A Psychologically Plausible Representation for Reasoning about Knowledge*

Anthony S. Maida

Program in Cognitive Science, T4
UC Berkeley
Berkeley, CA 94702

Richard B. Millward

Center for Cognitive Science,
Box 1911
Brown University
Providence, RI 02912

1. INTRODUCTION.

Designing a computer program to reason about the knowledge states of cognitive agents is a difficult matter. To ask that this reasoning be done in a human-like way is even more difficult. This paper describes a psychologically plausible representation and algorithm for representing and processing information about other cognitive agents' knowledge states.

1.1. The Fregean Approach to Reasoning about Knowledge

We are concerned with keeping track of coreferent terms in a memory which contains assertions about the knowledge states of cognitive agents. The situation can be illustrated by McCarthy's (1979) "telephone number problem" in which there are two phone numbers, Mike's phone number and Mary's phone number, which are coreferent. It follows that if a person, say Ed, dials Mike's phone number, he also dials Mary's phone number. However, if he knows Mike's phone number, it does not follow that he knows Mary's phone number. The Fregean approach (e.g., McCarthy, 1979) toward solving this problem is to claim that the phrase "Mike's phone number" means its referent, in normal contexts, but means something else (called the "sense") in contexts involving knowledge.

Creary (1979) and Barnden (1983) have shown that use of the Fregean approach requires more than just sense and reference; it requires a hierarchy of concepts. For example, in sentences (1)-(3) below, the phrase "Mike's phone number" must be represented as: 1) the number itself; 2) the concept of the number; and, 3) the concept of the concept of the number. For each embedding in a knowledge context, we must go one level deeper in the concept hierarchy.

(1) Ed dials Mike's phone number.
(2) Pat knows Mike's phone number.
(3) Tony knows that Pat knows Mike's phone number.

2. COGNITIVE SIMULATION APPROACH

We now describe another approach to this problem. Any concept that a human is capable of contemplating at the level of introspection should be representable as a node, or some equivalent kind of cognitive unit, in a simulation's memory. We will interchangeably call these

The authors acknowledge Nigel Ward for careful criticisms of an earlier draft of this paper. The first author was supported by the A.P. Sloan Foundation.

nodes "cognitive units" or "mental OBJECTs," to signify that they correlate with, or represent, manipulable objects of thought in a human mind. There should be no mental OBJECTs which represent concepts, ideas, or distinctions that humans do not use. These premises are a version of the so-called Uniqueness Principle described in Maida and Shapiro (1982). According to this principle, then, the phrase "Mike's phone number" should not be mapped into three distinct mental OBJECTs or cognitive units in sentences (1)-(3) unless there is introspective evidence that humans make this same kind of distinction. The remainder of this paper describes how to make this simple representation work.

We must augment this representation with processing that manages knowledge states. This involves two components. One is a scheme to maintain canonical names for equivalence classes of coreferent OBJECTs. The other is a scheme to maintain knowledge contexts for the mental OBJECTs. Each knowledge context will associate a mental OBJECT with a canonical name. Assertions will be stored with the canonical name selected by the current knowledge context.

3. THE KNOWLEDGE CONTEXT ALGORITHM.

The Knowledge Context Algorithm processes assertions of knowing with respect to the substitution of coreferent terms. The algorithm handles nested knowing such as the assertion: "Tony knows that Pat knows Mike's telephone number." The algorithm draws all valid inferences which follow strictly from the substitution of coreferent terms while drawing no invalid inferences (cf Maida, 1984).

3.1. Assigning Knowledge Contexts to Mental OBJECTs.

Expression (4) below makes reference to three concepts which will be represented as distinct mental OBJECTs. They are Tony, Pat, and Mike's-phone-number.

(4) (know-that Tony (know-value-of Pat Mike's-phone-number))

We must assign a knowledge context to each of these OBJECTs. An OBJECT in an assertion acquires its knowledge context from two sources. They are: 1) the OBJECT's knowledge context within the assertion (i.e., whether it is the second argument of a "know-that" or "know-value-of"), and 2) who happens to believe the assertion. Assuming the assertion resides in the top level of the system's memory, the knowledge contexts for each of the objects in the assertion appear in Table 1.

Table 1

Knowledge Contexts Assigned to the
OBJECTs appearing in Expression (4)

OBJECT	Knowledge Context
Tony	System
Pat	System-Tony
Mike's-phone-number	System-Tony-Pat

A hyphenated context such as "System-Tony" should be interpreted as the system's knowledge of Tony's knowledge.

3.2. Context Relative Equivalence Classes

Multiple cognitive units can turn out to be coreferent. Sets of OBJECTs known to be coreferent create equivalence classes. This scheme as it stands however can lead to a serious cross-referencing problem because multiple facts about a single real-world object can be sprinkled among any of the OBJECTs in the equivalence class.

We use a technique first employed by McAllester (1980) which involves assigning canonical-names (c-names) to equivalence classes. We shall augment McAllester's scheme with a manager for knowledge contexts. Henceforth, a mental OBJECT in an equivalence class will have a canonical name with respect to its knowledge context. Figure 1 depicts a situation in which the system knows the coreference of the three units in the set {the Morning Star, the Evening Star, Venus}, whereas it knows that Pat knows the coreference of only two of these, namely the units in the set {the Morning Star, the Evening Star}. Note that the unit representing the Morning Star has a canonical name which depends on the currently active knowledge context, i.e., in this case, whether it is the system's knowledge or the system's knowledge of Pat's knowledge.

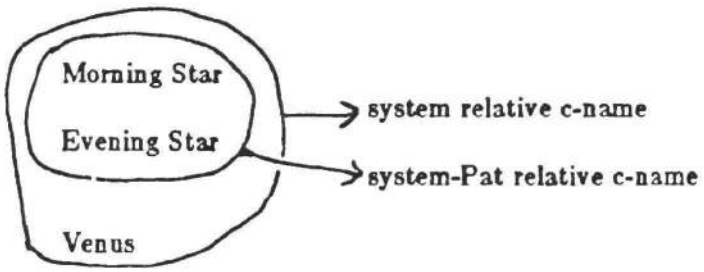

Figure 1

A OBJECT can be in Multiple Context-Relative Equivalence Classes each having their own Canonical Name.

3.3. Processing with the Knowledge Context Algorithm

Consider the units representing "Mike's phone number" in (5a) and (5b).

(5a) Pat dials Mike's phone number.
(5b) Pat knows Mike's phone number.

For a given OBJECT, information about it is stored with a canonical name that depends on the OBJECT's current knowledge context. In sentence (5a) the knowledge context for the OBJECT representing Mike's phone number is **system** and its canonical name in that context will necessarily be the same as the canonical name for the data base OBJECT representing Mary's phone number. Since extensional information pertaining to either Mike's or Mary's phone number will be stored with this canonical name, inferences which follow from substitution of coreferent terms are made implicitly by the storage scheme. However, in sentence (5b) the relevant knowledge context is **system-Pat** and the canonical name for the OBJECT representing Mike's phone number in this context will only be the same as the one for the OBJECT representing Mary's phone number if the system has an assertion residing in its data base asserting that Pat knows the phone numbers are coreferent. If the assertion resides in memory, the inference goes through, otherwise not; these are exactly the performance characteristics that we want.

3.4. Related Psychological Evidence

Anderson (1978) studied the question of what happens when two structures in human memory, previously believed to correspond to distinct real-world entities, are learned to be coreferent. When the subject learns of the coreference of distinct cognitive units, he or she gradually

migrates the factual information from one unit to the other, abandoning one unit. The unit previously used the most dominates. The subject's choice, based upon amount of previous use, is an arbitrary choice from a semantic or conceptual standpoint.

4. CONCLUSION

We presented a method of representing information about the knowledge states of other cognitive agents which is psychologically more plausible than the Fregean method. Purely Fregean methods for representing knowledge about knowledge make more distinctions than a human thinking about the same problem would make. The present method makes exactly the right number of distinctions.

References

Anderson, J.R. The processing of referring expressions within a semantic network. In TINLAP-2, 1978, 51-56.

Barnden, J.A. Intensions as such: an outline. In IJCAI-83, 1983, 280-286.

Creary, L.G. Propositinal attitudes: Fregean representation and simulative reasoning. In IJCAI-79, 1979, 176-181.

Maida, A.S. Selecting a humanly understandable representation for reasoning about knowledge. To appear, International Journal of Man Machine Studies, 1984.

Maida, A.S. & Shapiro, S.C. Intensional concepts in propositional semantic networks. Cognitive Science, 1982, 6, 291-330.

McAllester, D. The use of equality in deduction and knowledge representation. AI-TR-550, MIT Artificial Intelligence Lab, 1980.

McCarthy, J. First order theories of individual concepts and propositions. In J. Hayes & D. Michie (Eds.) Machine Intelligence 9, New York: Halsted Press, 1979.