w) \wedge$
$SAMEVIEW(S2, S1)$
$\Rightarrow KNOWS(S2, w)$;

4.4 The proof steps

FOL is an interactive proof checker, and every proof step is accomplished by the application of an inference rule, such as the natural deduction rules, as well as powerful automatic proof checking procedures (e. g. the evaluator). The proof presented below requires context switching in order to simulate the reasoning carried on by the agents. Here we give an outline of the proof of the puzzle done interactively; a proof strategy for automatic proof generation could be devised at the meta-meta-level. The proof is organized along the reasoning that the three wisemen perform when attempting to answer the question of the king "which is the color of your hat?". (FOL:: is the FOL prompt, while the generated facts are progressively numbered).

FOL:: SWITCHCONTEXT $meta1$;

FOL:: EVAL $CANPROVE(ownC1, white1)$;
1 - $\neg CANPROVE(ownC1, white1)$

The first wiseman tries to prove $white1$ in $ownC1$, but fails. The second wiseman hears the answer of the first one and tries to understand why he could not answer.

FOL:: SWITCHCONTEXT $meta2$;

FOL:: ASSUME $notKNOWS(S1, white1)$;
1 - $\neg KNOWS(S1, white1)$

FOL:: EVAL $CANPROVE(addfact(black23, C21),$
$white1)$;
2 - $CANPROVE(addfact(black23, C21), white1)$

FOL:: EVAL $ACCEPTVIEW(S1, black23)$;
3 - $ACCEPTVIEW(S1, black23)$

By modus ponens applied to the axiom $reason1$, instantiated with $white1$ and $black23$, we obtain:

4 - $KNOWS(S1, mknot(black23))$

FOL:: EVAL $CANPROVE(C21,$
$mknot(mknot(black23)))$;
5 - $\neg CANPROVE(C21, mknot(mknot(black23)))$

By modus ponens applied to the axiom $knows1$, instantiated with $mknot(black23)$, we obtain:

6 - $HOLDS(C21, mknot(black23))$

At this point we update $C21$; this is done with the function $updatectx$ and the command LET, using the axiom $update$.

FOL:: LET $C21 = updatectx(C21, mknot(black23))$
BY $\{update\}$;

By modus ponens applied to the axiom $confidence1$, instantiated with $mknot(black23)$, we obtain:

7 - $HOLDS(ownC2, mknot(black23))$

FOL:: LET $ownC2 = updatectx(ownC2,$
$mknot(black23))$ BY $\{update\}$;

The second wiseman updates his knowledge and the knowledge he has about the first one. Now he tries to prove $white2$ with his own knowledge.

FOL:: EVAL $CANPROVE(ownC2, white2)$;
8 - $\neg CANPROVE(ownC2, white2)$

The second wiseman has not enough knowledge to prove $white2$, therefore his answer is "I don't know". The third wiseman hears the answer of the second one and starts his reasoning.

FOL:: SWITCHCONTEXT $meta3$;

FOL:: ASSUME $SAMEVIEW(S2, S1)$;
1 - $SAMEVIEW(S2, S1)$

FOL:: ASSUME $\neg KNOWS(S1, white1)$;
2 - $\neg KNOWS(S1, white1)$

FOL:: ASSUME $\neg KNOWS(S2, white2)$;
3 - $\neg KNOWS(S2, white2)$

With the same reasoning of the second wiseman about the first one the following steps are generated.

4 - $KNOWS(S1, mknot(black23))$

5 - $HOLDS(C31, mknot(black23))$

FOL:: LET $C31 = updatectx(C31, mknot(black23))$
BY $\{update\}$;

6 - $HOLDS(ownC3, mknot(black23))$

FOL:: LET $ownC3 = updatectx(ownC3,$
$mknot(black23))$ BY $\{update\}$;

The third wiseman, before reasoning about the answer of the second one, needs to update his view of the knowledge of the second one. By modus ponens applied to the axiom $same$, instantiated with $mknot(black23)$, we obtain:

7 - $KNOWS(S2, mknot(black23))$

By modus ponens applied to the axiom $knows2$, instantiated with $mknot(black23)$, we obtain:

8 - $HOLDS(C32, mknot(black23))$

FOL:: LET $C32 = updatectx(C32,$
$mknot(black23))$ BY $\{update\}$;

Now, after the update of $C32$, the third wiseman can reason about the answer given by the second one. By modus ponens applied to the axiom $reason2$, instantiated with $white2$ and $mknot(mknot(white3))$, we obtain:

9 - $KNOWS(S2, mknot(mknot(white3)))$

By modus ponens applied to the axiom $knows2$, instantiated with $mknot(mknot(white3))$, we obtain:

10 - $HOLDS(C32, mknot(mknot(white3)))$

FOL:: LET $C32$ = $updatectx(C32,$
$mknot(mknot(white3)))$ BY $\{update\}$;

By modus ponens applied to the axiom $confidence2$, instantiated with $mknot(mknot(white3))$, we obtain:

11 - $HOLDS(ownC3, mknot(mknot(white3)))$

FOL:: LET $ownC3$ = $updatectx(ownC3,$
$mknot(mknot(white3)))$ BY $\{update\}$;

At this point, the third wiseman tries to answer the question of the king, and finally succeeds.

FOL:: EVAL $CANPROVE(ownC3, white3)$;
12 - $CANPROVE(ownC3, white3)$

5 DISCUSSION

We have presented an architecture for reasoning about reasoning in a multi-agent scenario; it is based on a representation of knowledge in first order logic and has been experimented using Weyhrauch's FOL for solving the three wisemen puzzle. In our solution, each agent is represented as a FOL system consisting of a meta-level context, where the knowledge relative to reasoning about reasoning is represented, and three object-level contexts, where the agent's own knowledge and its understanding of the other agents' knowledge is represented.

This architecture is more general and flexible than the one we presented in [4]. In that proposal, the knowledge relative to reasoning about reasoning of all the agents is represented in a single meta-context and the object-level knowledge of each agent is represented in a separate context, divided into a "private" and a "visible" part. The private knowledge is only accessible to the agent that owns it, the visible knowledge is accessible also to the other agents.

Actually, the generality of the architecture proposed in the present paper is not fully exploited in the solution of the three wisemen puzzle, because the knowledge bases of the three wisemen are identical, but for the constant symbols denoting the various agents. Nevertheless, it can also be used in formalizing problems where each agent has his own meta- knowledge, different from the other ones, and has different views on the knowledge of the others.

The three wisemen puzzle has often been used as an example to illustrate various approaches to the formalization of reasoning about reasoning. In the following we compare our proposal with others that recently appeared in the literature.

Modal logic has been proposed as the representation formalism for problems related to knowledge and belief; both Konolige [12,17] and Farinas del Cerro [10], have solved the three wisemen puzzle using a modal system.

The modal approach has the merit of identifying the essential features of the problem; the logic chosen is specific to a limited class of problems and it is not evident how to deal with situations where not all the agents are perfect reasoners or use different deduction strategies. A modal solution would require a different set of modal operators, which practically means to use a different system. A meta-level architecture allows to specify modal operators as meta-level knowledge associated with different agents, therefore providing a great degree of flexibility. Modal operators are expressed as meta-level predicates: once the problem has been formalized it reduces to a set of contexts linked to one another, and the reasoning process is performed by means of the standard deduction mechanisms of first order logic.

Our point in favour of a meta-level architecture based on first order logic is reinforced by considerations about other representation problems such as: inference control, non-monotonic and default reasoning, self- evaluation and self-modification, etc. (see [2] for a general discussion, [15] and [19] for examples of use of a meta- level architecture for non-monotonic reasoning and belief revision, respectively). A meta-level architecture allows to build a system where such different issues can be dealt with in a unified framework.

The three wisemen puzzle has been proposed as an example of the use of knowledge base management facilities described by Coscia et al. in [8]. Their solution presents a structure of meta-descriptions, based on different types of links that can be established among theories. In particular, the common knowledge can be inherited by the theories representing the three wisemen. The proposed deduction is obtained by a meta-level program written in an extended PROLOG. The code implicitly deals with many aspects of the problem that are instead expressed explicitly in our solution. For example, the implications of the "I don't know" answers given by the wisemen are not explicitly deduced by the others, but simply asserted in each theory; furthermore, inside each theory there is no distinction between what is own knowledge and knowledge about the others.

An analogous solution based on Horn Clause Logic has been proposed by Nait Abdallah [21], who, instead of relying on an amalgamation between object-level and meta-level knowledge, uses special operators to deal with local-

ity of proofs. His notion of locality somehow corresponds to our contexts.

Another solution to the puzzle has been given within the framework of the OMEGA system [5]. In this case, the knowledge of the three wisemen is represented by three viewpoints, and the deduction is done according to a set of rules that specify general properties of the viewpoints. This provides a clean specification, but we think there may be a problem with the application of the Indirect Proof axiom, which is a property of the viewpoints corresponding to the closed world assumption. In fact, the axiom should only be applied to what we called acceptable views, i.e. the second wiseman cannot assume that the first one knows the color of his hat, because otherwise an inconsistency is obtained.

Our solution to the three wisemen puzzle brings up a general architecture for representing agents capable of reasoning about other agents' knowledge and reasoning. In particular, the axiom *confidence* captures the notion of loyal agent, while the axiom *reason* describes a method for reasoning by contradiction.

We believe that a very important notion, not yet fully exploited, is that of agent. Agents have been characterized as a set of theories and meta-theories linked together, and implemented using *FOLsystems*. While in the solution of the puzzle we explicitly referred to agent/system components, a more general framework should allow to reason about agents as a whole. In particular, this would allow us to provide a suitable formalization of properties about agents such as the one expressed by the axiom *same*; in such a framework this axiom can be derived from more primitive statements about agents. In this case, a meta-meta-level is needed in order to specify the structure of agents/systems.

Acknowledgments

We acknowledge Bob Kowalski for useful comments on an earlier version of this work.

The work reported here has been partially supported by Ministero Pubblica Istruzione with the project Intelligenza Artificiale:ASSI, ENIDATA, Project COST13 n.21 "Advanced issues in knowledge representation".

Bibliography

[1] Aiello L., "Automatic Generation of Semantic Attachments", *Proc. of AAAI 80*(1980), pp. 90-93.

[2] Aiello L., Cecchi C., Sartini D., "Representation and Use of Metaknowledge", *Proceedings of the IEEE*, 74:10 (1986), pp. 1304-1321.

[3] Aiello L., Levi G., "The Uses of Metaknowledge in AI Systems", *Proc. of ECAI 84*, O'Shea T. (Ed.), North-Holland (1984), pp. 705-717.

[4] Aiello L., Nardi D., Schaerf M., "Yet another solution to the three wisemen puzzle", to appear in *Proc. of ISMIS 88* (1988).

[5] Attardi G., Simi M., "Reasoning across viewpoints", *Proc. of ECAI 84*, O'Shea T. (Ed.), North-Holland (1984), pp. 315-325.

[6] Batali J., "Computational Introspection", MIT AI Memo 701 (1983).

[7] Bowen K. A., Kowalski R. A., "Amalgamating Language and Metalanguage", in *Logic Programming*, Tarnlund (Ed.), Academic Press, New York (1982), pp. 153-173.

[8] Coscia P., Franceschi P., Levi G., Sardu G., Torre L., "Object level reflection of inference rules by partial evaluation", in [20], pp. 313-327.

[9] des Rivieres, J., Levesque H. J., "The consistency of syntactical treatments of knowledge" *Proc. of the 1986 Conference on Theoretical Aspects on Reasoning about Knowledge*, in Halpern J. (Ed.), Morgan Kaufman, (1986), pp. 115-130.

[10] Farinas del Cerro, L., "MOLOG: a system for modal logics", *New Generation Computing*, (1986).

[11] Feferman S., "Transfinite Recursive Progressions of Axiomatic Theories", *Journal of Symbolic Logic*, 27:3 (1962), pp. 259-316.

[12] Geissler C., Konolige K., "A resolution method for quantified modal logics of knowledge and belief", *Proc. of the 1986 Conference on Theoretical Aspects on Reasoning about Knowledge*, in Halpern J. (Ed.), Morgan Kaufman, (1986), pp. 309-324.

[13] Genesereth M. R., "An Overview of Meta-level Architectures", *Proc. of AAAI 83* (1983), pp. 119-124.

[14] Genesereth M. R., Nilsson N, "Fundamentals of Artificial Intelligence", Morgan-Kaufman (1987).

[15] Giunchiglia F.,Weyhrauch R. W., "A Multi-Context Monotonic Axiomatization of Inessential Non-Monotonicity", in [20] (1988), pp 271-285.

[16] Hayes P. J., "In Defense of Logic", *Proc. of IJCAI 77* (1977), pp. 559-565.

[17] Konolige K., "Circumscriptive Ignorance", *Proc. of AAAI 82* (1982), pp. 202-204.

[18] Konolige K., "Belief and Incompleteness", in Hobbs J., Moore R.C. (eds.) *Formal Theories of the Commonsense World*, Ablex Pub. Corp, (1985), pp. 358-403.

[19] Lenzerini M., Nardi D., "Belief revision as meta-reasoning", *Proc. of the Workshop Machine Learning, Meta Reasoning and Logics* Sesimbra, Portugal (1988), pp. 257-263.

[20] Maes P., Nardi D., (Eds.) *Meta-level Architectures and Reflection*, North Holland, 1988.

[21] Nait Abdallah M. A., "Logic Programming with ions", Proc. of the 14th Int. Coll. on Automata Languages and Programming, Ottmann T. (Ed.), LNCS 267, Springer Verlag (1987), pp. 11-20.

[22] Nardi D., "Evaluation and Reflection in FOL", in [20], (1988), pp. 195-207.

[23] Perlis D., "Languages with self-references I", *Artificial Intelligence*, 25, (1985), pp. 301-322.

[24] Perlis D., "Meta in Logic", in [20], (1988) pp. 37-49.

[25] Perlis D., "Languages with self-references II", *Artificial Intelligence*, 34, (1988), pp. 179-212.

[26] Simi M., Motta E., "OMEGA: an integrated reflective framework", in [20], (1988) pp. 209-226.

[27] Smith B. C., "Varieties of Self-Reference", *Proc. of the 1986 Conference on Theoretical Aspects on Reasoning about Knowledge*, in Halpern J. (Ed.), Morgan Kaufman, (1986), pp. 19-43.

[28] Weyhrauch R. W., "Prolegomena to a Theory of Mechanized Formal Reasoning", *Artificial Intelligence*, 13,1 (1980), pp. 133-170.

[29] Weyhrauch R. W., "An Example of FOL Using Metatheory: Formalizing Reasoning Systems and Introducing Derived Inference Rules, *Proc. of CADE 82*, Loveland D. W. (Ed.), LCNS 138, Springer Verlag (1982), pp. 151-158.