Using Abstract Resources to Control Reasoning

Richard W. Weyhrauch
IBUKI Inc., PO Box 1627, Los Altos, CA 94022.
E-mail: `rww@ibuki.com`

Marco Cadoli
Dipartimento di Informatica e Sistemistica, Università di Roma "La Sapienza", Via Salaria 113, I-00198 Roma, Italy.
E-mail: `cadoli@dis.uniroma1.it`

Carolyn L. Talcott
Department of Computer Science, Stanford University, Stanford, CA 94305.
E-mail: `clt@cs.stanford.edu`

(Received ..... ; Accepted in final form .....)

Abstract. Many formalisms for reasoning about knowing commit an agent to be logically omniscient. Logical omniscience is an unrealistic principle for us to use to build a real-world agent, since it commits the agent to knowing infinitely many things. A number of formalizations of knowledge have been developed that do not ascribe logical omniscience to agents. With few exceptions, these approaches are modifications of the "possible-worlds" semantics. In this paper we use a combination of several general techniques for building non-omniscient reasoners. First we provide for the explicit representation of notions such as problems, solutions, and problem solving activities, notions which are usually left implicit in the discussions of autonomous agents. A second technique is to take explicitly into account the notion of *resource* when we formalize reasoning principles. We use the notion of resource to describe interesting principles of reasoning that are used for ascribing knowledge to agents. For us, resources are abstract objects. We make extensive use of ordering and inaccessibility relations on resources, but we do not find it necessary to define a metric. Using principles about resources without using a metric is one of the strengths of our approach.

We describe the architecture of a reasoner, built from a finite number of components, who solves a puzzle, involving reasoning about knowing, by explicitly using the notion of resource. Our approach allows the use of axioms about belief ordinarily used in problem solving —such as axiom **K** of modal logic— without being forced to attribute logical omniscience to any agent. In particular we address the issue of how we can use resource-unbounded (e.g., logically omniscient) reasoning to attribute knowledge to others without introducing contradictions. We do this by showing how omniscient reasoning can be introduced as a *conservative extension* over resource-bounded reasoning.

1. Introduction

In the AI community, the notion of an autonomous agent capable of knowing facts about the world or about the knowledge of other agents is of particular interest. We are interested in the general problem of building autonomous agents that can reason, solve problems, and interact with their environment. To do this we must address issues such as how to represent what an agent knows, and what mechanisms to provide in order for an agent to reason about what he knows and what other agents know.

Many formalisms for reasoning about knowing commit an agent to be *logically omniscient*, that is, they commit an agent to know all the logical consequences of what he knows. It is our belief that you can not build a real-world agent that is omniscient. If a real-world agent is a finite structure, then it is *prima facie* unreasonable that he should have an infinite number of components.

A commonly accepted principle for reasoning about an agent's knowledge is that agents are capable of performing modus ponens type inference. This principle says that if an agent knows an implication 'A implies B' and the agent also knows the premiss 'A', then the agent knows the conclusion 'B'. We shall refer to this principle as axiom **K**. Many philosophers have accepted this principle as reasonable (cf. (Montague, 1968)), and many AI researchers have incorporated some variant of axiom **K** in their theories of knowledge (cf. (Levesque, 1984; Halpern and Moses, 1992)). However axiom **K**, as it appears in standard formalisms, implies logical omniscience. Consider the following variant of axiom **K**.

$$\textbf{(K)}\quad \forall a\ w_1\ w_2\ .\ \mathit{Knows}(a, w_1) \wedge \mathit{Knows}(a, w_1 \overset{\circ}{\Rightarrow} w_2) \Rightarrow \mathit{Knows}(a, w_2)$$

where a ranges over agents, w_1 and w_2 range over formulas of propositional logic and $w_1 \overset{\circ}{\Rightarrow} w_2$ is the formula 'w_1 implies w_2'. The quantification is first-order, i.e., we considered formulas to be data structures. $\mathit{Knows}(a, w)$ expresses the fact that an agent a knows that the formula w is true, where $\mathit{Knows}(-,-)$ is not a modal operator, but a <u>relation</u>. In (Weyhrauch *et al.*, 1995) we show that axiom **K** implies logical omniscience by giving an `FOL` proof of the following formula:

$$\forall a\ w.(\mathit{Taut}(w) \Rightarrow \mathit{Knows}(a, w)),$$

where $\mathit{Taut}(w)$ expresses the fact that the formula w is a tautology. The `FOL` proof makes use of induction on the length of proofs.

It seems clear to us that an agent may fail to know whether a long formula is a tautology, even if he knows a complete set of axioms and rules for propositional logic. If you can imagine that there is some tautology that agent A does not know, then a formal contradiction is straight forward. Namely, for some tautology w, you can conclude both $\neg Knows(A, w)$, and $Knows(A, w)$. *Notice that, in this case, it is you, not the agent A, that is inconsistent.*

In fact one does not need the power of induction or even axiom **K** to conclude that an agent knows an unbounded number of things. It only requires a deductive system that generates an unbounded number of consequences. For example just a single axiom scheme of propositional logic can be instantiated to obtain infinitely many consequences. Thus even without axiom **K**, allowing unrestricted use of the usual deductive mechanisms can be problematic. For example, it is reasonable to imagine that an agent might not know a very long formula even if it is just an instance of an axiom scheme, because he might not be able to match it against the generating scheme. On the other hand, everyday use of principles like axiom **K** to reason about the knowledge of others does not cause us to be inconsistent!!! Why?

A number of formalizations of knowledge have been developed that do not ascribe logical omniscience to agents (see for example (Cadoli and Schaerf, 1992)). With few exceptions, these approaches are modifications of the "possible-worlds" semantics (see (Hintikka, 1962; Kripke, 1963)) for epistemic logic. Excellent surveys of existing solutions to the logical omniscience problem can be found in (Fagin *et al.*, 1995) and in (Halpern, 1991).

In our view, the problem of logical omniscience is not grounded in the details of any particular formal assumptions, but rather in how we allow ourselves to use these assumptions to attribute knowledge to some agent. In particular, the problem is not having axiom **K** available, but how we might use this axiom to make attributions of knowledge to other agents. Thus, we do not expect to find "solutions" to the logical omniscience problem based on alternative logics. In this paper we present a solution that involves the use of abstract resources to control the use of logic. We are concerned with the attribution of knowledge to agents rather than with a theory of knowledge or knowing, since attribution of knowledge is all that we can hope to do. Our solution allows us to use axioms such as **K** in problem solving without being forced to attribute logical omniscience to any agent, including ourselves.

We use a combination of several general techniques for building reasoners. First we explicitly represent notions such as problems, solutions, and problem solving activities, notions which are usually left implicit in the discussions of autonomous agents. By representing problems and solutions as objects, agents can have a theory of their own problem solving techniques, and theories of other agents' knowledge and problem solving capabilities. Each of these theories is represented as an FOL context.[1] FOL contexts are finite data structures that include both linguistic and model theoretic representations of information. The model theoretic information takes the form of a *simulation structure* which contains data structures representing individuals, and procedures for computing functions and predicates over these data structures. Contexts can be constructed and reasoned about by other contexts, and can appear as objects in the simulation structures of other contexts. These contexts play many different roles in the formalizations below. FOL systems are finite collections of FOL contexts, possibly with constraints on the relations between the constituent contexts or on the rules for change over time. Detailed descriptions of the FOL context data structures and their semantics can be found in (Weyhrauch and Talcott, 1994; Weyhrauch and Talcott, 1995a)

A second technique is to take explicitly into account the notion of *resource* when we formalize reasoning principles. We use the notion of resource to describe interesting principles of reasoning that are used for ascribing knowledge to agents. An example of such a principle is the *you are an expert* principle: *if I am a beginner in a certain discipline and I know something about that subject, and if you are an expert in the same discipline, then I can safely conclude that you also know whatever I know about the subject.* This principle says that an expert in a subject always has more resources for discovering facts about his subject than a beginner. This principle can be captured as part of our formalism because we are able to talk explicitly about resources in our language. We can imagine a novice has very limited resources to attack a subject but an expert in that subject has, by comparison, almost unlimited resources. We can also state that having lots of resources guarantees success in proving some simple theorems.

For us, resources are abstract objects. These objects have features suggestive of many diverse things: memory; computing time; knowing an algorithm or an heuristic. We could consider the knowledge that another agent communicates to us to be a resource, since we may use it provided we trust the agent. In general, resources are very complex objects, but we believe that we do not need a detailed theory of resources. Our theory of resources is expressed using general principles

[1] In this paper FOL does not stand for First Order Logic in general, but rather the specific system built by Weyhrauch and co-researchers, its data structures, and its underlying philosophy. Phrases like FOL context refer not to some general idea, but refer to well-defined data types that are used in FOL, a program for building FOL systems.

such as *the resource needed to know* $p \overset{\circ}{\Rightarrow} p$ *is small compared to the resources usually applied to problems*. We make extensive use of ordering and inaccessibility relations on resources, but we believe that it is a mistake to define a metric on resources. Using principles about resources without assigning a metric is one of the strengths of our approach.

Other researchers have described interesting formalizations of resource-bounded agents which are able to reason about their own resources. In (Elgot-Drapkin and Perlis, 1990; Elgot-Drapkin, 1991) the reasoner has a kind of self-reference that affects the course of its deductions: the resources of an agent —which can be axioms or inference rules— evolve over time, and the agent is conscious of their evolution. The agent —by reasoning in an introspective way and considering the resources he currently has and that he had in the past— is able to perform complex forms of reasoning about knowledge. A detailed comparison with our approach is given in § 5.2.

In order to clarify the intuitions underlying our ideas, we illustrate the techniques discussed above as we analyze in detail a famous puzzle dealing with reasoning about knowledge: the "muddy children problem". The muddy children problem —a variant of the "wise-men problem"— has been studied by many researchers, who have proposed a variety of solutions. The point of this paper is not to provide a "better" or "more elegant" solution to the problem, but rather to use it as a vehicle for presenting our ideas about resource bounded reasoning and how they allow an agent to use principles such as axiom **K**, and yet not have logical omniscience as a consequence. Our solution has some similarities with other work, however the set of reasoning principles that we use is —to our knowledge— novel.

We describe the architecture of a finite reasoner who solves the problem by explicitly using the notion of resource. In particular we address the issue of how we can attribute knowledge to others using demonstrably unsound resource-unbounded (e.g., logically omniscient) reasoning without introducing contradictions. We do this by showing how such "dangerous" ideas can be introduced as a *conservative extension* over resource-bounded reasoning. This allows us to mix both resource-bounded (safe) reasoning and resource-unbounded (unsafe) reasoning. We also state a principle called the *comparable resources* principle, which formalizes the conservative extension, and gives criteria for safely switching from one mode to another, when appropriate.

The structure of the paper is as follows. To provide background and motivation for our approach, in § 2 we discuss the overall goals of the `FOL` project and some of the key issues to be addressed in order to achieve these goals. In § 3 we present the muddy children puzzle and what we consider a reasonable, but still informal, solution to it. In § 4 we describe the construction of a reasoner able to solve the muddy children puzzle. The key ideas that we use involve principles for resource-bounded reasoning and principles for relating resource-bounded and resource-unbounded reasoning. In § 5 we discuss related work. In § 6 we summarize and indicate some future research topics.

2. `FOL` Problem Solving Contexts

Our solution builds on the basic structures provided by the `FOL` framework (Weyhrauch, 1977; Weyhrauch, 1980; Talcott and Weyhrauch, 1990; Weyhrauch and Talcott, 1994; Weyhrauch and Talcott, 1995a). The long term goal of the `FOL` project is to be able to build actual agents that can come to know things and that can reason. There is an important difference between giving a description of what a reasoner did, and building a reasoner that does do it. A formal system that is adequate for talking about reasoners will not necessarily be an adequate blueprint for building them. Because our goal/dream is to build agents that can reason about other agents, and about themselves, it is necessary that `FOL` meets both criteria.

The `FOL` project addresses such questions as: "What is the architecture of a knower?" and "What kinds of data structures are needed by a knower to represent knowledge and to support his reasoning activities?" Since real knowing agents, either human or computer, are made from a finite amount of matter, their reasoning capabilities and knowledge must be *finitely representable*, that is, presented using only a finite number of finite structures. Thus, in `FOL` we insist that every idea and notion can be represented as a finite structure.

One of the basic activities of reasoning agents is problem solving. In order to make substantial progress in being able to build even moderately competent reasoning agents it is necessary to develop a theory of problem solving. To make advances in the theory of problem solving we need theories that explicitly mention problems and solutions and problem-solving principles. One of the things that is conspicuously missing from the AI literature is theories that consider problems and their solutions themselves to be objects. No principles for justifying that proposed solutions actually solve the problem are set forth. In all cases, the fact that a solution works is entirely in the head of the author or reader of the paper. Some ideas about what a theory of problems should provide are discussed in (Weyhrauch and Talcott, 1995c). The key notion is that of a general *problem solving context* which embodies a theory of problem solving. The role of the general problem solving context is to extract information from a problem statement and to construct a *problem context* appropriate for solving the problem. Solving a problem means to construct a context that is appropriate for the problem and that is capable of answering questions relevant to the problem. General problem solving contexts will contain general principles for problem solving, while problem specific contexts will contain additional principles appropriate for the particular problem.

For example, if a problem involves reasoning about other agents, an `FOL` problem solving context might contain information about the questions: under what circumstances will a problem solver be willing to conclude that some other agent a knows that w is true —$Knows(a, w)$; and, what further conclusions is the problem solver willing to draw on the basis of this conclusion? Nothing in the proposed theory of problem solving requires anyone's "knowledge" to be absolute. Different agents may have different problem solving contexts. Thus answers to these questions are relative to the problem solver as well as to the problem.

An example of a reasoning principle that an agent might use to decide that another agent knows something is: *an agent A knows all the tautologies of the propositional calculus.* We might represent this with the single formula

$$\forall w.(Taut(w) \Rightarrow Knows(A, w)) \tag{1}$$

If we believe that agent B should use the above principle for ascribing knowledge to agent A, then, in the formalism described below, we would put formula (1) in our *problem solving context* for B. The act of putting (1) into the problem solving context of B is thus an assumption we have made about how B reasons. Because we must explicitly attribute this principle to B, if later, for any reason, we come to believe that our ideas about A's knowledge are incorrect, then we can look to this assumption as a possible source of error. In this way, we can account for the fact that we can be wrong about what A knows.

Another essential task in order to make progress in our ability to build reasoners with improved problem solving behavior is to find new kinds of reasoning principles. One kind of such principles make statements about how we can use our present knowledge to conclude new things. Such principles are almost never noticed when we discuss solutions to problems, but it may be essential to use them explicitly when we try to find solutions. An example is the *you are an expert* principle mentioned in § 1. One of the objectives of this paper is to begin articulating such principles and demonstrate their use.

3. The muddy children problem

In this section we analyze the muddy children problem, which concerns a group of children that are supposed to answer some questions asked by their father. We assume that the children are smart and are able to reason not only using their own knowledge, but also about what someone else knows. This puzzle is a variant of both the well-known *unfaithful wives* and *wise-men* puzzles, discussed in (Gamow and Stern, 1958). The version we present here is taken from (Barwise, 1981) (cf. also (Fagin *et al.*, 1995)), which deals with an arbitrary number of children. Many interesting aspects of this problem arise in the case of just two children, so we decided to use this simple case.

3.1. The puzzle

Two children (A and B) are playing. Their father (Dad) comes in and says "At least one of you has mud on his forehead". After a few seconds, Dad asks "Does any of you know whether he has mud on his forehead?". Both A and B answer "no". After few seconds, Dad asks again "Does any of you know whether he has mud on his forehead?". Both A and B answer "yes". What happened? Why do A and B answer "yes"?

3.2. A possible explanation

The following paragraph is one interpretation of what happened, told from the point of view of child A. It is a natural language description of one possible reasoning path but we will use it below as the basis for our more formal analysis.

I was playing with my brother, B, when Dad came in and said "At least one of you has mud on his forehead". I noticed that my brother had mud on his forehead. Then Dad asked "Does any of you know whether he has mud on his forehead?". I said "no", because I had really no reason to think that I was muddy. Actually, I had no reason to think that I was <u>not</u> muddy, either! It could have been one of the two: muddy or not, and there was no way to know which one was right. My brother also said "no". After few seconds Dad asked the same question again, and I did not understand why he was doing that again: we just answered to it few seconds before! But I considered the fact that usually Dad does not ask questions for no reason, and maybe it was worth to think a little bit more about it. This is how I reasoned: "Well, I could be muddy, since Dad said that at least one of us is muddy. I could be not muddy as well though, ... But, ... wait a minute! It's not true! If I were not muddy, then my brother B would have noticed that, and he could not have answered "no" to Dad's question last time, since he is smart and he is always sincere to Dad, and he would have understood that he was the muddy one". Hence I answered "yes".

4. Building a resource bounded reasoner

We imagine a problem solving agent, YB,[2] interested in solving the muddy children problem. In (Weyhrauch and Talcott, 1995b) we make the claim that `FOL` contexts can be used as the building blocks for building the mind of such an agent. In this section we describe `FOL` contexts that make up part of the mind of YB and show the changes they undergo as YB concludes that, if A is smart enough he could have correctly concluded that he is muddy and thus answered "yes". We show how `FOL` contexts can embody the knowledge of agents and how the process of reasoning about their knowledge can be carried out using ordinary forms of reasoning, while avoiding the inconsistency of logical omniscience. The key idea is to show how to use conclusions resulting from omniscient reasoning to provide proofs using resource-bounded reasoning.

We analyze the process of reasoning involved, formulate some general problem solving principles, and show how YB uses them. We emphasize the role played by resources in YB's reasoning; this gives us the opportunity to formulate general principles for resource-bounded reasoning.

We stress that YB is to be thought of as an actual agent, not a description of a theoretical agent. YB is different from Dad, A and B, but he has the same kind of limitations that they have (e.g., he is finite in structure, he is not omniscient, and all his awareness of the world comes through sensors).

4.1. How YB solves the puzzle - the critical step

The task of YB is to make sense of the information exchanged by the various agents, on the assumption that they are smart and sincere. In particular YB is supposed to understand why A answers "no" to the

[2] YB stands for "Yellow Box". YB got his name because of how he appeared on the blackboard in lectures given about this subject by Richard Weyhrauch in the early 1980's, where the objects in the mind of an agent were drawn inside a box drawn in yellow chalk that surrounded the entire blackboard. This represented the idea that we were to think of being inside the agent's mind rather than simply looking at it. (Blue and Red chalk were used to describe data structures that represented use and mention.)

first question and "yes" to the second one. Since A can not see the spot on his head, YB must convince himself that A has deduced that he is muddy.

YB knows very little about A's reasoning capabilities, except that he is smart and sincere. From YB's point of view it is not reasonable to ascribe to A the knowledge of a particular inference rule or reasoning pattern. On the other hand YB himself is able to prove to himself that A is muddy by means of a simple case analysis.

We are left with the question of how YB can ascribe this knowledge to A. We propose looking at the following scenario. YB can reason about his own problem solving activities. In particular, he is able to see that his own proof that A is muddy is not very complex, and assumes that A should be able to arrive at the same conclusion, not necessarily by using the same steps. On the assumption "*A is smart*", YB assumes that A has arrived at the same conclusion. YB therefore attributes knowledge of this conclusion to A.

The reasoning principle that YB is using might be called "*X is as smart as I am*" (for the purpose of this problem). The notion of resource plays a role in this principle, since YB must be able to evaluate the resources he spent on the proof, and to ascribe comparable resources to A. More generally, for an agent to apply the "*X is as smart as I am*" principle, the agent must estimate the resources he expended on a problem, then he must ascribe to X a comparable amount of resource for solving the problem. In § 4.3 we describe in detail YB's proof that A is muddy and give a formal account of the above principle.

The use of such a reasoning principle is —to our knowledge— novel. In existing solutions to reasoning about knowledge puzzles (see for example (Aiello *et al.*, 1991; Elgot-Drapkin, 1991) and the discussion in § 5) the solution is carried out ascribing precise reasoning patterns to agents.

As we can see, YB is a complex reasoner, who at some point needs to reason about his own reasoning. YB's head must contain structures which can carry out this meta-reasoning. We imagine that it contains both a *Problem Context* and a *Problem Solving Context*. The Problem Context contains all the information that YB needs for proving that A is muddy, and Problem Solving Context is able to perform meta-reasoning on it, in particular by extracting proofs from the context and evaluating the resources they use.

This structure is not *ad hoc*, but rather an instance of a very general paradigm for representing problems and their solutions using collections of contexts called `FOL` systems. Problem contexts are part of problem solving contexts. Problem solving contexts have the capability of creating and manipulating problem contexts. In addition, problem solving contexts may contain special purpose principles for constructing, reasoning about, and operating on problem contexts. These include general problem solving principles, as well as means for looking at problems and extracting problem specific reasoning principles. The nature of general problem solving contexts is discussed further in (Weyhrauch and Talcott, 1995c).

4.2. The structure of YB

In this subsection we describe in more detail the structure of the agent, YB, and some of our methodology for modeling reasoning about agents in YB's mind. The Problem Context of YB must contain all the relevant information about the reasoning agents (the father and the children), the messages they exchange and how they interact.

YB sees the children and their father as independent agents, capable of performing complex, albeit finite reasoning. He models their knowledge and reasoning using contexts. YB puts these structures in the simulation structure of his Problem Context, thus building a model of the situation.

So far, the picture of YB is as follows

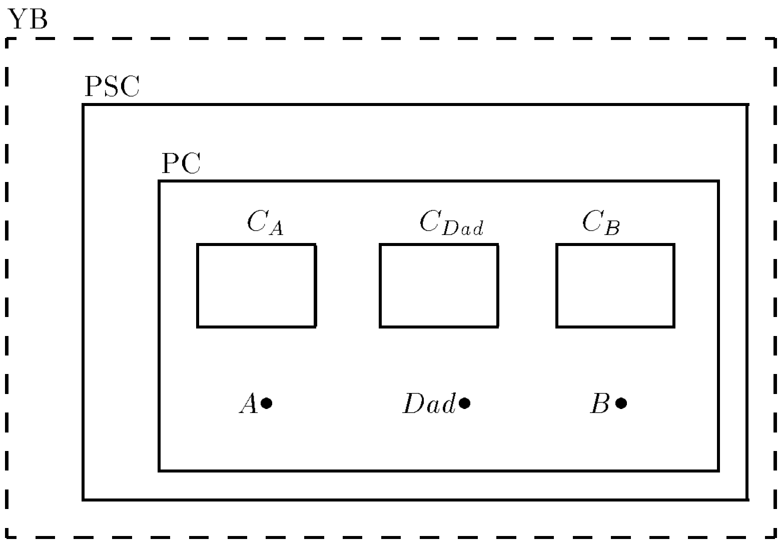

Figure 1: Structure of the Yellow Box

In the above picture, contexts are represented by solid rectangles. We use a dashed line for the outer rectangle since we can not make a yellow box. PSC is YB's general Problem Solving Context, PC is YB's Problem Context for the Muddy Children Problem. A, B and Dad are objects in PC that are YB's mental images of the agents A, B and Dad, respectively. C_A, C_B and C_{Dad} are `FOL` contexts in PC that constitute what YB is willing to assert about the knowledge of these agents.

YB reasons about the agents and their knowledge contexts. There is a constant symbol in the language of YB's Problem Context for each agent: $\hat{A}$ for child A, $\hat{B}$ for child B, $\widehat{Dad}$ for the father. YB uses the constant symbols $\hat{C}_A$, $\hat{C}_B$ and $\hat{C}_{Dad}$ to talk about the contexts C_A, C_B and C_{Dad}, respectively. $KnowledgeOf(a, c)$ means that context c contains what agent a knows. The assertions $KnowledgeOf(\hat{A}, \hat{C}_A)$, $KnowledgeOf(\hat{B}, \hat{C}_B)$, and $KnowledgeOf(\widehat{Dad}, \hat{C}_{Dad})$ are found amongst the facts of YB's Problem Context. Thus, in this problem, $\hat{C}_A$ is YB's understanding of the knowledge of $\hat{A}$.

YB knows that the other agents have names for each other, but he does not know what names the agents A, B, and Dad use for one another. *For the purpose of solving this problem we imagine that YB appeals to a problem solving principle that allows him to assume that from his point of view, A, B, and Dad agree with each other and with him about who is who.* Therefore he puts appropriate constant symbols in the contexts C_A, C_B and C_{Dad}.

From YB's point of view, the occurrences of a symbol in different contexts may or may not have the same interpretations. YB, in his solution of the puzzle, makes practical use of his assumption that A, B, and Dad agree with him about who is who. In his mind, YB attaches the same object to each occurrence of $\hat{A}$, $\hat{B}$, and $\widehat{Dad}$. In the same spirit, YB attributes to each of the agents A, B, and Dad, a common idea of the notion of muddy people. He puts the same predicate symbol, $\hat{M}$, in each agents language (including that of his own Problem Context), and assumes that all the agents mean the same when they talk about being muddy.

Notice that having the same symbols occur in different contexts means that expressions built from these symbols also belong to the languages specific to each of these contexts. YB needs to mention and to assert such formulas. For this, we use the following 'meta' notation. $\mathtt{mud}(\hat{A})$ is the formula $\hat{M}(\hat{A})$. Small circles over symbols signify propositional connectives that are used to mention the corresponding formulas. Thus $\mathtt{mud}(\hat{A}) \mathring{\vee} \mathtt{mud}(\hat{B})$ is the formula $\hat{M}(\hat{A}) \vee \hat{M}(\hat{B})$.

YB says $AssertableIn(w, c)$ in a problem context to mean that he is willing to assert w in c, for purposes of solving that problem.

A stronger assertion is $IsaFactOf(w, c)$, which means that w is the formula asserted by a fact of the context c (informally, w is a fact of c). Since facts may have dependencies —assumptions from which they are derived— we may make these explicit by saying $IsaFactOf(w, c)$ under assumptions D. The distinction between willingness to assert a formula in a context (for purposes of solving some problem) and the occurrence of a formula as the formula asserted by a fact in a context is important when we want to control the use of reasoning about what an agent might know in the attribution of knowledge.

YB can put $AssertableIn(\mathtt{mud}(\hat{A}) \overset{\circ}{\vee} \mathtt{mud}(\hat{B}), \hat{C}_B)$ in his Problem Context to indicate his willingness to assert $\mathtt{mud}(\hat{A}) \overset{\circ}{\vee} \mathtt{mud}(\hat{B})$ in C_B, for purposes of the Muddy Children problem. Another general problem solving principle allows YB to construct a new model, C'_B, of B's knowledge by doing the assertion operation on C_B. Then it will be the case that $AssertableIn(\mathtt{mud}(\hat{A}) \overset{\circ}{\vee} \mathtt{mud}(\hat{B}), \hat{C}'_B)$.

4.3. Carrying out the reasoning

In this subsection we describe the reasoning that YB does in order to conclude that A knows that he is muddy. As we said previously, the main steps of this reasoning are:

1. YB convinces himself, i.e., produces a "proof" that A is muddy;

2. YB ascribes to A the capability of doing the same using the "*X is as smart as I am*" principle, applied to A.

The description of the reasoning is informal. We make use of various problem solving principles, which are the phrases written in italic. We emphasize that this discussion is from the point of view of YB. Thus, when we say 'B' we mean YB's mental image of B, i.e., the semantic entity that is found *in YB's Problem Context*.

Each reasoning step is the application of a logical rule of deduction, or of a reasoning principle taken from the Problem Solving Context of YB. Such steps lead to the addition of some fact to the Problem Context of YB. Technically speaking, the evolution of the reasoning consists in a sequence of contexts $\mathrm{PC}_0, \ldots, \mathrm{PC}_n$, that are the evolution of the Problem Context of YB: PC_0 is the starting context, PC_n, is the last context, which contains a fact asserting that A knows that he is muddy. The explicit construction of contexts in this manner is described in more detail in (Weyhrauch, 1980). In the following, 'Problem Context' names the current state of evolution as there is no need, either for us or for YB, to make the history of the evolution explicit.

YB looks at the proof and determines it is "simple". He knows A is as "smart" as he is and concludes that A can do a proof too. If A is trying to solve this problem, A himself will have found a proof and therefore concluded that he is muddy.

4.3.1. *YB convinces himself that A is muddy*
This is a description, from YB's point of view, of the sequence of reasoning steps that lead YB himself to understand that A has a muddy forehead. In some reasoning steps the notion of resource plays an important role. This will be discussed later on.

- YB is told that Dad said "At least one of you has mud on his forehead". As a consequence, YB reasons in this way:
  - B knows that $M(A) \vee M(B)$. (YB uses the reasoning principle *B trusts and understands Dad*, which is in YB's Problem Solving Context). Formally, YB puts $\mathtt{mud}(\hat{A}) \overset{\circ}{\vee} \mathtt{mud}(\hat{B})$ in C_B.
  - YB uses the principle in his Problem Context that *B has looked at A and has determined whether or not he has mud on his forehead.* Formally, YB adds the fact

    $$AssertableIn(\mathtt{mud}(\hat{A}), \hat{C}_B) \vee AssertableIn(\overset{\circ}{\neg}\mathtt{mud}(\hat{A}), \hat{C}_B) \tag{2}$$

    to his Problem Context. This expresses the disjunction logically without having to build the two possible contexts, or to reason explicitly about alternative worlds.

- YB is told that Dad asked "Does either of you know whether there is mud on your forehead?" and that B answered "no". YB infers that B does not know whether B is muddy (using the principle *B is smart and sincere*).

  Formally YB can represent 'B does not know' in two ways:

  resource-unbounded version: $C_B \not\models \mathtt{mud}(\hat{B})$, i.e., the information available to B (in YB's model) does not logically imply that B is muddy;

  resource-bounded version: $C_B \not\vdash_{ask} \mathtt{mud}(\hat{B})$, where $\vdash_{ask}$ denotes the (resource-bounded) procedure that B has used for answering the question. This procedure may be a decision procedure for some epistemic logic, or any other technique that B might choose to use. YB has no way to know the procedure, but knows only its result. YB can assume that some maximal amount of resource was used in trying to answer (since B is sincere), and hence that no amount of resource available to B would yield the answer $\mathtt{mud}(\hat{B})$.

- At this point YB uses the principle of reasoning by cases based on the disjunction (2). Hence he must consider the two possibilities according to whether B observed that A is muddy or not. YB starts with the case in which he imagines that B observed that A is not muddy.

- Thus, YB assumes (in his Problem Context) that $AssertableIn(\mathring{\neg}\mathtt{mud}(\hat{A}), \hat{C}_B)$.

- As stated before, YB knows that (i.e., has as a fact in his Problem Context) $AssertableIn(\mathtt{mud}(\hat{A}) \mathbin{\mathring{\vee}} \mathtt{mud}(\hat{B}), \hat{C}_B)$. Depending on the interpretation of B's response, "I don't know", YB reasons as follows:

  resource-unbounded version: using simple properties of $\models$ YB infers that $C_B \models \mathtt{mud}(\hat{B})$, which is a contradiction.

  resource-bounded version: YB infers that $C_B \vdash_{mp} \mathtt{mud}(\hat{B})$, where mp represents resources sufficient to carry out one application of modus ponens rule. By using the principle *B is smart*, YB sees that B has sufficient resources to insure equivalent reasoning capability. Since YB has previously concluded that no amount of resource available to B would yield an answer, YB infers that this too is a contradiction.

  Either interpretation of "I don't know", leads to a contradiction. Hence it must be the case that B has observed that A is muddy. Formally, YB adds the fact $AssertableIn(\mathtt{mud}(\hat{A}), \hat{C}_B)$ to his Problem Context.

- YB now concludes that A is indeed muddy, using the principle that *$\hat{B}$ perceives correctly.* Formally, YB uses the principle that for this problem he can believe that $\hat{M}(\hat{A}) \Leftrightarrow AssertableIn(\mathtt{mud}(\hat{A}), \hat{C}_B)$. Thus $\hat{M}(\hat{A})$ is added to the facts of YB's Problem Context.

4.3.2. *YB ascribes reasoning capabilities to A*

In the previous subsection we showed a possible way in which YB could convince himself that A is muddy. In this subsection we focus on how it is possible for YB to attribute the same knowledge to A, on the basis that *A is smart*, hence he is able to carry out an argument of comparable complexity leading to the same conclusion.

It is easy to prove that logically omniscient agents starting from the same set of premises arrive always at the same conclusions. Thus, if YB is willing to attribute logical omniscience to A, YB could conclude that A knows that A is muddy from the fact that he (YB) has a proof that A is muddy *in his Problem Context* starting from facts that A also knows.

We want a formalization of the *X is as smart as I am* principle, that does not rely on logical omniscience. As we already pointed out, resources seem to be implicit in this principle. Below we propose a very general principle for ascribing knowledge to agents that is an alternative to logical omniscience. In order to describe this principle we need to introduce some notation. Let a range over agents, pr over problems, and r, r_1, r_2 over resources. As noted above a solution to a problem is a context, appropriate for the problem, that can answer relevant questions correctly. The construction of a solution context shows the reasoning principles used to obtain the solution from some initial context formalizing the problem. We let π range over contexts intended as solutions to a problem.

The following predicates appear in YB's Problem Solving Context:

- $Solves(pr, \pi)$: π is a context that is a solution to the problem pr.

- $Restricted\text{-}proof(pr, \pi)$: the proof in π uses only inference rules from a certain set, and uses only problem solving principles appropriate for pr;

- $r_1 \lll r_2$: anything that can be done using resource r_1 can be done using resource r_2, without noticeable decrease of r_2. (r_2 is inaccessible from r_1.)

- $Resource\text{-}bound(pr, \pi, r)$: the solution π to problem pr uses resources r_1 dominated by r, that is $r_1 \lll r$.

- $Has\text{-}resources(a, pr, r)$: the agent a has resources r for solving problem pr;

- $Has\text{-}found(a, pr, \pi)$: the agent a has found a solution π for the problem pr.

The axiom **P1** below formalizes the "*X is as smart as I am*" principle.

$$(\mathbf{P1}) \quad \begin{array}{l} \mathit{Solves}(pr, \pi) \wedge \\ \mathit{Restricted\text{-}proof}(pr, \pi) \wedge \\ \mathit{Resource\text{-}bound}(pr, \pi, r) \wedge \\ \mathit{Has\text{-}resources}(a, pr, r) \end{array} \Rightarrow \begin{array}{l} \exists \pi_1 . \mathit{Solves}(pr, \pi_1) \wedge \\ \quad \mathit{Has\text{-}found}(a, pr, \pi_1) \end{array}$$

P1 says that for any problem *pr*, agent *a*, and resources *r*; if we have a solution π to *pr* with a restricted kind proof, which uses resources dominated by *r*, and if the agent *a* has resources *r* for solving *pr*, then there is a solution π_1 to *pr* that the agent *a* is able to find. Notice that the proofs in two contexts π, π_1 need not be the same. Note also that we do not need to know exactly either the amount of resources used to obtain the solutions, or which of the two amounts is larger. Using this principle we can ascribe knowledge to other agents without having to know their reasoning patterns. At the end of this section we will introduce principles that provide some constraints on the *Restricted-proof* predicate and give some further insight as to the nature of the predicate.

We use **P1** in the following way. If we have constructed a solution, π, to a problem, *pr*, and if we imagine that agent *a* has resources that dominate those that we used to solve the problem, then we can conclude that *a* has solved this problem as well.

In addition to **P1** we need some principles relating contexts and knowledge of agents modeled by these contexts. Suppose *pr* belongs to a class of problems with the property that there is a formula w_{pr}, the answer to the question inherent in *pr*, and a set of formulas Γ_{pr}, the facts or assumptions of *pr*, such that if $\mathit{Solves}(pr, \pi)$, then $IsaFactOf(w_{pr}, \pi)$ under assumptions Γ_{pr} and there is a function *concl* on solutions such that $\mathit{concl}(\pi) = w_{pr}$.

The "is A muddy?" problem belongs to this class. Assume that *c* is a context that is the model of the knowledge of agent *a*, i.e., $\mathit{KnowledgeOf}(a, c)$, in some problem context. **r1** says that if agent *a* has found a solution π to *pr* which uses resources dominated by *r*, and if *a* has resources *r*, then we can assume that the conclusion of π is among the facts of *c*. **r2** says that we are willing to attribute knowledge of any fact of c to a.

$$\textbf{(r1)} \quad \textit{Has-found}(a, pr, \pi) \wedge \textit{Solves}(pr, \pi) \wedge (\forall \gamma \in \Gamma_{pr})(\textit{IsaFactOf}(\gamma, c)) \wedge \textit{KnowledgeOf}(a, c) \Rightarrow \textit{AssertableIn}(\textit{concl}(\pi), c)$$

$$\textbf{(r2)} \quad \textit{IsaFactOf}(w, c) \wedge \textit{KnowledgeOf}(a, c) \Rightarrow \textit{Knows}(a, w)$$

YB uses principles **P1, r1, r2**, which are in his Problem Solving Context, for ascribing to A the knowledge that he is muddy, on the basis that he is able to generate a proof with the same conclusion as YB's, from the facts of the problem. This can be done provided that:

1. YB is able to evaluate the resources r_0 that he spent in his proof, and to determine that r_0 is dominated by the resources r that he has. This is needed for instantiating the *Resource-bound* predicate.

2. YB is willing to ascribe resources r, to A.

Concerning the first prerequisite, YB could have in his Problem Solving Context a function *resources-used*, which can compute the resources used in a proof. Alternatively, the resources used in each step of reasoning could be made explicit, as part of the justifications of facts in YB's Problem Context. Concerning the second prerequisite, YB could ascribe the amount of resources r to A on the basis that A is smart. Assuming these capabilities for YB, we see that YB can add to the Problem Context the assertion $\textit{Knows}(\hat{A}, \texttt{mud}(\hat{A}))$.

If we do not want YB to be committed to be able to evaluate the resources of its proof, then a possible solution is that YB asserts the existence of some such resource and attributes to agent A unlimited resources, still on the basis that he is smart. In other words YB may choose to use the axiom (**P2**):

$$\textbf{(P2)} \quad \forall r.\textit{Has-resources}(a, pr, r)$$

Let us consider what we have so far:

1. There are two different ways to represent YB's reasoning that leads him to conclude that $M(A)$ is true: YB could use reasoning principles for the evolution of the Problem Context that make explicit the resources used, or use the same reasoning principles in a version that does not mention resources at all.

2. The "*A is as smart as I am*" principle could be used by YB in two different ways: **P1**, which requires the evaluation of the resources spent in the proof, and **P1** $\wedge$ **P2**, which does not.

This situation is represented in the following diagram:

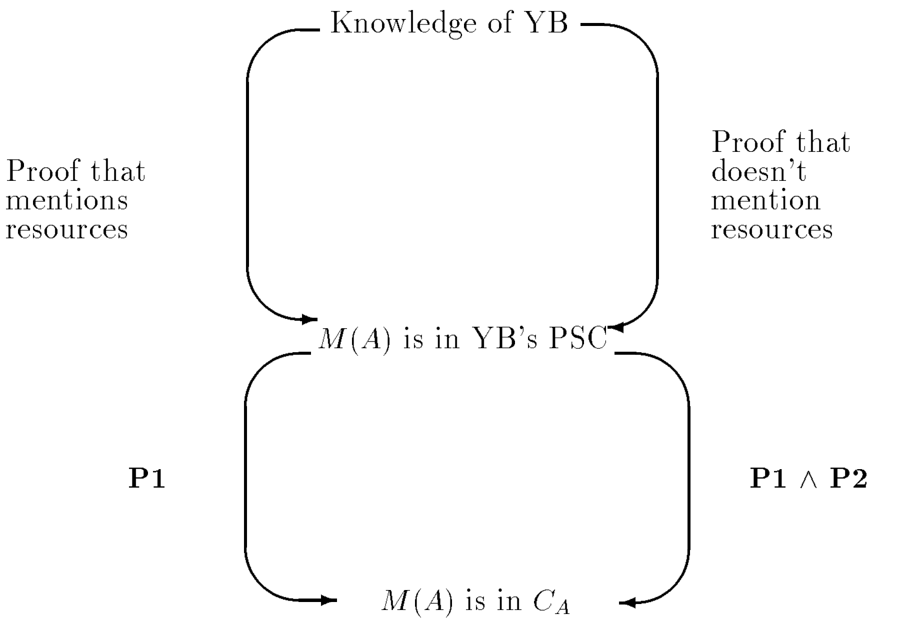

Figure 2: YB's possible ways of reasoning

In the above diagram the upper-left edge and the lower-right one suffer from two different problems: the upper-left edge requires a notion of proof in which we are required to build complex resource terms, that do not occur in our actual discussion of this problem. The lower-right ascribes to agent A an infinite amount of resources, which defeats the purpose of using resource-bounded reasoning, and is in general unsound.

It seems that "the best of both worlds", as far as resource-bounded (sound) versus resource-unbounded (unsound) reasoning is concerned, is represented by the other two edges: upper-right and lower-left. But now we have a question: if YB proves that A is muddy without mentioning resources at all, how can he use the axiom **P1**, which requires the evaluation of the resources spent in the proof? We claim that omission of explicit mention of resources in the upper-right edge is harmless. We may appeal to the fact that YB has a proof in his Problem Context. This proof is a finite data structure that can be examined, using the *resources-used* function, to determine the resources used. The process of determining the resources used effectively transforms the proof into one that makes resources explicit. In the Problem Solving Context we obtain $PC_n \vdash_{\mathtt{r}} \mathtt{mud}(\hat{A})$, for some concrete resource term $\mathtt{r}$ rather than simply $PC_n \vdash \mathtt{mud}(\hat{A})$. Because of the restricted nature of the proof YB can infer that the resources available to A dominate the resources expressed by the explicit term. The transformation of the argument given by YB making no explicit mention of resources available to A into an argument formulated in terms of resource bounded reasoning is actually an instance of a general property of restricted reasoning which we call the *comparable resources principle.*

(CR) If I **have** proof in hand, then I can determine the resources I used to construct it.

If I have a proof in hand whose resources I have determined are negligible compared to the resources I have for solving this problem, and I'm willing to attribute to X resources for solving this problem comparable to my resources, then I can reason that X knows the conclusion of this proof. This is because X can construct this or a similar proof using little of his resources.

Notice that the use of **CR** depends upon explicit assumptions about the resources of X. It is this principle that allows the use of apparently unsound resource unlimited reasoning to attribute knowledge to agents. And it also justifies mixing resource limited reasoning with resource unlimited reasoning.

To summarize, YB has at least three different ways to ascribe to A the knowledge that he is muddy:

1. The "quick and dirty" method is to ascribe an infinite amount of resources to A, by assuming **P2**. This essentially ascribes logical omniscience to A, but it is still an acceptable solution, since **P2** is an <u>explicit</u> assumption (that may be retracted later on).

2. YB may appeal to the fact that his particular proof was constructed using limited resources and that these same resources must be available to A for this problem, on the basis that A is smart. Therefore YB can attribute knowledge of his conclusion to A, using **P1**.

3. YB can use the *comparable resources* principle directly.

The principles **P2** and **CR** represent two directions of specializing **P1**. **P2** corresponds to assuming availability of unbounded resources and yields logical omniscience. **CR** corresponds to making stronger assumptions on the nature of available reasoning principles (assuming that only suitably restricted proofs can be constructed) and this means we can carry out an argument without worrying about resources in a way that is consistent with sound resource-bounded principles. Underlying **CR** is a property, **RP**, required of restricted proofs. **RP** says that such a proof can be transformed uniformly into proof based on sound resource-bounded principles.

(RP) Let $\mathcal{P}$ be a restricted proof-system which does not mention resources and let $\mathcal{P}_R$ be the corresponding proof system which uses explicit resources. Thus for any proof π in $\mathcal{P}$ there is a determinable amount of resource spent in π, and there is a corresponding proof π_R in $\mathcal{P}_R$ such that the resource explicit in the conclusion of π_R is comparable to the resource spent in π.

Technically, this says that addition of **CR** gives a *conservative extension principle*. It justifies use of principles that, used in an unconstrained manner, imply logical omniscience.

5. Related work

Many formal solutions to the "muddy children" puzzle have been proposed in the literature. We refer to (Aiello *et al.*, 1991) for a detailed review on papers on this topic appeared until 1990. More recent work can be found in (Fagin *et al.*, 1995; Halpern and Vardi, 1991).

In this section we consider other work related to two specific aspects of our solution: the use of FOL contexts and the use of resources to model and control reasoning about knowledge.

5.1. Reasoning about knowledge in FOL

FOL ideas have already been used for solving puzzles related to reasoning about knowledge. In (Aiello *et al.*, 1991) Aiello, Nardi and Schaerf studied the "three wise-men puzzle" proposing a solution based on multiple FOL contexts. In their approach a context is used for modeling each wise-man. Each context contains a theory of knowledge of the other agents, i.e., axioms and rules expressing necessary and sufficient conditions for the other agents to know facts. These axioms and rules model the precise reasoning patterns that are ascribed to the other agents. The solution to the puzzle is obtained using such a theory of knowledge.

One of the differences between this approach and the one we have used in this paper is that the agents we describe do not commit the other agents to any specific reasoning pattern. By using the *X is as smart as I am* principle described in § 4, agents are able to ascribe knowledge to other agents without having to fill in the details of how such knowledge is obtained.

Another difference between the two approaches is that we have explicitly used resources to constrain the reasoning used in problem solving.

Giunchiglia proposes in (Giunchiglia, 1993) a solution that uses multiple FOL contexts for representing the various agents. Beliefs about the world of interest are carried from a context into another (i.e., from a reasoner to another) by means of "bridge rules". Also in this work there is ascription of reasoning patterns, and resources are not considered.

We do not consider these as examples of alternative approaches, but rather as an illustration of the power of FOL to allow for a variety of solutions to a variety of problems.

5.2. Step-Logic

In (Elgot-Drapkin and Perlis, 1990; Elgot-Drapkin *et al.*, 1991) a general framework for taking into account time during evolution of reasoning is presented. The emphasis of the formalism —named Step-Logic— is on considering reasoning as a *process*. This means that Step-Logic does not focus on the end results of reasoning, but rather on its *history*: the formalism has a kind of real-time self-reference that affects the course of deduction itself. This gives a perspective on notions such as inconsistency which is rather different from the classical one: inconsistency at some specific step of reasoning is tolerated, and can even be useful for inferring new interesting facts in subsequent steps.

In FOL we are also interested in evolution of reasoning and not only in its final results. As a matter of fact, this is not the only general idea that Step-Logic and FOL have in common. They also share the assumption that the representation of the agent must always be finite, that the system must be flexible enough to allow any inference rule which is considered meaningful for a specific problem, and that the formal system should be able to model both positive and negative introspection, as well as retraction of knowledge. Finally, they both emphasize that the system's knowledge should always be presented in first-order syntax.

Nevertheless, there are also significant differences between the two approaches. Indeed, in Step-Logic there is no mention of key FOL ideas such as contexts and semantic attachments. Moreover, time must be mentioned explicitly in Step-Logic, while in FOL this is an option. In general, in Step-Logic the only way to ascribe some knowledge to an agent is to ascribe him a specific pattern of reasoning, since in such a formalism one always deals with explicit inference rules.

Such differences become evident when we compare the solution to the "three wise-men problem" provided in (Elgot-Drapkin, 1991) to the solution to the "muddy children problem" shown in the present paper. Although in both cases the notion of resource is essential, in Step-Logic the solution is attained without using several contexts. Moreover, each wise-man must ascribe a specific reasoning pattern, i.e., a set of inference rules, to the other ones. Although the presence of such inference rules is an explicit assumption, there are other assumptions (e.g., each wise-man is smart and reasons at the same speed of the other ones) that are not explicit, and therefore could not be retracted if there is evidence that they are dangerous.

5.3. Related work on contexts

`FOL` context data structures are the descendents of LS pairs (L for language and S for simulation structures) as presented in (Weyhrauch, 1977). `FOL` contexts as data structures were first described in lectures by Richard Weyhrauch in 1985 at Stanford and at the Meta-Level Architectures and Reflection Workshop held in Alghero in October 1986 (Maes and Nardi, 1988). In this workshop papers were presented using `FOL` contexts to describe non-monotonic reasoning, using `FOL` contexts for logic programming, and using reflection to evaluate meta-theorems in `FOL` systems.

The notion of viewpoints was discussed in 1984 by Attardi and Simi (Attardi and Simi, 1984; Attardi and Simi, 1997). In (McCarthy, 1987) John McCarthy suggested formalizing the notion of context by restricting the scope of the truth of statements. These ideas are elaborated in (McCarthy and Buvač, 1994; Buvač, 1996). Fausto Giunchiglia (Giunchiglia, 1991; Giunchiglia, 1993) has proposed multi-language systems as a form of contextualization.

Before we comment on these theories, we would like to reiterate that the aim of the theory of `FOL` systems is to provide a blueprint for the construction of an artificial mind, and as such has very different requirements from the other three. For this reason this paper does not present a theory of knowledge, in the usual sense, but presents a collection of data structures that might be in the head of a reasoner who needs to attribute knowledge to someone else. `FOL` systems are good for this because they are specific finite data structures that can capture not only one's mental idea of the theories of others, but by using simulation structures our reasoner can also imagine both the mental images of the objects mentioned in their theories and the relation of these to the objects imagined by our reasoner.

In the FOL approach satisfaction and satisfaction preserving maps are the fundamental semantic concepts. Validity (true in all models) as a notion finds little relevance. Using satisfaction allows for explicit treatment of operations such as forgetting, acquiring new language, and discovering new facts. The notion of consistency (has a model) for FOL systems is much simpler than that of other formalisms with multiple contexts [Attardi-Simi (AS), McCarthy- Buvač (MB), Giunchiglia (G)], while the notion of inference rule is much richer. Specifically, an FOL system is consistent if each of the contexts that it contains is consistent. A rule of inference is a *consistency preserving map* on FOL systems. Additional expressive capability is provided by the notion of a physics. A physics is a predicate, P, describing constraints on the contexts of a system and a physical system is a physics together with rules of inference that preserve P. Such a P represents a class of systems, and the rules tell you the ways in which a system can evolve. Many problems of philosophy, cognitive science, and artificial intelligence can be formulated as an appropriate choice of a property P and a set of P-preserving rules of inference. Some examples are simple theories (mathematical and common sense), meta theoretic reasoning, states of knowledge that change over time, the use of more than one modal operator at a time, non-monotonic principles of reasoning, etc.

Underlying the ideas presented in this paper is the use of FOL systems to formalize the activity of solving problems. There are contexts where very general principles of solving problems are stated, contexts where specific classes of problem are described and principles specific to solving such problems stated, and contexts describing specific problems. Problem solving contexts describe how information can be moved from one context to another and used to arrive at solutions.

FOL contexts can appear as objects of discourse of other FOL contexts. This happens explicitly when terms constructing contexts appear in a meta-level context or when a context occurs as an attachment to a constant symbol of another context (nesting). It also happens implicitly in FOL systems (cf. (Weyhrauch and Talcott, 1995a)). MB and AS consider nested contexts, while Gs work on multi-language systems has focused on the study of systems of contexts.

The Giunchiglia approach to contextualization has been used to formalize a theory of proof by abstraction 1992, modal logics 1994, and meta-reasoning 1996. This approach differs from the FOL approach in that the theory of multi-language systems uses, but has no way of explicitly mentioning, multi-language systems and in that sense does not deal with contexts as objects of discourse.

Even though MB often define a context by giving a finite number of axioms that need to be true in it, they explicitly say "Contexts are abstract objects." 1994, 1996 and do not identify these abstract objects with the sets of axioms, but rather a set of truth assignments, which in general are infinite sets. In AS (Attardi and Simi, 1997) "*Viewpoints* are *sets of sentences* which represent the assumptions of a theory." but later say "The possibility of recursive definitions of viewpoints (...) and of expressing self referential viewpoints leads to non finite viewpoints." Both of these provide a means for reasoning about truth and provability, but they do not address the FOL goal of providing blueprints for building artificial minds. They do not answer the question: what are the finite building blocks.

In MB they talk about contexts *associated* with a mental state of an agent and AS mention that viewpoints can be thought of as belief contexts which are aimed at *capturing* an epistemic state. These are remarks about what the notion of context means in their theories. In the FOL approach, FOL contexts are in one's head and FOL systems constitute the internal mental state of a thinker. Thus we speak of a thinker having a context as a physical part of its physical mind —in its straight forward meaning, its mental state. This is in contrast to a theory of what situation or context a thinkers is *in*. This distinction gives rise to to subtle differences that are interesting topics for future investigation.

The Problem Solving Principles of MB are good candidates for principles to be embedded in FOL problem solving contexts. However, axioms like (flat) are inconsistent with FOL interpretation of contexts.

The approach of ter Meulen (ter Meulen, 1995b; ter Meulen, 1995a) to representing the semantics of natural language communications has much in common with the FOL approach: the emphasis on the importance of concrete data structures, representation of partial information, and explicit rules for dynamic buildup of information. This work suggests possible blueprints for building blocks that provide artificial minds with the capability to acquire information by conversation rather than brain surgery.

6. Discussion and future research

We have described the construction of an agent (YB) that can solve the 2-muddy children puzzle. This construction uses a novel combination of techniques. It is an initial step in the task of developing general techniques for building actual agents that can reason, interact with their environment, acquire new knowledge, and solve problems with some minimal degree of competence.

The first technique is to provide for the explicit representations of notions usually left implicit in the discussions of autonomous agents. These notions include problems and solutions as objects, allowing for agents to have a theory of their own problem solving techniques, and providing for agents to have theories of other agents knowledge and problem solving capabilities. The explicit representation is based on the concepts of general problem solving contexts and of contexts appropriate for specific problems extracted from problem descriptions. Solutions are contexts that can answer relevant questions correctly, constructed using reasoning principles considered appropriate. Appropriate principles include principles found in general problem solving contexts as well as problem specific principles extracted from problem descriptions.

The second technique is to use the notion of *resource* in the formulation of principles for attribution of knowledge to agents. This allows an agent to use axioms such as **K** in problem solving without being forced to attribute logical omniscience to any agent, including himself. Our approach involves understanding the nature of resources and of the role that they play in controlling of the use of logic. We consider resources to be abstract entities equipped with ordering and inaccessibility relations such as $\lll$. The role of $\lll$ and the other predicates introduced in § 4 is not just to allow us to write down a cute axiom. The point is to find concepts that allow us to write principles that we can agree on, without a completely formal understanding of the meaning of the concepts. Such principles allow us to carry out proofs when sufficiently detailed information is known. They act as constraints on problem solving theories. Additional information can be added as long as it is compatible with these constraints.

One of the new principles we formulated above is the *comparable resources* principle. It is this principle that allows the use of apparently unsound resource unlimited reasoning to attribute knowledge to agents. And it also justifies mixing resource limited reasoning with resource unlimited reasoning.

There are many other principles about resources to be discovered. We believe that the following principle is reasonable. It says that no matter what resource you have, there is always a tautology that can not be proved using only this amount of resource. Thus, this principle expresses a form of finiteness of resources.

$$\forall r \exists w.(Taut(w) \wedge \forall \pi.\neg Proves_R(\pi, w, r))$$

where the predicate $Proves_R$ formalizes provability in a proof system with resources made explicit in the inference rules and π is a proof of the formula w that uses resources r. A similar principle,

$$\forall r_1 . \exists r_2 . r_1 \lll r_2$$

expresses that there is no bound on inaccessibility.

We have not found it necessary to define a metric on resources. Using principles about resources without using a metric is one of the strengths of our approach.

ACKNOWLEDGMENTS

We would like to thank Gigina Aiello, Alessandro Cimatti, Ian Mason, and Daniele Nardi for critical comments on early versions of this paper that lead to substantial improvements in the presentation. We also thank several anonymous reviewers for helpful criticisms and Saša Buvač for his editorial advice and patience.

This research was partially supported by ARPA contract NAG2-703, and ARPA/ONR grant N00014-94-1-0775, by the ESPRIT Basic Research Action N.6810-COMPULOG II, by the Progetto Finalizzato Sistemi Informatici e Calcolo Parallelo of the CNR (Italian Research Council), and by IBUKI.

References

L. Aiello, D. Nardi, and M. Schaerf. Reasoning about reasoning in a meta-level architecture. *Journal of Applied Intelligence*, 1:55–67, 1991.

G. Attardi and M. Simi. Metalanguage and reasoning across viewpoints. In *Proceedings of the European Conference on Artificial Intelligence, Pisa*, pages 315–324, 1984.

G. Attardi and M. Simi. Communication across viewpoints. *Journal of Logic, Language and Information*, 1997. This issue.

J. Barwise. Scenes and other situations. *Journal of Philosophy*, LXXVIII(7):369–397, 1981.

S. Buvač. Quantificational logic of context. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96)*, pages 600–606, 1996.

M. Cadoli and M. Schaerf. Approximate reasoning and non-omniscient agents. In *Proceedings of the Fourth Conference on Theoretical Aspects of Reasoning about Knowledge (TARK-92)*, pages 169–183. Morgan Kaufmann, 1992.

J. J. Elgot-Drapkin and D. Perlis. Reasoning situated in time I: basic concepts. *Journal of Experimental and Theoretical Artificial Intelligence*, 2:75–98, 1990.

J. J. Elgot-Drapkin, M. Miller, and D. Perlis. Memory, reason and time: The step-logic approach. In R. Cummins and J. Pollock, editors, *Philosophy and AI: Essays at the Interface*, chapter 4, pages 79–103. MIT Press, 1991.

J. J. Elgot-Drapkin. Step-logic and the three-wise-men problem. In *Proceedings of the Ninth National Conference on Artificial Intelligence (AAAI-91)*, pages 412–417, 1991.

R. Fagin, J. Y. Halpern, Y. Moses, and M. Y. Vardi. *Reasoning about Knowledge.* The MIT Press, Cambridge, Mass., 1995.

G. Gamow and M. Stern. *Puzzle Math.* Vintage Press, 1958.

F. Giunchiglia and L. Serafini. Multilanguage hierarchical logics (or: how we can do without modal logics). *Artificial Intelligence*, 65:29–70, 1994.

F. Giunchiglia and P. Traverso. A metatheory of a mechanized object theory. *Artificial Intelligence*, 80:197–241, 1996.

F. Giunchiglia and T. Walsh. A theory of abstraction. *Artificial Intelligence*, 56:323–390, 1992.

F. Giunchiglia. Multilanguage systems. In *Proceedings of AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, 1991. Also IRST-Technical Report no. 9011-17.

F. Giunchiglia. Contextual reasoning. *Epistemologia, special issue on I Linguaggi e le Macchine*, XVI:345–364, 1993.

J. Y. Halpern and Y. Moses. A guide to completeness and complexity for modal logic of knowledge and belief. *Artificial Intelligence Journal*, 54:319–379, 1992.

J. Y. Halpern and M. Y. Vardi. Model checking vs. theorem proving: A manifesto. In *Proceedings of the Second International Conference on the Principles of Knowledge Representation and Reasoning (KR-91)*, 1991. Also in Lifshitz V. *Artificial Intelligence and Mathematical Theory of Computation. Papers in Honor of John McCarthy*, Academic Press, San Diego, 1991.

J. Y. Halpern. Reasoning about knowledge: A survey circa 1991. In A. A. Kent and J. G. Williams, editors, *Encyclopedia of Computer Science and Technology.* Marcel Dekker, 1991. Preliminary version on *Proceedings of the Conference on Theoretical Aspects of Reasoning about Knowledge* (TARK), Morgan Kaufmann Publisher, 1986.

J. Hintikka. *Knowledge and belief.* Cornell University Press, Ithaca, New York, 1962.

S. A. Kripke. Semantical considerations on modal logic. *Acta Philosophica Fennica*, 16:83–94, 1963.

H. J. Levesque. A logic of implicit and explicit belief. In *Proceedings of the Fourth National Conference on Artificial Intelligence (AAAI-84)*, pages 198–202, 1984.

P. Maes and D. Nardi, editors. *Meta-level Architectures and Reflection.* Elsevier Science Publishers B.V. (North Holland), 1988.

J. McCarthy and S. Buvač. Formalizing Context (Expanded Notes). Technical Note STAN-CS-TN-94-13, Stanford University, 1994.

J. McCarthy. Generality in artificial intelligence. *Communications of the ACM*, 30, 1987.

R. Montague. Pragmatics. In R. Kalibansky, editor, *Contemporary Philosophy*, pages 101–121. La Nuova Italia Editrice, 1968.

C. L. Talcott and R. W. Weyhrauch. Towards a theory of mechanized reasoning I: FOL contexts, an extensional view. In *Proc. of the 8th European Conference on Artificial Intelligence, (ECAI-90)*, pages 634–639, 1990.

A. G. B. ter Meulen. Content in context. In *Formalizing Context*, AAAI-95 Fall Symposium, pages 97–101, 1995.

A. G. B. ter Meulen. *Representing Time in Natural Language.* The MIT Press, 1995.

R. W. Weyhrauch and C. L. Talcott. The logic of FOL systems: Formulated in set theory. In M. Hagiya, N. D. Jones, and M. Sato, editors, *Festschrift in honor of Professor Satoru Takasu*, number 792 in Lecture Notes in Computer Science. Springer-Verlag, 1994.

R. W. Weyhrauch and C. L. Talcott. FOL home page, 1995. URL = http://www-formal.stanford.edu/FOL/home.html.

R. W. Weyhrauch and C. L. Talcott. The logic of FOL systems: Introduction, 1995. URL = http://www-formal.stanford.edu/FOL/home.html.

R. W. Weyhrauch and C. L. Talcott. Problem solving contexts, 1995. URL = http://www-formal.stanford.edu/FOL/home.html.

R. W. Weyhrauch, M. Cadoli, and C. L. Talcott. Deducing logical omniscience, 1995. URL = http://www-formal.stanford.edu/FOL/home.html.

R. W. Weyhrauch. A Users Manual for Fol. Technical Report STAN-CS-77-432, Stanford University Computer Science Department, 1977.

R. W. Weyhrauch. Prolegomena to a theory of formal reasoning. *Artificial Intelligence*, 13:133–170, 1980.