Reprinted from Working Notes of the AAAI Spring Symp. on Log. Formalizations of Commonsense Reas.(1991)33--40

A Real-time Solution to the Wise-men Problem

Jennifer J. Elgot-Drapkin
Department of Computer Science and Engineering
College of Engineering and Applied Sciences
Arizona State University
Tempe, AZ 85287-5406

1 Background

Honest-to-goodness commonsense reasoners have finite resources and hence also limited reasoning capabilities. Traditional logically complete logics therefore cannot be used to satisfactorily represent the reasoning of these agents. This observation has brought about tremendous strides in the area of *limited reasoning*. We mention here several such approaches.

Konolige [Kon84] distinguishes three types of limited reasoning: relevance incompleteness, resource-limited incompleteness, and fundamental logical incompleteness. Konolige uses the idea of relevance incompleteness, the notion that a reasoner may not know all the relevant rules in order to solve a problem, to get an interesting solution to the *Three-wise-men problem* (see [Kon84]). In this paper we present our own solution to this problem[1], in which we focus on an aspect that Konolige has left unaddressed.

The second type of limited reasoning that Konolige describes is that of resource-limited incompleteness: an agent may have the inferential capability to derive some consequence of his beliefs, yet does not have the computational resources to do so. This is a difficulty that arises, for instance, in a chess-playing program.

The third and final type of limited reasoning that Konolige describes is that of fundamental logical incompleteness. This arises in a reasoner that may simply not reason about beliefs of other agents at all, making a solution to the *Three-wise-men problem* impossible.

Levesque [Lev84] has given an intuitively plausible semantic account of *implicit* and *explicit* beliefs. An agent's implicit beliefs include all valid formulas, his explicit beliefs, and the logical consequences of his explicit beliefs. His explicit beliefs, on the other hand, are closed under a much weaker set of conditions---in particular, they may not be deductively closed. For example, an agent does not necessarily explicitly believe all valid formulas, nor does it necessarily explicitly believe β, simply because it explicitly believes α and $\alpha \rightarrow \beta$. Using the set of explicit beliefs, Levesque is able to describe a limited reasoner.

Levesque's logic, however, does not allow meta-reasoning about one's own beliefs or reasoning about other agents' beliefs. These abilities are needed in many situations, including planning and goal-directed behavior, where one may have to reason about the knowledge that one has as well as the knowledge that others may possess. Fagin and Halpern [FH88] extend Levesque's notion of implicit and explicit beliefs to allow for multiple agents and beliefs of beliefs. They introduce a notion of *awareness*, based on the idea that one cannot have beliefs about something of which one has no knowledge. Interestingly, the implicit belief operator acts like the classical belief operator, so that, for instance, if one assumes that the agents are aware of

[1]For simplicity, we describe a solution to the less complex *Two-wise-men problem* in this paper. A similar, but more complex, solution to the *Three-wise-men problem* can be found in [ED88, ED91].

all formulas, the logic reduces to the classical logic of belief, weak S5 (see [Che80]).

These approaches all model *limited* reasoning, yet the process is in terms of the standard mold of *static* reasoning---although there is a restricted view of what counts as a theorem, the logic is still *final-tray-like*.[2] That is, all conclusions are taken to be drawn instantaneously, without regard to the sequence of deductions involved in actually producing those theorems. The emphasis instead is on some ''final'' state of reasoning. Although the final tray is smaller in a limited reasoning approach than in the conventional *omniscient* approach (it is catching less, if you will), it is still only the final set of consequences that are evident. In Fagin and Halpern's logic of general awareness [FH88], for example, $\alpha, \alpha \rightarrow \beta, \alpha \rightarrow \gamma$ and γ may all appear in the tray without β, given that the agent is unaware of β. Although the tray is catching less here, the over-simplification of a final state of reasoning is nonetheless maintained. All the conclusions are still drawn instantaneously; the individual deductive steps are not taken into consideration.

Step-logic was proposed in [ED88, EDP90] as an alternative to the approaches to limited reasoning just discussed, where it is *not* a final tray of conclusions in which one is interested, but rather the ever-changing set of conclusions drawn along the way. That is, step-logic was designed to model reasoning that focuses on the *on-going process* of deduction; there is no final state of reasoning---no final tray of conclusions.

There are many examples of situations in which the effort or time spent making deductions is crucial. Consider Little Nell who has been tied to the railroad tracks. A train is quickly approaching. Dudley must save her. (See [Haa85, McD82].) It is not appropriate for Dudley to spend hours figuring out a plan to save Nell; she will no longer need saving by then. Thus if we are to model Dudley's reasoning, we must have a mechanism that takes into account the passage of time as the agent is reasoning.

The *Wise-men Problem* (described in Section 3) is another example in which the effort involved in making deductions is critical. In this paper we show how step-logic can be used to arrive at a more temporally plausible account of the wise men's reasoning. In other formalizations of the *Wise-men Problem* this time aspect has been ignored. (See [Kon84, KL87, Kon90].)

[2] See [ED88] for a more detailed description of this phenomenon.

2 Step-logic

Step-logic serves as a model of a reasoning agent with the ability to reason about the passage of time *as* it is reasoning. Intuitively, we view a reasoning agent as an inference mechanism that may be given external inputs or observations. Inferred wffs are called beliefs; these may include certain observations. We regard the reasoning process as occurring over a sequence of discrete *steps*. The reasoner starts out with an empty set of beliefs at step 0. ''Observations'' (external inputs to the system) may arise at any step. When an observation appears, it is considered a belief. From these beliefs, new beliefs may be concluded. At some step i the reasoner may have belief α, concluded based on earlier beliefs or arising at step i directly as an observation. These time parameters i are allowed to figure in the on-going reasoning itself. In particular, in some step-logics---although not ones to be used in this paper---at step i the observation $Now(i)$ is made. We think of each step of reasoning as representing a given fixed interval of time, so that for instance, after 10 steps of reasoning have occurred, so have 10 units of time.

In [DP86, ED88] we defined a family of eight step-logics---SL_0, SL_1, ..., SL_7---arranged in increasing sophistication, each designed to model the reasoning of a reasoning agent. Each differs in the capabilities that the agent has. In an SL_0 step-logic, for instance, the reasoner has no knowledge of the passage of time as it is reasoning, it cannot introspect on its beliefs, and it is unable to retract former beliefs. (SL_0 is not very useful for modeling commonsense reasoners.) In an SL_7 step-logic, by contrast, the agent is capable of all three of these aspects that are so critical to commonsense reasoning. Most commonsense reasoners seem to need the full capabilities

of an SL_7 step-logic.

A step-logic is characterized by a language, observations, and inference rules. Step-logic is *deterministic* in that at each step i *all* possible conclusions from one application of the rules of inference applied to the previous steps are drawn (and therefore are among the wffs at step i). However, for real-time effectiveness and cognitive plausibility, at each step we want only a finite number of conclusions to be drawn.

Let $\mathcal{L}$ be a first-order or propositional language, and let $\mathcal{W}$ be the set of wffs of $\mathcal{L}$.

Definition 1 *An* observation-function *is a function* $OBS : \mathbf{N} \rightarrow \mathcal{P}(\mathcal{W})$*, where* $\mathcal{P}(\mathcal{W})$ *is the power set of* $\mathcal{W}$*, and where for each* $i \in \mathbf{N}$*, the set* $OBS(i)$ *is finite. If* $\alpha \in OBS(i)$*, then* α *is called an* i-observation.

Definition 2 *A* history *is a finite tuple of pairs of finite subsets of* $\mathcal{W}$*.* $\mathcal{H}$ *is the set of histories.*

Definition 3 *An* inference-function *is a function* $INF : \mathcal{H} \rightarrow \mathcal{P}(\mathcal{W})$*, where for each* $h \in \mathcal{H}$*,* $INF(h)$ *is finite.*

Intuitively, a history is a conceivable temporal sequence of belief-set/observation-set pairs. The history is a *finite* tuple; it represents the temporal sequence up to a certain point in time. The inference-function extends the temporal sequence of belief sets by one more step beyond the history.

Definition 4 *An* SL_n-theory over a language $\mathcal{L}$ *is a triple,* $< \mathcal{L}, OBS, INF >$*, where* $\mathcal{L}$ *is a first-order language,* OBS *is an observation-function, and* INF *is an inference-function. We use the notation,* $SL_n(OBS, INF)$*, for such a theory (the language* $\mathcal{L}$ *is implicit in the definitions of* OBS *and* INF*).*

For more background on step-logic, see [ED88, EDP90].

3 The Problem

We present a variation of the classic wise-men-problem which was first introduced to the AI literature by McCarthy in [McC78]. This version best illustrates the type of reasoning that is so characteristic of commonsense reasoners.

A king wishes to know whether his three advisors are as wise as they claim to be. Three chairs are lined up, all facing the same direction, with one behind the other. The wise men are instructed to sit down. The wise man in the back (wise man #3) can see the backs of the other two men. The man in the middle (wise man #2) can only see the one wise man in front of him (wise man #1); and the wise man in front (wise man #1) can see neither wise man #3 nor wise man #2. The king informs the wise men that he has three cards, all of which are either black or white, at least one of which is white. He places one card, face up, behind each of the three wise men, explaining that each wise man must determine the color of his own card. Each wise man must announce the color of his own card as soon as he knows what it is. (The first to correctly announce the color of his own card will be aptly rewarded.) All know that this will happen. The room is silent; then, after several minutes, wise man #1 says ''My card is white!''.

We assume in this puzzle that the wise men do not lie, that they all have the same reasoning capabilities, and that they can all think at the same speed. We then can postulate that the following reasoning took place. Each wise man knows there is at least one white card. If the cards of wise man #2 and wise man #1 were black, then wise man #3 would have been able to announce immediately that his card was white. They all realize this. Since wise man #3 kept silent, either wise man #2's card is white, or wise man #1's is white. At this point wise man #2 would be able to determine, if wise man #1's were black, that his card was white. They all realize this. Since wise man #2 also remains silent, wise man #1 knows his card must be white.

Thus we see that it is important to be able to reason in the following manner:

If such and such were true *at that time*,

then so and so *would have realized it by this time.*

For instance, if wise man #2 is able to determine that wise man #3 would have *already* been able to figure out that wise man #3's card is white, and wise man #2 has heard nothing, then wise man #2 knows wise man #3 does *not* know the color of his card. Step-logic is particularly well-suited to this type of deduction since it focuses on the actual individual deductive steps. Others have studied this problem (e.g. see [Kon84, KL87, Kon90]) from a final-tray perspective and thus are not able to address this temporal aspect of the problem: assessing what others have been able to conclude *so far*.

We show here how step-logic can be used to model a version of this problem in which there are only two men. A model of the problem in which there are three men is presented in [ED88, ED91].

4 Formulation using Step-logic

In the two-wise-men puzzle the king has just two wise men and two cards, at least one of which is white. Again wise man #2 sits behind wise man #1, so wise man #1 can see nothing, and wise man #2 can see wise man #1's card. Wise man #2 is unable to identify the color of his card. Wise man #1 is then able to determine that his card must be white.

The reasoning involved in this version of the puzzle is much simpler than in the three-wise-men version. Wise man #2 can see the color of wise man #1's card. If it were black, then wise man #2 would know, since there is at least one white card, that his card was white. Wise man #1 knows this. Yet wise man #2 says nothing. Since wise man #2 is silent, it must be the case that wise man #1's card is not black, but rather white.

The step-logic used to model this problem is defined in Figure 1. The problem is modeled from wise man #1's point of view. The observation-function contains all the axioms that wise man #1 needs to solve the problem, and the inference-function provides the allowable rules of inference.

We use what is called an SL_5 step-logic.[3] The language of SL_5 is first-order, having binary predicate symbols K_j and U, and function symbol s. $K_j(i, \text{`}\alpha\text{'})$ is interpreted as "$\alpha$ is known[4] by agent $j$ at step $i$". Note that this gives the agent the expressive power to introspect on his own beliefs as well as the beliefs of others. $U(i, \text{`}x\text{'})$ [5] expresses the fact that an utterance of x is made (by wise man #2) at step i. $s(i)$ is the successor function (where $s^k(0)$ is used as an abbreviation for $\underbrace{s(s(\cdots(s}_{k}(0))\cdots))$). W_i and B_i express the facts that i's card is white, and i's card is black, respectively.

In the particular version of step-logic that is used, the formulas that the agent has at step i (the *i-theorems*) are precisely all those that can be deduced from step $i-1$ using one application of the applicable rules of inference. As previously stated, the agent is to have only a finite number of theorems (conclusions, beliefs, or simply wffs) at any given step. We write:

$$\begin{aligned} i: &\quad \alpha \\ i+1: &\quad \beta \end{aligned}$$

to mean that α is an i-theorem, and β is an $i+1$-theorem. There is no implicit assumption that α (or any other wff other than β) is present (or not present) at step $i+1$. Wffs are not assumed to be inherited or retained in passing from one step to the next, unless explicitly stated in an inference rule. Rule 6 in Figure 1 provides an unrestricted form of inheritance.

The axioms in $OBS_{W_2}(1)$ have the following intuitive meaning. (Refer to Figure 1.) Wise man #1 knows the following:

a. Wise man #2 uses the rule of *modus ponens*.

b. Wise man #2 knows at step 1 that if my card is black, then his is white.

[3] For more details on SL_n step-logics, see [ED88].

[4] known, believed, or concluded. The distinctions between these (see [Get63, Per86, Per88]) are not addressed here.

[5] For simplicity, in the remainder of the paper we drop the quotes around the second argument of predicates U and K_j.

OBS_{W_2} is defined as follows.

$$OBS_{W_2}(i) = \begin{cases} \left\{ \begin{array}{l} (\forall i)(\forall x)(\forall y)[K_2(i, x \rightarrow y) \rightarrow (K_2(i, x) \rightarrow K_2(s(i), y))] \\ K_2(s(0), B_1 \rightarrow W_2) \\ (B_1 \rightarrow K_2(s(0), B_1)) \\ (\neg B_1 \rightarrow W_1) \\ (\forall i)[\neg U(s(i), W_2) \rightarrow \neg K_2(i, W_2)] \\ (\forall i)[\neg K_1(s(i), U(i, W_2)) \rightarrow \neg U(i, W_2)] \end{array} \right\} & \text{if } i = 1 \\ \emptyset & \text{otherwise} \end{cases}$$

The inference rules given here correspond to an inference-function, INF_{W_2}. For any given history, INF_{W_2} returns the set of all immediate consequences of Rules 1--6 applied to the last step in that history. Note that Rule 5 is the only default rule.

Rule 1 :	$\dfrac{i : \ldots}{i+1 : \ldots, \alpha}$	if $\alpha \in OBS(i+1)$
Rule 2 :	$\dfrac{i : \ldots, \alpha, (\alpha \rightarrow \beta)}{i+1 : \ldots, \beta}$	Modus ponens
Rule 3 :	$\dfrac{i : \ldots, P_1\overline{a}, \ldots, P_n\overline{a}, (\forall \overline{x})[(P_1\overline{x} \wedge \ldots \wedge P_n\overline{x}) \rightarrow Q\overline{x}]}{i+1 : \ldots, Q\overline{a}}$	Extended modus ponens
Rule 4 :	$\dfrac{i : \ldots, \neg\beta, (\alpha \rightarrow \beta)}{i+1 : \ldots, \neg\alpha}$	Modus tolens
Rule 5 :	$\dfrac{i : \ldots}{i+1 : \ldots, \neg K_1(s^i(0), U(s^{i-1}(0), W_2))}$	if $U(s^{i-1}(0), W_2) \notin \vdash_i,\ i > 1$
Rule 6 :	$\dfrac{i : \ldots, \alpha}{i+1 : \ldots, \alpha}$	Inheritance

Figure 1: OBS_{W_2} and INF_{W_2} for the Two-wise-men Problem

c. If my card is black, then wise man #2 knows this at step 1.

d. If my card is not black, then it is white.

e. If there is no utterance at a given step by wise man #2 that his card is white, then wise man #2 didn't know that his card was white (i.e. W_2) at the previous step.

f. If I don't know at a given step that there has been an utterance of W_2, then there was no utterance of W_2 at the previous step. (I would know that an utterance of W_2 was made one step after it is uttered.)

Note the following concerning the inference rules:

1. Rule 5 is a rule of introspection. Wise man #1 can introspect on what utterances have been made.

2. The rule of inheritance is quite general: *everything* is inherited from one step to the next.[6]

3. The rule for extended modus ponens allows an arbitrary number of variables.

5 The Solution

The solution to the problem is given in Figure 2. The step number is listed on the left. The reason (inference rule used) for each deduction is listed on the right. To allow for ease of reading, only the wffs in which we are interested are shown at each step. In addition, none of the inherited wffs are shown. This means that a rule appears to be operating on a step other than the previous one; the wffs involved have, in fact, actually been inherited to the appropriate step.

In step 1, we see that all the initial axioms ($OBS_{W_2}(1)$) have been inferred through the use of Rule 1. The wff in step 2 has been deduced through the use of Rule 3. It says that if wise man #2 knows B_1 at step 1, then wise man #2 will know W_2 at step 2. No new (significant) inferences are made in steps 3 and 4. At step 5, wise man #1 negatively introspects to determine that no utterance of W_2 was made at step 3. Note the time delay: wise man #1 is able to prove *at step 5* that he did not know *at step 4* of an utterance made *at step 3*.[7] At step 6, wise man #1 can then conclude that indeed no utterance of W_2 was made at step 3. The reasoning continues from step to step, and in step 10, wise man #1 is finally able to prove that, because nothing has been said by now (actually by step 3), his card must be white.

We see that step-logic is a useful vehicle for formulating and solving a problem of this kind in which the time that something occurs is important. Wise man #1 does indeed determine ''if wise man #2 knew the color of his card, he would have announced it by now.'' (See steps 6 and 7.) Wise man #1 then reasons backwards from here to determine that his card must not be black (step 9), and hence must be white (step 10).

It is interesting to note that wise man #1 needs to know very little about wise man #2 and how he reasons. We could, for instance, have given wise man #1 the information that wise man #2 is just as clever as he, and thus has all the same rules of inference. This much knowledge was not necessary, however. The only inference rule of wise man #2's of which wise man #1 had to be aware was *modus ponens*. Wise man #1 needed to know three additional facts about wise man #2:

1. that wise man #2 could see the color of wise man #1's card

2. that wise man #2 knew there was at least one white card

3. that wise man #2 would announce the color of his card as soon as he knew it.

Note that wise man #1 needs to know only that wise man #2 knows points 1 and 2 *at step 1*. Wise man #1 doesn't need to know that wise man #2 may know these two facts later in time as well. It turns out that this is *not* sufficient in the *Three-wise-men*

[6] For other commonsense reasoning problems, a more restrictive version of inheritance is necessary.

[7] For a detailed description of this phenomenon, see [ED88].

0:		$\emptyset$	
1:	(a)	$(\forall i)(\forall x)(\forall y)[K_2(i, x \rightarrow y) \rightarrow (K_2(i, x) \rightarrow K_2(s(i), y))]$	(R1)
	(b)	$K_2(s(0), B_1 \rightarrow W_2)$	(R1)
	(c)	$(B_1 \rightarrow K_2(s(0), B_1))$	(R1)
	(d)	$(\neg B_1 \rightarrow W_1)$	(R1)
	(e)	$(\forall i)[\neg U(s(i), W_2) \rightarrow \neg K_2(i, W_2)]$	(R1)
	(f)	$(\forall i)[\neg K_1(s(i), U(i, W_2)) \rightarrow \neg U(i, W_2)]$	(R1)
2:		$(K_2(s(0), B_1) \rightarrow K_2(s^2(0), W_2))$	(R3,1a,1b)
3:		(no new deductions of interest)	
4:		(no new deductions of interest)	
5:		$\neg K_1(s^4(0), U(s^3(0), W_2))$	(R5)
6:		$\neg U(s^3(0), W_2)$	(R3,5,1f)
7:		$\neg K_2(s^2(0), W_2)$	(R3,6,1e)
8:		$\neg K_2(s(0), B_1)$	(R4,7,2a)
9:		$\neg B_1$	(R4,8,1c)
10:		W_1	(R2,9,1d)

Figure 2: Solution to the Two-wise-men Problem

problem---in the more complex problem, wise man #1 must know that wise man #2 knows these facts at all points in time. (See [ED88, ED91].)

6 Summary

Step-logic is a suitable theory for modeling commonsense reasoning problems in which it is necessary to recognize that the reasoning process itself takes time. The *Two-wise-men problem* is one such example. The problem is modeled from wise man #1's point of view, where OBS_{W_2} represents the information that wise man #1 needs to know in order to solve the problem. The notion that wise man #1 is able to reason about how long it takes for wise man #2 to reason is missing from other approaches to this problem. A similar but more complicated solution is given for the *Three-wise-men problem* in [ED88, ED91].

Step-logic has been used to solve other commonsense reasoning problems, including Moore's *Brother Problem* (see [Moo83, ED88, EDP90]). The solutions to both the *Wise-men problem* and the *Brother Problem* have been implemented on an IBM PC-AT using Arity Prolog (see [ED88] for details).

7 Acknowledgments

We would like to thank Don Perlis, Kevin Gary, and Laurie Ihrig for helpful comments.

References

[Che80] B. Chellas. *Modal Logic*. Cambridge University Press, 1980.

[DP86] J. Drapkin and D. Perlis. Step-logics: An alternative approach to limited reasoning. In *Proceedings of the European Conf. on Artificial Intelligence*, pages 160--163, 1986. Brighton, England.

[ED88] J. Elgot-Drapkin. *Step-logic: Reasoning Situated in Time*. PhD thesis, Department of Computer Science, University of Maryland, College Park, Maryland, 1988.

[ED91] J. Elgot-Drapkin. Step-logic and the three-wise-men problem. Manuscript in progress, 1991.

[EDP90] J. Elgot-Drapkin and D. Perlis. Reasoning situated in time I: Basic concepts. *Journal of Experimental and Theoretical Artificial Intelligence*, 2(1):75--98, 1990.

[FH88] R. Fagin and Y. Halpern, J. Belief, awareness, and limited reasoning. *Artificial Intelligence*, 34(1):39--76, 1988.

[Get63] E. Gettier. Is justified true belief knowledge? *Analysis*, 23:121--123, 1963.

[Haa85] A. Haas. Possible events, actual events, and robots. *Computational Intelligence*, 1(2):59--70, 1985.

[KL87] S. Kraus and D. Lehmann. Knowledge, belief and time. Technical Report 87-4, Department of Computer Science, Hebrew University, Jerusalem 91904, Israel, April 1987.

[Kon84] K. Konolige. Belief and incompleteness. Technical Report 319, SRI International, 1984.

[Kon90] K. Konolige. Explanatory belief ascription. In R. Parikh, editor, *Theoretical Aspects of Reasoning about Knowledge: Proceedings of the Third Conference*, pages 85--96. Morgan Kaufmann, 1990. Pacific Grove, CA.

[Lev84] H. Levesque. A logic of implicit and explicit belief. In *Proceedings of the 3rd National Conf. on Artificial Intelligence*, pages 198--202, 1984. Austin, TX.

[McC78] J. McCarthy. Formalization of two puzzles involving knowledge. Unpublished note, Stanford University, 1978.

[McD82] D. McDermott. A temporal logic for reasoning about processes and plans. *Cognitive Science*, 6:101--155, 1982.

[Moo83] R. Moore. Semantical considerations on nonmonotonic logic. In *Proceedings of the 8th Int'l Joint Conf. on Artificial Intelligence*, 1983. Karlsruhe, West Germany.

[Per86] D. Perlis. On the consistency of commonsense reasoning. *Computational Intelligence*, 2:180--190, 1986.

[Per88] D. Perlis. Languages with self reference II: Knowledge, belief, and modality. *Artificial Intelligence*, 34:179--212, 1988.