This paper was selected by a process of
anonymous peer reviewing for presentation at

COMMONSENSE 2007

8th International Symposium on Logical Formalizations of Commonsense Reasoning

Part of the AAAI Spring Symposium Series, March 26-28 2007,
Stanford University, California

Further information, including follow-up notes for some of the
selected papers, can be found at:

www.ucl.ac.uk/commonsense07

Towards Realistic Autocognitive Inference

Fabrizio Morbini and **Lenhart Schubert**
University of Rochester, Rochester, NY 14627

Abstract

We propose methods of employing *autocognitive inference* as a realistic, feasible way for an agent to make many inferences about its own mind and about the world that have often been regarded in the past as depending on nonmonotonic reasoning. The keys to realism are **(1)** to use a computable notion of knowing; and **(2)** to employ specific, realistic assumptions about the sources of our knowledge, rather than simplified negation-as-failure assumptions. We illustrate our methods with a preliminary implementation of several reasoning examples in the EPILOG system.

Introduction

Consider questions such as the following, posed as tests of someone's commonsense knowledge. (Hypothetical answers are indicated in brackets).

1. Do pigs have wings? [*Of course not.*]
2. Do thrips have wings? [*Hmm, I don't know.*]
3. Can you find out from Wikipedia whether thrips have wings? [*I believe so.*]
4. Did the phone ring within the last 10 minutes? [*No.*]
5. If the phone rings within the next 10 minutes, will you hear it? [*Yes.*]
6. Is Bill Clinton sitting right now? [*I don't know.*] (see (McCarthy 1995))

The hypothetical answers seem reasonable for a human respondent (under plausible assumptions about the respondent's state of knowledge). A common feature in the requisite reasoning seems to be a reliance on *knowledge about one's own cognitive faculties*. By this we mean not only knowledge of what one knows and doesn't know, but more broadly what one learns and perceives under various circumstances or through various actions (e.g., consider (3-5)). For this reason we use the term *autocognitive inference*, rather than the usually more narrowly construed term *autoepistemic inference*.

Our goal is to be able to answer questions like those above easily with the aid of autocognitive inference. This goal strikes us as important not only for the purposes of the self-awareness project we are engaged in (for earlier theoretical and methodological remarks, see (Schubert 2005)), but more generally for developing more realistic versions of certain kinds of commonsense reasoning often treated within nonmonotonic reasoning (NMR) frameworks. A central claim is that realism demands a definition of "knowing" according to which any proposition that is known can be derived swiftly, rather than in the limit of an unbounded computational process. Thus we are committed to a formal computational model of belief along the lines of (Kaplan 2000; Kaplan & Schubert 2000), which overcomes the intractability and undecidability problems that beset classical NMR methods. Further, realism requires that any knowledge completeness assumptions should be explicitly stated and practically defensible, instead of being left implicit in the rules of inference or in axioms that minimize predicate extensions.

In the next section we elaborate on the perceived shortcomings in current methods for examples like (1), and on our proposed approach. This is followed by a discussion of further examples (3-6), and, in the "Examples and Results" section, by some details about examples implemented in the EPILOG system. We then state our conclusions and future agenda.

Pigs, Wings, Knowledge and Realism

In answering (1), a reasoner based on default logic (Reiter 1980) might use a rule to the effect that a creature can be assumed to lack wings whenever that assumption is consistent with the KB. This method and other NMR and autoepistemic approaches depend on verifying *nonentailment* of propositions by a KB – which even in the propositional case is in general intractable and for FOL is in general impossible. This is strikingly at odds with the fact that human introspection about what one knows (as opposed to what one can figure out with protracted thought) is virtually instantaneous.

Our approach here, as already indicated, is to borrow some essential features of the *computational model of belief* developed in (Kaplan 2000; Kaplan & Schubert 2000). This model makes a clear distinction between knowing and being able to infer. Knowing is formalized in terms of an algorithmic ASK-TELL mechanism (constrained by AGM-like metaaxioms) that always reponds quickly; whereas reasoning may go on indefinitely, as in the case of people. For example, most educated people know that Egypt is more southerly than Finland, most can figure out, but don't know, that 28 is the sum of its proper divisors, and none know, nor can figure out, whether the first player in chess can force a win, even if they know the rules of chess. The computational model in certain respects generalizes and in other respects constrains Konolige's *deduction model of belief* (Konolige 1986), and supports sound reasoning about the beliefs of other agents using *simulative inference*.

Even though our implementation of the ASK mechanism in EPILOG is still in progress (as is an overhaul of EPILOG itself), it already has many intuitively desirable properties, including timely termination and various properties that traditionally have made possible-worlds models nearly irresistible, such as precluding knowing both ϕ and its negation, order-independence of conjuncts in knowledge of conjunctions, indifference to double negation, knowledge of any disjunction if either disjunct is known, etc.

The second issue we are concerned with here is lack of realism in the knowledge closure assumptions underlying the motivating examples in much of the NMR/autoepistemic reasoning literature. For example, a closure assumption of the type, "*If any creature has wings, this follows from my knowledge*", as a means for answering (1), is thoroughly unrealistic. We think that a practically usable approach to answering questions like (1-5) requires close attention to the plausibility of the underlying knowledge and metaknowledge assumptions.

As an initial attempt to characterize the reasoning involved in (1) more realistically, we might suppose that to answer the question we just examine some internal representation or prototype of a pig, and failing to find wings, answer negatively. However, things cannot be quite so simple – for instance, we might also fail to find, say, an appendix in our representation of pig anatomy, yet might want to plead ignorance in that case. So when does absence of a part from the representation indicate its actual absence, as opposed to mere incompleteness of the representation? It seems that only "major parts" (such as plainly visible ones and major internal ones) can be assumed to be known, if they exist in reality. But as (2) illustrates, for less familiar creatures such as thrips, we may not even know all the major parts, even if we have some direct or indirect acquaintance with the creatures. So a further assumption seems to be required that the species in question should be a familiar one, if our knowledge of its major parts is to be trusted as complete, enabling the negative answer in (1). As a somewhat plausible basis for answering (1), we thus propose the following sorts of postulates (formalized later) for familiar kinds of entities, where major parts include wings, and familiar kinds include pigs:

7. If (K Q) is a familiar natural kind, and (K P) a major kind of bodypart, and kind (K Q) has-as-part (K P), then I know (that ((K Q) has-as-part (K P)).
8. (K pig) is a familiar natural kind, and (K wing) is a major kind of bodypart.

Note that not only the explicit knowledge completeness assumption, but also the assumption of familiarity with certain kinds, is an autocognitive assumption. We think that familiarity-knowledge is quite important in introspection about what we know. Unlike a tacit metabelief that we can prove the presence of wings for all creatures that have them, metaknowledge qualified by familiarity assumptions could quite plausibly be acquired from experience. For instance, we might have learned early on that certain kinds of external physical parts (such as head, body, wings, or legs in the case of animals) that are referred to in language are generally visible when present. So if we have observed some instances of a kind of creature (which is one of the ways of increasing our familiarity with it), we have surely noted what parts it possesses, among these overt sorts of parts.

Premises (7-8) indicate that our knowledge representations will have some unusual features, such as predicate reification (K P), forming the individual kind corresponding to P, proposition reification (that ((K Q) has-as-part (K P)), and application of a predicate (*has-as-part-of*) that is intended to relate physical objects to a pair of abstract individual kinds. The reification operators are part of the *episodic logic (EL)* knowledge representation (e.g., (Schubert & Hwang 2000; Hwang 1992)); the "misapplication" of the *has-as-part* predicate can be thought of as a macro (or as "type-coercion"), to be expanded by meaning postulates (as will be illustrated).

Further Aspects of Autocognitive Inference

(3-5) illustrate the important role of knowledge about how we come to know things, and how our perceptual faculties function. In (3), metaknowledge about the *kinds* of knowledge to be found in Wikipedia is crucial (cf. the autoepistemic metaknowledge in 7). In addition, autocognitive knowledge is required about the effects of looking up and reading information contained in an encyclopedic source.[1]

In example (4), as in (1), it is again tempting to simply assume that "if this proposition (that the phone rang) were true I would know it". But since we don't hear every ringing phone, we again need to ask what deeper knowledge would lead us to accept such an assumption. In general, to hear a ringing phone (or other prominent sounds) it seems sufficient to be conscious, not hearing-obstructed, and within earshot of the phone (or other source). In addition, we know that if we hear certain meaningful sounds such as human speech or ringing phones, we will remember hearing them for a matter of hours or more (at least in summary form, if densely spaced in time). So if the conditions for hearing the phone held, then lack of recollection of the phone ringing indicates that it did not ring. The knowledge used in deriving this negative conclusion is again clearly autocognitive, much as in (1).

We recognize that we run into the qualification problem at every turn here (McCarthy 1986), i.e., the success conditions we have sketched for hearing and remembering a ringing phone are not absolutely reliable. For example the phone

[1] A similar knowledge assumption was made early in the history of AI by McCarthy & Hayes in their discussion of making a phone call (McCarthy & Hayes 1969).

might have malfunctioned and rung too softly to be heard, or a memory lapse might prevent recollection of its ringing. We propose to deal with the qualification problem probabilistically, i.e., we treat conclusions such as that the phone will be heard under certain conditions as uncertain (even if highly probable), much as causal consequences in causal networks are regarded as uncertain in general. We will include a simplified version of (4) handled by EPILOG in the next section.

Example (5) is much like (4), but it illustrates that the sort of autocognitive knowledge we have already discussed, concerning conditions under which a ringing phone is heard, also permits prediction, assuming that the respondent has reason to believe that these conditions will be satisfied (perhaps because of personal commitment to being at the relevant location at the relevant time, and expectations about remaining conscious and hearing-enabled).

(6) is in a way trivial, given our notion of knowing: the positive and negative self-query both yield NO, so the answer is UNKNOWN. But insofar as McCarthy's conception of knowing is akin to inferrability, there is a deeper puzzle here: not only don't we know whether Clinton is sitting right now, we also know that we can't *infer* a definite answer, no matter how deeply we think about it. How is it possible to know this in advance, without actually putting in the reasoning effort?

We think that two sorts of autocognitive evidence may contribute to the expectation that further reasoning would be futile. First, a natural approach to assessing the difficulty of an intellectual (or other) task is to try to quickly sketch a plan for completing it, and evaluate the likelihood of success of any plan(s) found. In the present case, it seems likely that such a planning attempt would quickly reach an impasse (simply for "lack of ideas" – perhaps reflected in an empty agenda).

The success probability may be lowered further by the knowledge that certain properties vary capriciously, including ones like other people's body posture, motions, and speech; and such capriciously variable properties, if not immediately known (i.e., obtainable through the self-query mechanism), can generally be ascertained only by personal observation or "live" reports from others, not by pure thought. We think it likely that people possess large amounts of this sort of metaknowledge, facilitating the productive allocation of intellectual (and other) resources.[2]

Examples and Results

In this section we consider two of the above question-answering (QA) problems, which resemble common examples in the NMR literature. We show simplified solutions to these problems implemented in the "legacy" version of the EPILOG inference engine for EL (Schaeffer *et al.* 1993), currently being overhauled. The overhaul of EPILOG and the experimentation with various autocognitive reasoning problems are intended as steps towards creation of a system with explicit self-awareness in the sense of (Schubert 2005), which we will call EPI2ME.[3]

Pigs, Wings, and EPILOG

The first example concerns the question "*Do pigs have wings*", and some variants. The point here is to show how metaknowledge about the completeness of some limited aspect of an agent's knowledge can lead to conclusions similar to ones often obtained in NMR through negation-as-failure or similar methods.

As noted earlier, we use a kind-forming operator K in formalizing this example. A point often neglected in the literature in discussions of examples such as that birds generally fly, is that in ordinary discourse we really understand such examples in terms of kinds of entities (e.g., kinds of animals, or species of birds). Kinds and generic sentences have been much discussed in linguistic semantics (e.g., (Carlson & Pelletier 1995)), and our "language-like" EL representation readily accommodates this intuition. (The operator is intensional, but we set aside semantic details here.) We also use an operator KindOf that maps a monadic predicate P to a kind-level predicate that is true of the subkinds of the kind (K P). Predicate modification is a very useful feature of EL.

We will ask EPILOG not only whether pigs have wings, but also whether Gerda, a particular pig, has wings, thus showing that the connection between kinds and their instances can be made. We also pose the question whether Gerda has a tail, again obtaining an answer based on generic knowledge. The details of the knowledge supplied to EPILOG are as follows:

```
;; Pigs are a natural kind.
(kn '((K pig) NaturalKind))

;; The kind 'wing' is a major kind of bodypart:
(kn '((K wing) (Major (KindOf BodyPart))))

;; Epilog is familiar with the kind of animal, 'pig'
(kn '(EpilogSystem FamiliarWith (K pig)))

;; The following is the limited knowledge-completeness
;; assumption supplied to Epilog.
;;
;; If Epilog is familiar with a natural kind (y) and this
;; kind has a major kind of bodypart (x) then Epilog knows it.
;; 'KnownByMe' is the predicate triggering introspection.
;; It simply starts an embedded question/answering
;; process and returns:
;; -YES if the embedded q/a returns YES;
;; -NO if the embedded q/a returns NO or UNKNOWN.
(kn '(A y ((y NaturalKind) and (EpilogSystem FamiliarWith y))
      (A x (x (Major (KindOf BodyPart)))
        ((y HaveAsPart x) =>
         ((that (y HaveAsPart x)) KnownByMe)))))

;; Now we ask whether pigs have-as-part wings:
(dq '((K Pig) HaveAsPart (K Wing)))
```

[2]Note that we are not assuming that the possibility of inferring an answer to (6) will be ruled out in all cases. We already allowed that the answer may be known by personal observation or live report, but even if it is not, the planning process we hypothesized may yield a plan that is likely to succeed. For example, suppose that at the time the question is posed, we know from a puzzle-loving friend at the scene, keeping us informed by phone, that Clinton is sitting at that moment if and only if the square root of 123,454,321 is an integer containing a 1. Then we probably still won't *know* the answer to the question – but we'll easily construct a plan for figuring it out.

[3]pronounced *e·pit′ ŏ·mē*

The answer returned is "NO with probability 1", and the justifications given are that EPILOG has no knowledge that pigs have-as-part wings; for every major kind of bodypart that pigs have, EPILOG knows that they do; and wings are a major kind of bodypart.

An important point here is that the knowledge-completeness assumption cannot be used freely in proofs by Assumption of the Antecedent (AA), in the presence of an introspection mechanism. In particular, we could not soundly prove an instance of the conditional

((y HaveAsPart x) ⇒ ((that (y HaveAsPart x)) KnownByMe))

(having established the natural-kind, familiarity, and major bodypart portions of the antecedent) using AA, i.e., assuming the instance of (y HaveAsPart x) and then applying introspection to confirm the consequent. Introspection would trivially confirm the consequent once the antecedent has been added to the knowledge base – but this is clearly unsound.[4] Instead, *modus tollens* can be used soundly here: if the knowledge claim in the consequent of the conditional is found to be false by introspection, then we can conclude that the antecedent, (y HaveAsPart x), is false as well. This is handled uniformly by the general goal chaining (GC) rule of EPILOG.

Now, for simplicity we will take the semantics of "kind of creature y has kind of bodypart x" to be that *all* instances of the kind y have as part an instance of the kind of bodypart x. (See below, in the axioms for the question whether Gerda has a tail.) In a more careful treatment, we would at least weaken the quantifier to something like "virtually all", and derive conclusions with some non-unit probability. But interestingly, neither the strong version with *all* nor a weaker one with *virtually all* will let us conclude from the negation of "Pigs have wings" that Gerda lacks wings, because even if it is false that (virtually) all pigs have wings, it remains possible that some pigs do. One remedy would be to construe "Pigs don't have wings as something like ((K Pig) (InNoCase (HaveAsPart (K Wing)))).[5] The converse strategy is to strengthen the knowledge-completeness premise to "Whenever a familiar natural kind *in some cases* has a certain major kind of bodypart, I know it". We use the latter to answer the question whether Gerda has wings:

```
(kn '(A y ((y NaturalKind) and
           (EpilogSystem FamiliarWith y))
      (A x (x (Major (KindOf BodyPart)))
       ((y (InSomeCases (HaveAsPart x))) =>
        ((that (y (InSomeCases (HaveAsPart x))))
         KnownByMe)))))

;; If it is false that a kind (K y) in some cases
;; has a kind of part (K x), then no instances of
;; (K y) have that kind of part:
(mp '(A x_pred (A y_pred
       (((qq (not ((K y)
                  (InSomeCases
                    (HaveAsPart (K x)))))) true)
        =>
        ((qq (A z (z InstanceOf (K y))
                  (not (z HaveAsPart (K x))))) true)))))

;; Connect predicates with the kinds formed from them:
(mp '(A x_pred
      (A y_term (((qq (y x)) true) =>
                 ((qq (y InstanceOf (K x))) true)))))
;; Gerda is a pig
(kn '(Gerda Pig))

;; Now we ask, "Does Gerda have wings?"
(dq '(Gerda HaveAsPart (K Wing)))
```

The answer is again "NO", with the justification that EPILOG has no knowledge that pigs *in some cases* have wings; for every major kind of bodypart that pigs have in some cases, EPILOG knows it; wings are a major kind of bodypart; if it is false that pigs in some cases have wings then no instances of pigs have wings; and Gerda is an instance of pigs.

Finally, to answer the question whether Gerda has a tail we add

```
;; Pigs (in all cases) have tails
(kn '((K Pig) HaveAsPart (K Tail)))

;; If a kind has a kind of part, then all its instances
;; have that kind of part
(mp '(A x_pred
      (A y_pred
       (((qq ((K y) HaveAsPart (K x))) true)
        =>
        ((qq (A z (z InstanceOf (K y))
                  (z HaveAsPart (K x)))) true)))))
;; Does Gerda have a tail?
(pq '(Gerda HaveAsPart (K Tail)))
```

A "YES" answer is immediate because of forward inference performed by EPILOG. We could also get an immediate affirmative answer to `(pq '(E x (x Tail) (Gerda HaveAsPart x)))` By adding a further axiom that if an object-level entity (not a kind) has a kind of part, it has an instance of that kind of part.

Did the phone ring?

The second example shows how an agent's knowledge about how it acquires knowledge through perception can be used to answer the following question negatively: "*Did the phone ring (during some particular episode E1)?*". Note that as discussed in the previous section, it would be unjustified to answer "*no*" simply on the grounds that the agent doesn't know that it rang. The knowledge used by EPILOG is as follows:

```
(kn '(P1 Phone))
(kn '(A e2 (e2 During E1)
         ((Me Within-Earshot-of P1) @ e2)))
(kn '(A e2 (e2 During E1) ((Me Conscious) @ e2)))
(kn '(A e2 (e2 During E1)
         (((Hearing-Ability-Of Me) Normal) @ e2)))

;; Autocognitive assumption about conditions for hearing
;; a phone (approximate -- the conditional should be
;; probabilistic not universal):
(kn '(A x (x Phone)
      (A ev ((x Ring) ** ev)
       ((((Me Within-Earshot-Of x) @ ev) and
```

[4]This is related to the fact that the rule of necessitation in modal logic, $\frac{\vdash \phi}{\vdash \Box \phi}$, cannot be cast as an axiom $\phi \Rightarrow \Box\phi$.

[5]This again makes use of the predicate-modification syntax of EL.

```
        ((Me Conscious) @ ev)
        (((Hearing-Ability-Of Me) Normal) @ ev))
       => ((Me Hear ev) @ ev)))))

;; I know what I've heard (approximate -- Know should
;; really be time-dependent):
(kn '(A ev ((Me Hear ev) @ ev)
          (Me Know (That ((Me Hear ev) @ ev)))))

;; Ask whether P1 rang during E1:
(dq '(E ev (ev During E1) ((P1 Ring) ** ev)))
```

The answer is "NO with probability 1", with the justification that P1 is a telephone, and EPILOG was always within earshot of P1, conscious, and of normal hearing during E1, and whenever such conditions hold and the phone rings, EPILOG will know about it, and EPILOG doesn't know whether P1 rang during E1.

The "consciousness" and "normal hearing" assumptions could themselves be conclusions from autocognitive reasoning along the lines, "*If I had been unconscious or my hearing had been obstructed during E1, I would know it, but I don't know it, so I wasn't*". The specific knowledge completeness assumptions involved here are quite plausible, if for instance the agent registers and remembers points of transition between waking and sleeping states, and episodes of auditory (and other sensory) disruptions during waking states, such as loud masking noises, covered or injured ears, etc.

Conclusions and Further Work

We have advocated greater realism in formalizing the kinds of commonsense reasoning that rely on assumptions about how complete an agent's knowledge is in certain respects, and how it acquires knowledge through perception and other means. In particular, we have suggested that knowledge introspection should be based on a fast self-query algorithm (as in the computational model of belief (Kaplan 2000; Kaplan & Schubert 2000)), and should use explicit knowledge-completeness premises, familiarity premises, and premises about how the agent acquires knowledge. We have referred to this style of reasoning as *autocognitive* reasoning.

As preliminary evidence of the feasibility of autocognitive reasoning we mentioned some desirable properties that our initial implementation of the self-query algorithm possesses; we outlined approaches to a number of specific QA problems; and we presented some simplified working examples implemented in EPILOG. This also provided a glimpse of our ongoing effort to build an explicitly self-aware system, EPI2ME.

Our self-query algorithm (associated with KnownByMe) cannot yet call itself arbitrarily, and allowing it to do so is one of our immediate goals (in tandem with the EPILOG overhaul). Syntactic quantification and the use of quasi-quotes (for metaknowledge and meaning postulates) also needs further study and revision. Beyond such technical issues, it will also be a major goal to build up a sizable knowledge base (by hand, borrowing from various sources, and text-mining) so that commonsense and autocognitive reasoning over a reasonably broad range of topics can be attempted.

Acknowledgements

This work was supported by NSF IIS grants 0328849 and 0535105.

References

Carlson, G. N., and Pelletier, F. J. 1995. *The Generic Book*. Univ. of Chicago Press.

Hwang, C. H. 1992. *A Logical Approach to Narrative Understanding*. Ph.D. Dissertation, University of Alberta.

Kaplan, A. N., and Schubert, L. K. 2000. A computational model of belief. *Artif. Intell.* 120(1):119–160.

Kaplan, A. 2000. *A Computational Model of Belief*. Ph.D. Dissertation, University of Rochester.

Konolige, K. 1986. *A Deduction Model of Belief*. San Francisco, CA: Morgan Kaufmann.

Lakemeyer, G., and Levesque, H. 1988. A tractable knowledge representation service with full introspection. In *Proc. of the 2nd Conf. on Theoretical Aspects of Reasoning about Knowledge (TARK-88)*, 145–159.

Lakemeyer, G., and Levesque, H. J. 2005. Only-knowing: Taking it beyond autoepistemic reasoning. In Veloso, M. M., and Kambhampati, S., eds., *AAAI*, 633–638. AAAI Press AAAI Press / The MIT Press.

McCarthy, J., and Hayes, P. J. 1969. Some philosophical problems from the standpoint of artificial intelligence. In Meltzer, B., and Michie, D., eds., *Machine Intelligence volume = 4*. Edinburgh University Press. 463–502.

McCarthy, J. 1980. Circumscription—a form of non-monotonic reasoning. *Artificial Intelligence* 13:27–39.

McCarthy, J. 1986. Applications of circumscription to formalizing common-sense knowledge. *Artif. Intell.* 28(1):89–116.

McCarthy, J. 1995. Making robots conscious of their mental states. In *Machine Intelligence 15*, 3–17.

Moore, R. C. 1985. Semantical considerations on non-monotonic logic. *Artif. Intell.* 25(1):75–94.

Reiter, R. 1980. A logic for default reasoning. *Artificial Intelligence* 13:81–132.

Schaeffer, S.; Hwang, C.; de Haan, J.; and Schubert, L. 1993. EPILOG, the computational system for episodic logic: User's guide. Technical report, Dept. of Computing Science, Univ. of Alberta.

Schubert, L., and Hwang, C. 2000. Episodic logic meets little red riding hood: A comprehensive, natural representation for language understanding. In Iwanska, L., and Shapiro, S., eds., *Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language*. Menlo Park, CA: MIT/AAAI Press. 111–174.

Schubert, L. K. 2005. Some knowledge representation and reasoning requirements for self-awareness. In *Metacognition in Computation*, 106–113.