CIRCUMSCRIPTIVE IGNORANCE

Kurt Konolige

Artificial Intelligence Center
SRI International

ABSTRACT

In formal systems that reason about knowledge, inferring that an agent actually does *not* know a particular fact can be problematic. Collins [1] has shown that there are many different modes of reasoning that a subject can use to show that he is ignorant of something; some of these, for example, involve the subject reasoning about the limitations of his own information-gathering and memory abilities. This paper will consider a single type of inference about ignorance, which we call *circumscriptive ignorance*. We present a partial formalization of circumscriptive ignorance and apply it to the Wise Man Puzzle.[1]

1. Circumscriptive Ignorance

The premise that there is a limited amount of information, resources, or strategies available for the solution of a problem is often an unstated but essential part of the problem abstraction. For example, in the Missionaries and Cannibals Puzzle, it is important that only a single boat is available to ferry people across the river; one cannot invoke a helicopter brigade from the Sudan to solve the puzzle. McCarthy [5] has investigated the first-order formalization of problem statements such as this, using a *circumscription schema* to capture unstated limitations on resources.

In puzzles that involve reasoning about the knowledge agents possess, there are often unstated conditions on the initial information given an agent, as well as on the information he can acquire. In the Wise Man Puzzle (see Section 3 below for a full statement of this puzzle), it is common knowledge that each man can see his neighbors' spots and knows from the king that there is at least one white spot. It is an unstated condition of the puzzle that this is the only knowledge that wise men have about the initial situation; in an adequate formalization of the puzzle it should be possible to prove that each wise man is ignorant of the color of his own spot in the initial situation. In effect, the knowledge that is available to the agents in the puzzle is being circumscribed; informally one would say "The only facts that agent S knows about proposition p are F." If from the facts F it is not possible for S to infer p (or $\sim p$), then S does not know whether p is true. Proving ignorance based on a limitation on the knowledge available to an agent will be called reasoning by circumscriptive ignorance.

Circumscriptive ignorance, especially for the Wise Man Puzzle, has been formalized in first-order axiomatizations of a possible-world semantics for knowledge (Goad [2], McCarthy [6], and Sato [7]). However, there has been no formalization in a modal logic of knowledge. The advantages of using a modal formalization are clarity of expression and the ability to use inference procedures, including decision procedures, that have been developed for modal logic. In the next few sections we outline a modal formalization of circumscriptive ignorance.

2. The Modal Logic $KI4$

The modal logic we shall use is a propositional modal logic based on Sato's $K4$ [7], which includes an axiomatization of common knowledge.[2] $K4$ is a family of languages parameterized by the choice of propositional letters Pr and agents Sp. $0 \in Sp$ is a reserved name for *Fool*, a fictitious agent whose knowledge is common knowledge. For a particular choice of Pr and Sp, the language $K4$ is the propositional calculus over Pr, together with a set of indexed unary modal operators $[S]$, $S \in Sp$. The intended meaning of $[S]\alpha$ is that agent S knows α. The axiom schemata for $K4$ are

$$\begin{array}{ll} (A1) & \text{All propositional tautologies} \\ (A2) & [S]\alpha \supset \alpha \\ (A3) & [S]\alpha \supset [S][S]\alpha \\ (A4) & [S](\alpha \supset \beta) \supset ([S]\alpha \supset [S]\beta) \\ (A5) & [0]\alpha \supset [0][S]\alpha \quad , \end{array} \tag{1}$$

where α and β denote arbitrary sentences, and S denotes an arbitrary agent. Axioms $A1$–$A4$ give the system $S4$

[1] This short note describes current research collected in [4]. The work presented here was supported by grant N0014-80-C-0296 from the Office of Naval Research.

[2] For simplicity we use $K4$ rather than Sato's more complicated $KT4$, which deals explicitly with time.

for each modality $[S]$, while $A5$ is the common knowledge axiom: what any fool knows, any fool knows everyone knows. The two rules of inference are modus ponens and necessitation (from α, infer $[S]\alpha$).

In $K4$ and other modal logics of knowledge, an agent's knowledge is described as a theory, that is, as a set of formulas that contain the axioms and are closed under the rules of inference. To see this, note that all instances of $[S]\alpha$ for which α is an axiom of $K4$ are provable, and that modus ponens is implemented by $A4$. We shall use the term *agent's theory* to mean the set of formulas α for which $[S]\alpha$ can be proven in $K4$.

There is a difference between the axiomatization of an agent's theory and $K4$ itself. $K4$ is a Hilbert system in which no proper axioms are allowed; an agent's theory allows proper axioms (*i.e.*, we can assert formulas of the form $[S]p$).[3] We define the *α-theory of $K4$*, for a fixed sentence α, as the set of formulas β for which $\alpha \supset \beta$ is a theorem of $K4$. We write $\alpha \vdash_{\overline{K4}} \beta$ if β is in the α-theory of $K4$, *i.e.*, $\alpha \supset \beta$ is a theorem of $K4$.

$K4$ by itself is not sufficient to represent circumscriptive ignorance, since there is no way of limiting the proper axioms that could conceivably be used to derive knowledge in an agent's theory. If we look at a particular theory for an agent, where the only proper axiom is α, it is impossible to derive proofs of certain formulas within that theory; but there is also no way to express this in $K4$ itself.

To express circumscriptive ignorance, $K4$ is extended by a family of new unary modal operators indexed by sentences of $K4$. These are called *circumscriptive modalities*, and are written as $[\alpha]$, where α is a sentence of $K4$. The extended language is called $KI4$. In informal terms, $[\alpha]$ is intended to mean the α-theory of $K4$, that is, $[\alpha]\beta$ holds just in case β is in the α-theory of $K4$. Thus the notion of provability is explicitly introduced into $KI4$.

The axiomatization of $[\alpha]$ is problematic, since it involves formalizing not only which sentences of $K4$ are provable, but also which sentences are not. However, a sufficient set of axioms for $KI4$ can be obtained by making use of the intended interpretation of $[\alpha]$ as provability in $K4$. The axioms of $KI4$ are simply those of $K4$, together with the schemata

$$\begin{array}{ll} (A6) & [\alpha]\beta, \text{ where } \alpha \vdash_{\overline{K4}} \beta \\ (A7) & \sim[\alpha]\beta, \text{ where } \alpha \not\vdash_{\overline{K4}} \beta \quad . \end{array} \tag{2}$$

The rules of inference for $KI4$ are the same as $K4$.

The axioms (2) only form a recursive set if $K4$ is decidable; further, it is obvious that if $K4$ is decidable, then so is $KI4$. Sato [7] gives a proof that $K4$ is decidable, and decision procedures with reasonable computational properties can be found by analogy with those given by Kripke [3] for $S4$. These procedures are detailed in [4].

An alternative characterization of $A6$ that is more in the style of typical modal language axiomatization could be given as follows:

$$\begin{array}{ll} (A8) & [\alpha]\beta, \ \beta \text{ an instance of } A1\text{–}A5 \\ (A9) & [\alpha]\alpha \\ (A10) & [\alpha](\beta \supset \gamma) \supset ([\alpha]\beta \supset [\alpha]\gamma) \\ (A11) & [\alpha]\beta \supset [\alpha][S]\beta \quad . \end{array} \tag{3}$$

$A8$ and $A9$ ensure that all instances of the logical axioms of the α-theory of $K4$ are present; $A10$ and $A11$ are modus ponens and necessitation, respectively, for $K4$. However, there is no similar characterization of $A7$. We will use the schemata $A6$ and $A7$ directly in the remainder of this paper.

The circumscriptive quality of $[\alpha]$ comes from the restriction of its meaning to the α-theory of $K4$. If we take α to be a knowledge operator, this translates into a circumsciption of an agent's theory. For example, if $\alpha = [S]q$, $[\alpha]$ picks out the agent's theory for which q is the only proper axiom. More complicated statements are possible; for example, to say that the only knowledge S has about p is that he knows either q_1 or q_2, assert $[\alpha][S]p$, with $\alpha = [S]q_1 \vee [S]q_2$.

$KI4$ has several interesting properties related to its circumscriptive nature. For every atom of the form $[\alpha]\beta$, either it or its negation is provable in $KI4$. Every subset of the language $KI4$ whose atoms are all circumscription operators is *complete*: it has no consistent proper extension. Thus $KI4$ (with circumscription atoms only) has a single model, given by the theorems of $K4$.

3. The Wise Man Puzzle

The Wise Man Puzzle can be stated as follows [6]: *A king wishing to know which of his three wise men is the wisest, paints white dots on each of their foreheads, tells them that at least one spot is white, and asks each to determine the color of his own spot. After a while the wisest announces that his spot is white, reasoning as follows: "Suppose my spot were black. The second wisest of us would then see a black and a white and would reason that if his spot were black, the least wise would see two black spots and would conclude that his spot is white on the basis of the king's assurance. He would have announced it by now, so my spot must be white."* We simplify this puzzle by having the king ask each wise man in turn what color his spot is, starting with the least wise.

[3] This difference becomes apparent in the rules of inference: necessitation is *not* included as a rule of inference in an agent's theory, since it would allow the derivation of $[0]p$ from $[S]p$.

To formalize the puzzle, we use the language $KI4$ with $Sp = \{0, S_1, S_2, S_3\}$ and $Pr = \{p_1, p_2, p_3\}$. p_i is the sentence asserting that S_i has a white spot on his forehead. A handy abbreviation is $[\![S]\!]p \doteq [S]p \vee [S]\sim p$, *i.e.*, S knows whether or not p is true. Then the following axioms suffice for the initial conditions of the puzzle (without worrying about time):

$$\begin{aligned}
&(W1)\quad p_1 \wedge p_2 \wedge p_3 \\
&(W2)\quad [0](p_1 \vee p_2 \vee p_3) \\
&(W3)\quad [0]([\![S_1]\!]p_2 \wedge [\![S_1]\!]p_3 \wedge [\![S_2]\!]p_1 \wedge [\![S_2]\!]p_3 \\
&\qquad\qquad \wedge [\![S_3]\!]p_1 \wedge [\![S_3]\!]p_2) \\
&(W4)\quad [\alpha][\![S_1]\!]p_1 \equiv [\![S_1]\!]p_1, \\
&\qquad \text{where } \alpha = W2 \wedge W3 \wedge [S_1]p_2 \wedge [S_1]p_3.
\end{aligned} \tag{4}$$

$W2$ says that it's common knowledge that at least one spot is white; $W3$ says that everyone knows that each can see the spots of the others. $W4$ is the circumscription axiom: it says that S_1 knows whether p_1 holds solely on the basis of common knowledge ($W2$ and $W3$) and his own observations.

In this initial situation, it is possible to prove, through the use of $W4$, that S_1 does not know the color of his own spot. To see if $[\alpha][\![S_1]\!]p_1$ or its negation is provable in $KI4$, it suffices by $A6$ and $A7$ to apply the decision procedure for $K4$ to the sentence $\alpha \supset [\![S_1]\!]p_1$. The decision procedure looks for a $K4$-model of the negation of this sentence, and indeed finds one. Hence $\alpha \supset [\![S_1]\!]p_1$ is not provable in $K4$, and $\sim[\alpha][\![S_1]\!]p_1$ is a theorem of $KI4$ by $A6$. This in turn implies, by $W4$, that $\sim[\![S_1]\!]p_1$, *i.e.*, S_1 does not know the color of his own spot.

The second situation is similar, except that S_2 has heard S_1's reply to the king that he does not know his spot's color. The axioms for this situation are $W1$–$W3$, along with

$$\begin{aligned}
&(W5)\quad [S_2]\sim[\![S_1]\!]p_1 \\
&(W6)\quad [\alpha][\![S_2]\!]p_2 \equiv [\![S_2]\!]p_2, \\
&\qquad \text{where } \alpha = W2 \wedge W3 \wedge W5 \wedge [S_2]p_1 \wedge [S_2]p_3.
\end{aligned} \tag{5}$$

Again, by reasoning in the decidable theory $K4$, it is possible to show that $\sim(\alpha \supset [\![S_2]\!]p_2)$ has a model; thus S_2 does not know the color of his spot.

In the final situation, S_3 knows that S_2 does not know the color of his spot after hearing S_1's reply, and so S_3 is able to deduce that his own spot is white. The axioms here are $W1$–$W3$ and

$$\begin{aligned}
&(W7)\quad [S_3][S_2]\sim[\![S_1]\!]p_1 \\
&(W8)\quad [S_3]\sim[\![S_2]\!]p_2 \\
&(W9)\quad [\alpha][\![S_3]\!]p_3 \equiv [\![S_3]\!]p_3, \quad \text{where} \\
&\qquad \alpha = W2 \wedge W3 \wedge W7 \wedge W8 \wedge [S_3]p_1 \wedge [S_3]p_2.
\end{aligned} \tag{6}$$

It is possible to prove in $K4$ that $\alpha \supset [S_3]p_3$, and so $W9$ simply asserts that S_3 knows the color of his spot.

4. Conclusion

This brief note has introduced the idea of circumscriptive ignorance, showing how it could be formalized in a modal logic of knowledge called $KI4$. It then becomes possible to infer nonknowledge in the Wise Man Puzzle within this logic. The utility of $KI4$ is not limited to this puzzle, however; it should be possible within this logic to do general reasoning about the state of knowledge needed by an agent to derive various facts.

An interesting modification of $KI4$ occurs if the sentences in the circumscriptive operator are taken from $KI4$ itself, rather than $K4$. In this way agents could reason about the limitations of the state of knowledge of other agents. However, it is no longer obvious that $KI4$ would be decidable under this modification; further work needs to be done on this problem.

References

[1] Collins, A., "Fragments of a Theory of Human Plausible Reasoning," in *Proceedings of Theoretical Issues in Natural Language Processing* , Nash-Webber, B. and Schank, R. (eds.), Cambridge, Massachusetts (June 1975).

[2] Goad, C., "A Formal Representation for Situations Involving Knowledge," *unpublished note* , Stanford University, Stanford, California (December 1976).

[3] Hughes, G. E. and Cresswell, M. J., *An Introduction to Modal Logic,* Methuen and Company Ltd., London, 1968.

[4] Konolige, K., "Modal Logics for Belief," *unpublished notes* (February 1982).

[5] McCarthy, J., "Circumscription–A Form of Non-Monotonic Reasoning," *Artificial Intelligence* **13** (1980).

[6] McCarthy, J. *et al.*, "On the Model Theory of Knowledge," *Memo AIM-312* , Stanford University, Stanford (1978).

[7] Sato, M., *A Study of Kripke-type Models for Some Modal Logics by Gentzen's Sequential Method,* Research Institute for Mathematical Sciences, Kyoto University, Kyoto, Japan, July 1976.