PROVING FACTS ABOUT " I "

Michael Miller[a,b] and Donald Perlis[a-C]

[a]Computer Science Department [b]Systems Research Center [c]nstitute for Advanced Computer Studies

University -of Maryland College,Park, Maryland USA 20742

ABSTRACT

We study the *Knights and Knaves* problem, and find that for a proper treatment via theorem-proving, an interaction with natural language processing research is helpful. In particular, we discuss Ohlbach's claim that first-order logic is not well suited to handling this problem. Then we provide another interpretation of the problem using indexicals, and axiomatize it so that the desired result follows. We conclude by suggesting a broader context for dealing with "self-utterances" in automatic theorem-proving. Fuller details of automated proofs are given in a longer paper.

I. INTRODUCTION

The *Knights and Knaves* problem [Smullyan 1978] can be stated as follows: An island exists whose only inhabitants are knights, knaves, and a princess. The knights on the island always tell the truth, while the knaves always lie. Some of the knights are poor and the rest of them are rich. The same holds for the knaves. The princess is looking for a husband who must be a rich knave. In uttering one statement, how can a rich knave convince the princess that he is indeed a prospective husband for her?*

[Ohlbach 1985] is devoted to the framing and solution of this problem in a formal theorem-proving context using first-order logic (FOL). Though trying to write the problem in FOL may not appear to be difficult at first, it is shown by Ohlbach not to be entirely elementary. He examines, and finds inadequate, two different approaches before he finally settles on a third. This final approach, though successful in that it gets the desired "solution", is unsatisfactory because it involves constructs not faithfully related to the original problem.

This research has been supported in part by the following institutions:
National Science Foundation Grant #OIR-8500108
The U.S. Army Research Office (DAAG29-85-K-0177)
The Martin Marietta Corporation

*The intended solution is the statement "I am a poor knave" (or as used below, "I am not-rich and a knave"). The reader can readily verify that this indeed is a solution. Note the self-referential nature of the statement; this feature is a special case of indexicality, which we address below. For general treatments of self-reference, see [Perlis 1985], where another of Smullyan's puzzles is treated, and [Smith 1985].

We contend that there is a straightforward treatment of the problem that is faithful to its intent and that does allow a formal proof of the desired result. However, it requires employing concepts into the formalism that are not usually found in the context of problem-solving via resolution theorem-provers, namely ideas from natural language processing. Nevertheless, we are not replacing one trick by another, but rather introducing a well-understood and general formalism for problems of this sort.

II. PROBLEM REPRESENTATION

Finding a suitable representation for problems in artificial intelligence (AI) is often a difficult task. However, the *formalism* used to represent a problem is not necessarily the cause of the difficulty, though we grant that sometimes it is. Often it is the problem itself that is resisting representation and, when this occurs further insight into the problem is necessary.

The *Knights and Knaves* problem is a prime example of this. Ohlbach's interpretation of the problem results in him asking "Is there a statement x that I (being a rich knave) can say to convince the princess that I am indeed a rich knave?" Formally this might look like (and does in Ohlbach's second treatment):

OHL: $(\exists x)[CanSay(I,x) \leftrightarrow T(and(knave(I),rich(I))))]$**

where T is the predicate meaning True and "and" although a function symbol, intuitively takes two statements as arguments and returns another single conjunctive statement.***

This may appear to be a reasonable interpretation given the English statement of the problem. But as Ohlbach discusses, this representation (along with other associated axioms) is *not* sufficient to derive the intended result. Indeed, it is not hard to see the problem. The constant "I" stands for a fixed person (who is a rich knave). The point of the biconditional in OHL, and especially of the right-hand side, is to test whether the speaker is a rich knave, based on the

**Ohlbach's first treatment involved the goal $(\exists x)[CanSay(I,x) \rightarrow T(and(knave(I),rich(I))))]$ which (in addition to yielding a trivial and unhelpful answer) does not seem to correspond to his English interpretation "There exists a statement which I can say and which implies that I am really a rich knave." In fact, it seems to us that the goal statement $(\exists y)[CanSay(I,y) \& (\forall x)[CanSay(x,y) \rightarrow T(and(rich(x),knave(x))))]]$ comes much closer to the English.

***Actually, a name of the statement.

ability to say x. That is, the problem really seems to be asking, "What statement, when made by *anyone,* will convince the princess that *the person making the statement* is a rich knave." This then is the first problem with that representation: " I " should not be bound to a fixed individual, but should represent any "man in the street" who might state x. We then suggest the alternative version:

$$G_1: \quad (\exists x)(\forall y)[CanSay(y,x) \leftrightarrow T(and(rich(y),knave(y)))]$$

We claim to have now adequately represented the goal statement; but this is still not enough. For although this goal statement expresses what we want, there are other problems arising from the truth conditions of utterances which enter into other axioms in the problem representation.

HI. UTTERANCE INSTANCES

This brings us to what we think is the key issue in this pussle, and which has broader significance as well. Specifically, it is that utterances are instances of statement uses, and it is these instances that, in general, have truth-values, rather than the statement in and of itself. In particular, terms in a statement may have no definite reference outside of the context of an utterance. Although this is familiar to linguists* and philosophers (it is the so-called problem of indexicals) it is worth going into detail in the current paper, since the issue of representing knowledge in the *Knights and Knaves* problem hinges on this very phenomenon.

Typically, we think of a statement as being either true or false. This, however, is not always the case. For example, the statement:

"I am a knave"

will have a truth-value dependent upon who the speaker is; and so would be falsely uttered by any knight and truly by any knave.** Thus statements that contain indexicals (the word " I " in the above example) have meanings, and hence truth-values, that depend upon context.

Because of the indeterminacy of truth-values of sentences that contain indexicals, we will only refer to the truth-value of utterance-instances of such statements. An utterance-instance of a statement contains a context in which the statement was (or is) made including who the utterer is.

IV. "WHO AM I?"

If we look closely at any of Ohlbach's representations of the *Knights and Knaves* problem, we notice that the constant " I " seems to be playing two different roles. In all of his goal statements " I " is presumably used as a particular person. For example Ohlbach's second goal statement, OHL, illustrates this usage. On the other hand, in the intended solution to the problem, the "x" of the goal statement is bound to *anriki:*[***]

and(not(rich(I)),knave(I))

where, the same symbol " I " appears as before but now what is of interest is its presence inside the "x" in CanSay(I,x); i.e., as part of a potential utterance whose truth value depends upon who the speaker is. That is, any number of people might utter *anriki,* and its meaning would be different in each case. We now have an utterance-instance and need to know who " 1 " is before assigning a truth-value. Thus, " I " must be viewed as a pronoun and not a person here. In particular, it is the knighthood or "knavehood" of " I " which determines the truth of *anriki.* Of course, in the world in question, only knaves (and rich ones at that) could utter *anriki.* But that is the point; the princess must be able to deduce precisely that fact: that anyone at all who utters *anriki* must consequently be a rich knave.

In what follows, we have removed this ambiguity by introducing a new predicate (TU) into the language of *Knights and Knaves.* TU is used as a 2-place predicate expression with its first argument being a person and its second an utterance. Intuitively, TU(p,t) is true if and only if t is true when uttered by person p. More precisely, we say TU(p,t) is true if and only if the substitution-instance of t resulting from replacing all occurrences of " I " in t by "p" is true. Thus the statement:

TU(John,'i am six feet tall")

is true if and only if John (the utterer) is indeed six feet tall.-"

V. NOTATION AND AXIOMS

We now introduce our notation for representing the problem. We use a first-order theory which contains the following:

I: constant (the word "I")
knave: function letter (knave(x) stands for the term "x is a knave")
rich: function letter

'Including those who work in natural language processing; for Tor instance [Allen 1984], (Allen and Perrault 1980], (Harper and Charniak 1988].

"Hence, this statement can be uttered by neither knights ner knaTes, in the *Knights sni Kneves* problem!

***Throughout the remainder of this paper we use "anriki" as a short-hand for: and(not(rich(I)),knave(I)).

****This is somewhat comparable to the formulation of Barwise and Perry [1983] when they speak of an utterance in a "situation" concerning "I": u[I am six feet tall]s is true (where u is the utterance "I am six feet tall" and s is a situation in which John is present and makes utterance u) iff John is indeed six feet tall in situation s.

knight: function letter
not: function letter
and: 2-place function letter
CanSay: 2-place predicate letter (CanSay(x,y) means "x can say y")
TU: 2-place predicate expression (TU(p,t) means term t would be true if occurrences of "I" in t are replaced by p)
T: predicate expression (T(t) means the term t is true)

Given the above notation, we can now present the axioms which will capture the *Knights and Knaves* problem as we see it. For simplicity, we suppose all variables range over knights, knaves, the princess, and utterances.*

We require only three first-order axioms and one functional schema in order to sufficiently represent the needed facts about the world in which the knights, knaves, and princess live, namely,

(1) $(\forall p)(\forall u)\{T(knave(p)) \leftrightarrow [CanSay(p,u) \leftrightarrow \neg TU(p,u)]\}$
I.e., u is a knave iff the things t that u can say are precisely those which would be false if u uttered them.

(2) $(\forall y)(\forall z)[T(and(y,z)) \leftrightarrow T(y)\&T(z)]$
This captures the meaning of the function letter 'and'.

(3) $(\forall s)[T(s) \leftrightarrow \neg T(not(s))]$
This axiom captures the meaning of the function letter 'not'.

(4) $(\forall p)[TU(p,f(I)) \leftrightarrow T(f(I))]$
For example, this intuitively corresponds to TU(Bill,rich(I)) ↔ Rich(Bill), where *f* is "rich".**

If, in line with our earlier discussion, we take G to be our goal statement, then in using the above axioms we have been able to use resolution to give us the desired solution.*** A version of our resolution proof can be found in the longer paper [Miller and Perlis 1987].

*This follows the convention of Ohlbach. The use of either typed or relativized variables would eliminate unusual readings at the expense of more complex formulas.

**Note that this replaces Tarski's Convention T: T"α" ↔ α, in cases of α having the indexical "I".

***As discussed axiom 4 is really a functional schema. As such, our axiomatization of *Knights and Knaves* is not one that is computationally practicable without a mechanism for supplying substitution instances to the theorem-prover. An alternative representation of the problem at hand that requires no such mechanism but nonetheless follows closely our use of indexicals, is given in [Miller and Perlis 1987]. Namely, this alternative approach introduces a finite axiomatization of "TU". In that paper we recursively establish *all* possible instances of axiom 4 in terms of the leftmost function symbol occurring in TU's second argument. This in conjunction with our resolution proof [Miller and Perlis 1987] gives a computational solution to the *Knights and Knaves* problem.

VI. DISCUSSION

Ohlbach has pointed out an interesting problem for knowledge representation. We agree in principle with his conclusion that knowledge representation is hard. In fact, if someone has to invent a new trick each time they wished to represent a problem, the task would become hopeless. Furthermore, if the language used by the AI practitioner forced the need for tricks, then there would certainly be an argument for selecting another language.

We feel, however, that neither first-order logic nor automatic theorem-proving imposes any such restriction on the *Knights and Knaves* problem. The complexity that Ohlbach discovered in trying to represent this problem is due to indexicals. In fact, his second argument to "T" might be dealing with indexical-binding in some way. We have found that a proper treatment of indexical-binding makes for a natural and correct (in that a proper solution is found) representation of *Knights and Knaves.*

Furthermore, we feel that this problem is indicative of a whole class of problems that can be handled in a similar fashion, i.e., not dependent upon isolated or ad hoc tricks. In *Knights and Knaves* we defined TU in terms of the indexical " I " only. This is because " I " is the only indexical of importance in this problem. In broader contexts, however, this would be insufficient. It will be interesting to see how well generalizations of TU handle other indexicals and other problems.

REFERENCES

(1) Allen, J. [1984] Toward a general theory of action and time, *Artificial Intelligence,* Vol. 23, 123-154.

(2) Allen, J. and Perrault, R. [1980] Analyzing intentions in utterances, *Artificial Intelligence,* Vol. 15, 143-178.

(3) Barwise, J. and Perry, J. [1983] *Situations and Attitudes.* MIT Press, Cambridge, Massachusetts.

(4) Harper, M. and Charniak, E. [1986], *Proceedings, ACL Annual Meeting.*

(5) Miller, M. and Perlis, D. [1987] Proving self-utterances. Submitted for publication, *Journal of Automated Reasoning.*

(6) Ohlbach, H. J. [1985] Predicate logic hacker tricks, *J. of Automated Reasoning* Vol. 1, No. 4, 435-440.

(7) Perlis, D. [1985] Languages with self reference I, *Artificial Intelligence,* Vol. 25, 301-322.

(8) Smith, B. [1985] Varieties of self-reference. *Proceedings, Theoretical Aspects of Reasoning About Knowledge,* ed. J. Halpern.

(9) Smullyan, R. [1978] *What is the Name of this Book?.* Prentice-Hall, Englewood Cliffs, N.J.