Some Knowledge Representation and Reasoning Requirements for Self-Awareness

Lenhart Schubert
Department of Computer Science
University of Rochester
Rochester, NY 14627-0226

Abstract

This paper motivates and defines a notion of *explicit self-awareness*, one that implies human-like scope of the self-model, and an explicit internal representation susceptible to general inference methods and permitting overt communication about the self. The features proposed for knowledge representation and reasoning supporting explicit self-awareness include natural language-like expressiveness, autoepistemic inference grounded in a computable notion of knowing/believing, certain metasyntactic devices, and an ability to abstract and summarize stories. A small preliminary example of self-awareness involving knowledge of knowledge categories is attached as an appendix.

1 INTRODUCTION

The current surge of interest in the AI community in "self-awareness" and "consciousness" (e.g., see McCarthy & Chaudhri 2004; Aleksander 2004; Aleksander *et al.* 2003; Holland 2003; Franklin 2003; Sanz, Sloman, & Chrisley 2003; Koch *et al.* 2001) is likely to spawn multiple, competing definitions of these terms. In the following I motivate and explain a very strong concept of self-awareness, that of *explicit* self-awareness, and distinguish this from weaker notions. In section 3 I then discuss the KR&R requirements for explicit self-awareness, and in the conclusion I call for a control structure based on a "life plan". The Appendix contains a transcript of a small implemented example of self-awareness.

2 EXPLICIT SELF-AWARENESS

I wish to focus here on the kind of self-awareness that could be described as both human-like (in operation, not necessarily design) and explicit (in several related senses).

Human-like self-awareness entails the possession, in usable form, of a well-elaborated self-model[1] that encompasses the agent's physical characteristics, autobiography, current situation, activities, abilities, goals, knowledge, intentions, etc. In addition it entails the possession of very general representational, reasoning, and goal/utility-directed planning abilities, scalable to a large KB; for if the "self" being modelled lives a cognitively impoverished life, unable to conceive of objects, events or situations in the world in human-like terms, or unequipped for human-like inferencing or planning, then the self-model too is bound to be impoverished.

Explicit self-awareness in an artificial agent, besides presupposing the above human-like competencies, also entails (i) that self-knowledge be encoded in a form that is *readily examinable and interpretable*; (ii) that it can be *overtly displayed*, by the agent itself, preferably through ordinary language (and perhaps other modalities); and (iii) that it lends itself to the *same* inferential processes as all other knowledge, i.e., it is not compartmentalized. Point (i) is a way of stating a commitment to a "representationalist" approach. This has the important advantage of being conducive to theoretical transparency and practical modifiability, and providing a clear path towards achieving points (ii) and (iii), overt display of self-knowledge and integrated reasoning about self.

Why investigate explicit self-awareness?

Building artifacts with explicit self-awareness promises to be interesting and useful in several related respects. First, it would help to **push the AI envelope** since self-awareness appears to have important "boot-strapping" potential with regard to *metacontrol*, *error recovery*, *learning of facts or skills* (based on perceived knowledge gaps and strategy/outcome analysis), and *autoepistemic reasoning*. (Concerning the last item, I will suggest that certain important instances of nonmonotonic reasoning (NMR) based on closure assumptions can be replaced by monotonic reasoning based on explicit self-knowledge.)

In addition, in practical applications such as task planning, medical advising, and tutoring, explicitly self-aware agents would be able to keep users informed about the assumptions on which they are basing particular responses; this sort of **transparency in interactions** (as distinct from mere theoretical transparency) is considered crucial for trustworthy agents (e.g., Norman 2001). A related point is that interaction with such agents, even by naive users, would potentially be very **natural and engaging**, insofar as such interactions depend on the user and system having a shared context – including an understanding of each other's capabilities and limitations, each other's beliefs, plans, intentions

[1] For an early essay on the need for a self-model, see (Minsky 1965)

and desires and the reasons behind these, and awareness of recent and ongoing events in the interaction.

Finally, explicitly self-aware artifacts would provide consciousness theorists of various philosophical persuasions with **state-of-the-art exemplars** of entities that seem operationally self-aware, and whose internal basis for this self-awareness can be scrutinized and understood.

Some related, but weaker notions

Previous workshops such as (McCarthy & Chaudhri 2004) herald the emergence of the following sorts of notions of self-awareness, which from the present perspective could be viewed as aimed towards some *aspects* of explicit self-awareness, but in no case subsuming it.

First, **self-monitoring** agents monitor, evaluate, and intervene in their internal processes in a purposive way (e.g., to recover from error or optimize performance); but in itself, self-monitoring does *not* presuppose a self-model integrated with a general reasoning system (consider operating systems, thermostats). Second, **self-explanation** entails recounting and justifying actions and inferences; though desirable, this does *not* entail an elaborate self-model or its integration into general reasoning (e.g., Winograd's SHRDLU didn't really understand its own explanations). **Global workspace systems (blackboard systems)** make a shared workspace globally available to and influential upon numerous processes; though human consciousness seems to involve such an architectural feature (Baars 1988), that feature alone does not presuppose reflective thought, or indeed any specific kinds of inferencing. Finally, **adaptive, robust, goal-directed systems** by definition have some of the "human-like" features criterial to explicit self-awareness; but they do not, in themselves, imply any self-model or significant reasoning (consider the lowly spider, or viral populations).

3 KR&R REQUIREMENTS

In an evolving paper that is thematically very relevant to our proposal, John McCarthy argues that for machines to be conscious of their mental states, they need to form internal sentences about ongoing mental and physical events and about stored sentences and clusters of sentences, as well as about abilities, methods, goals, intentions, and other conceptual entities (McCarthy 1999). This work is unique in that it concretely and formally tackles many aspects of the problem of enabling an agent to reason about its own internal states. The present proposal can be viewed as suggesting some variants and augmentations of McCarthy's proposals. I will open each of the subsections that follow with some motivating examples of reflective knowledge that a futuristic robot with explicit self-awareness might possess.

3.1 Basic Logical Framework

I am a robot. That I am a robot implies that I am not human.
I am almost human in some respects.
I can do housework and bookkeeping, and I can help a person set up and maintain schedules.
As accessories for entertaining children, I carry a fake light saber and a fake phaser.
All my programs were written by Jane D. and John S., or under their direction.
Most of my programs are probably bug-free.

Explicit representation of particular and general factual knowledge, whether of self or not, calls at least for predicates, names, connectives, quantifiers, and identity – in other words, a logical framework. A frequent choice, including in (McCarthy 1999), is FOL (or a subset), which allows for the above devices, and no more. However, it seems to me that the desideratum of communicability of self-knowledge in an easily understandable form – preferably, NL – calls for something closer to NL, matching at least the expressive devices shared by all natural languages. These include, besides the resources of FOL, generalized quantifiers (including ones like "*most*" and "*usually*""), predicate and sentence modifiers (e.g., adverbs in English, including adverbs of uncertainty such as "*probably*""), and predicate and sentence nominalization operators (e.g., the "*-ness*" ending and the complementizer "*that*" respectively in English). Moreover the logic should be intensional, since for example the meaning of a modified predicate does not depend on the extension of the predicate alone (e.g., "*fake light saber*" and "*fake phaser*" apply the same modifier to two extensionless predicates, yet may yield nonempty, distinct extensions); and the denotation of a nominalized sentence certainly does not depend on the truth of the sentence alone.

One of the goals in the development of *episodic logic* (EL) and its implementation in the EPILOG system (e.g., Hwang & Schubert 1993; Schubert & Hwang 2000) has been to allow for the above types of devices in the representation language and inference machinery. Some of these occur in the transcript in the Appendix. EPILOG's reasoning methods allow input-driven and goal-driven inference for the full EL language, and are integrated with a dozen specialist reasoners, including ones for type hierarchies, part hierarchies, times, numbers, and sets, via a uniform specialist interface.

3.2 Events and Situations

My development began in 2007. By the end of 2013, no more funds were available for my development. This situation persisted for two years, causing postponement of my first field trials.

Another important requirement in a general representational framework is some means for associating event/situation terms with event/situation-describing sentences, so that the described events and situations can be referred to, temporally modified, causally related, or otherwise qualified. The Situation Calculus (SC) (McCarthy & Hayes 1969) has in recent years been shown to be more expressive than had casually been assumed for decades, for instance with respect to concurrent, extended, and temporally qualified actions as well as causation (e.g., Schubert 1990; Pinto 1994). However, its ontology of situations (global states of affairs) makes no direct allowance for mundane events such as "the

football game", "Bob's car accident", or "the drop in stock prices", and the resultant remoteness from language and ordinary intuitions makes it unattractive for general, transparent, communicable KR&R.

Another popular approach is the use of Davidsonian event variables as "extra" arguments of event/situation predicates (e.g., Hobbs *et al.* 1993). However, I gave detailed arguments in (Schubert 2000) demonstrating the need to refer not only to situations described by atomic predications, but ones described by complex sentences such as the negatively described situation in the box above, or quantified ones like "*the situation of each superpower fearing annihilation by the other*". In that paper I developed a generalization of Davidsonian event semantics (with some resemblance to the approaches of (Reichenbach 1947) and later (Barwise & Perry 1983) as well) that allows for these complexities via two operators ('*' and '**') that connect sentences with event terms. EL and EPILOG (mentioned above) also make heavy use of those operators.

3.3 Attitudes & Autoepistemic Inference

I know that I am not human, and I know that you know it.
I remember that my first field trial went poorly.
I intend to avoid the mistakes I have made in the past.
There were no phone calls while you were away.
Yes, of course I know that cats don't have wings, though I've never considered that proposition.
No, I don't immediately know whether 76543 is a prime number. Let me think ... Yes, it is; now I know.

One obviously important issue for self-awareness is the representation of believing, intending, remembering, and other attitudes. If we have a sentence nominalization operator like English "*that*", which we can interpret as reifying the meaning of the sentence, yielding a proposition, then we can naturally treat the attitudes as predicates. But whether we regard attitudes as predicates or modal sentential operators, a crucial problem is that of determining what beliefs and other attitudes follow from given ones. The position develped in (Kaplan & Schubert 2000) is that any realistic, practically usable logic of belief must allow for the computational nature of belief "retrieval"; our claim is that belief retrieval can involve very complex computations (e.g., consider whether you currently believe that Copenhagen is north of Rome), but not *arbitrarily* complex ones (e.g., consider whether you currently believe that 4567 is a prime number). The proposed *computational model of belief* avoids both the extreme of logical omniscience (that if ϕ is believed, then all consequences of ϕ are believed) and that of vacuity (that no consequences of ϕ need be believed). Furthermore it permits sound belief ascription to other "like-minded" agents by use of simulative inference, based on rather weak axiomatic assumptions about the belief "storage" and "retrieval" mechanisms, TELL and ASK, that are part of our model.

For the type of computationally tractable ASK mechanism we assume, determining whether or not I know that ϕ (positive and negative introspection) is just a matter of running the ASK mechanism on ϕ, and concluding that I know ϕ if the answer is YES, and that I don't know ϕ if the answer is NO.

The example of knowing that there were no phone calls illustrates the potential of autoepistemic reasoning. An NMR approach might suggest the strategy of assuming the negative if no positive answer can be derived. But this is extremely hazardous, not only because it risks nontermination, but because it proceduralizes *tacit* assumptions about the completeness of the agent's knowledge. Perhaps the agent was out of earshot of the phone for a while and isn't aware that someone called, and for *that* reason can't derive a positive answer! Rather than relying on *ad hoc* rules or completions, we should base inferences on an agent's explicit beliefs or assumptions about the scope of its own knowledge and how its knowledge is acquired. In the case of the telephone calls, the robot obtains a negative answer by reasoning, "*If I am within earshot of a phone and am conscious, and the phone rings, I'll hear it. Further, I remember such conspicuous perceptual events for at least several days. I was in fact conscious and within earshot of the phone during the relevant time (I record my periods of consciousness and my locations), and so, since I don't remember a call, there was none.*" Similarly, in the case of the question whether cats have wings, autoepistemic reasoning could be used, hinging on knowledge to the effect that "*I am familiar with (the biological kind) cats, and know what all the major body parts of biological kinds familiar to me are, and do not know cats to have wings (which are major body parts)*".[2] Note that these are *monotonic* inferences, to the extent that the premises used are believed and not simply assumed. And if any of them *are* simply assumed, then though the inferences are in a sense nonmonotonic, they are at least transparent, allowing speculation about where the fault may lie if the conclusion turns out to be mistaken (i.e., truth-maintenance).

3.4 Generic Knowledge

When people talk to me, they usually have a question or request.
When I intend to do something, more often than not I do it.
When someone tells me a joke, I rarely get it. So I probably won't get the joke you are insisting on telling me.
When I meet someone I haven't met before, I usually greet them and introduce myself and then ask for their name. When they respond, I tell them that I am pleased to meet them. If they are family members or guests that I am expected to serve, I usually describe my capabilities briefly and offer my services.

[2] One can imagine imagistic methods for answering this question, in which case the completeness premise would be about the way the "cat image" encodes major body parts.

Much of our general world knowledge appears to consist of *generic* and *habitual* sentences (e.g., Carlson & Pelletier 1995), ranging from simple ones like the first two examples above to complex generic passages (Carlson & Spejewski 1997) such as the final example. Note also the inference from a generality to an uncertain particular in the third example. So an explicitly self-aware agent needs to be able to represent such knowledge.

Roger Schank and his collaborators employed *script*-like knowledge in promising ways in the 70's and 80's, but were limited, I believe, by insistence on conceptual decomposition,[3] and lack of well-founded methods for uncertain inference based on script-like and other knowledge.[4] Developments in NMR have also provided some relevant tools, with relatively well-developed foundations, but in most cases generic knowledge has been cast in the form of rules or closure operations rather than interpretable sentences, and none of the tools are subtle enough to allow for complexities like those above.

EL has some basic capabilities in this direction, viewing generic sentences as statistical generalizations with associated conditional frequency bounds. Moreover, EPILOG incorporates mechanisms for drawing probabilistically qualified conclusions from such generalizations along with other (possibly also probabilistically qualified) premises. Such bounds are useful, providing not only a convenient handle for inference control (disprefer pursuing consequences of highly improbable propositions), but also a basis for rational decision making. However, we still lack well-founded, practical methods for general uncertain inference. I believe that practical methods must exploit causal independence assumptions wherever possible, in Bayes-net-like manner, to minimize the demand for numerical data. Recent work such as (Poole 1993; Ngo & Haddawy 1996; Pfeffer 2000; Pasula & Russell 2001; Halpern 2003; Schubert 2004) can be considered steps in that direction.

3.5 Metasyntactic Devices

When something is **very** *so-and-so, then of course it is so-and-so. If something is* **almost** *so-and-so, then in fact it is* **not** *so-and-so.*
I know what Bill's phone number is. It is 123-4567.
The answer to the summation problem 123+321 is 444.
I can tell you what Bill looks like. He is about 6 feet tall and rather thin, 30 years old, tanned, with dark wavy hair, brown eyes, wearing rimless glasses, ...

Several kinds of knowledge (about self and the world) appear to require that the representation language be able to refer to its own syntactic expressions, or expressions with independent significance (such as NL terms, mathematical expressions, or computer programs). It appears that two metasyntactic devices suffice for the phenomena of interest: *substitutional quantification* and *quotation*.

Axiom Schemas

The first pair of sentences above illustrate *schematic meaning postulates*. "So-and-so" refers to an arbitrary monadic predicate, and we can formalize the sentences by using substitutional quantification over monadic predicates:

$$(\forall x)(\forall_{pred} P)\ \mathrm{very}(P)(x) \Rightarrow P(x),$$

$$(\forall x)(\forall_{pred} P)\ \mathrm{almost}(P)(x) \Rightarrow \neg P(x).$$

Here "$\forall_{pred}$" sanctions all uniform substitutions of well-formed monadic predicate expressions of the object language for the metavariable that it binds (those expressions include atomic predicates and modified ones, and perhaps lambda-abstracts).

We can see the need for quotation if we consider a generalization of the first axiom, applicable not only to "*very*" but also to other *monotone* modifiers such as "*truly*", "*intensely*", or "*delightfully*":

$$(\forall x)(\forall_{pred} P)\ (\forall_{pmod} M{:}\ \mathrm{monotone\text{-}mod}(\text{`M'}))\ M(P)(x) \Rightarrow P(x).$$

Here the quantification over M is restricted by the predication following the colon, and "*monotone-mod*" is a syntactic predicate (metapredicate) that can be evaluated procedurally whenever its argument is a metavariable-free expression. The availability of procedural evaluation can itself be made explicit through an axiom

$$(\forall_{subst} M)\ \mathrm{monotone\text{-}mod}(\text{`M'}) \Leftrightarrow \mathrm{APPLY}(\text{`monotone-mod?'}, \text{`M'}) = \text{`T'},$$

where "$\forall_{subst}$" is the most general substitutional quantifier, iterating over all symbolic expressions of a (Lisp-like) expression syntax, including well-formed expressions of the object language, but exclusive of ones containing metavariables. We take quotation to be transparent to substitution for metavariables. "APPLY" is unique in that it is the only metafunction for which the fact that it can be procedurally evaluated remains tacit, and it provides the interface between the logic and any special evaluation procedures. If a metapredicate or metafunction can be evaluated procedurally (for at least some arguments), then we state this using "APPLY". Its value, as a function in the logic, is defined to be whatever expression is returned as value when the procedure named by the first argument (in the above case, "*monotone-mod?*")

[3] We do need to canonicalize, and we do need ways to *infer* that eating x entails ingesting x through the mouth, or that dining at a restaurant (normally) entails numerous specific actions such as placing oneself at a table, but *reducing* everything to primitive representations is in my view counterproductive. Later work influenced by Schank, particularly in case-based reasoning, has generally dropped the insistence on primitive-only representations (e.g., Forbus, Gentner, & Law 1994; Cox & Ram 1999).

[4] By well-founded methods I means ones grounded in probability theory, such as Bayesian network techniques or probabilistic logics (e.g., Halpern 2003). Of course, an inference method may have much to recommend it (e.g., intuitive plausibility, or proven efficacy in cognitive modelling or practical applications) while still lacking formal foundations. But the field of KR&R as a whole has apparently become rather wary of techniques that have not (yet) been fully formalized.

is applied to the remaining arguments.[5]

Knowing a Value

Consider now the example of knowing a phone number. Assume that the fact

has-phone-number(Bill,'123-4567')

is available, from which it should follow that the possessor of that fact knows Bill's phone number. The quotation is appropriate here since a phone "number" actually provides a dialling sequence, and in some cases (like an address) may be mingled with non-numeric characters.

Now, it is tempting to represent the claim that I know Bill's phone number as

$\exists x$. *Know*(ME,*has-phone-number*(*Bill,x*)),[6]

and indeed the above fact lets us prove this formula by positive introspection and existential generalization. However, this would be a mistake, since the same proof is also supported by less "revealing" versions of the given fact, such as

has-phone-number(Bill,THE-EASIEST-PHONE-NUMBER),

where if I don't know that THE-EASIEST-PHONE-NUMBER = '123-4567', I don't necessarily know Bill's phone number. Once again, substitutional quantification provides a way out of this difficulty (though other ways are known as well):

($\exists_{subst}$ x) Know(ME,has-phone-number(Bill,'x').

It is then clear that the "revealing" fact *has-phone-number*(*Bill,*'123-4567') verifies the knowledge claim via substitution of 123-4567 for x and introspection, while the less revealing version does not.

Deriving a Value

Next consider the sample "summation problem". This may appear trivial, but the problem is not simply that of proving

($\exists$x) x = 123 + 321,

as the proof may not give the required binding (e.g., the statement is verified by 123+321 = 123+321). The point is that when we ask for the value of a sum, we are tacitly requiring the result to be expressed in a certain form, in this case using a standard decimal number representation. Again, a direct way to formalize such a *syntactic* constraint is through quotation and substitutional quantification:

($\exists_{term}$ x) decimal-number-rep('x') $\wedge$ x=(123+321).

It is easy to see that proof of this existential statement from 444 = (123+321) leads to the desired binding for x while (123+321) = (123+321) does not, assuming that we treat *decimal-number-rep*('(123+321)') as false. Note that we need to confirm *decimal-number-rep*('444') along the way, and I assume that this would be done through procedural attachment, again explicitly invoked through an attachment axiom that relates *decimal-number-rep*(...) to APPLY('*decimal-number-rep?*, ...).

[5]For nonterminating computations we can take the value to be some fixed object that is not an expression.

[6]Whether "ME" should be regarded as indexical is discussed briefly in (Kaplan & Schubert 2000); see also the perspectives in (Lespérance & Levesque 1995) and (Anderson & Perlis 2005).

It remains to say how the value 444 for (123+321) would be obtained. With our uniform attachment policy, a natural attachment axiom is

($\forall_{term}$ x: decimal-number-rep('x'))
($\forall_{term}$ y: decimal-number-rep('y'))
($\forall_{term}$ z: decimal-number-rep('z')
(x = y + z) $\Leftrightarrow$ 'x' = APPLY('add','y','z').

The solution would then easily be obtained by matching of (123+321) to $(y + z)$, and use of APPLY on '*add*' and '*decimal-number-rep?*'. I think this approach has quite general applicability to the integration of logic with computation, in a way that allows "subgoaling" via knowledge about syntactic predicates and functions that can be procedurally evaluated.

Knowledge Categorization

Turning now to the statement about Bill's appearance, the first clause seems beyond the scope of any familiar formal representation, and the remaining clauses beyond the capabilities of any inference system (except perhaps ones equipped with *ad hoc* subroutines for generating just such descriptions). The problem lies in interpreting "looking like", and connecting it with the sorts of facts listed. Here I will set aside the issue of formalizing question nominals such as "*what Bill looks like*", and simply assume that what is at issue is Bill's *appearance*. Then it becomes clear that the system itself needs to be able to determine which of its beliefs about an individual are *appearance propositions*. Quotation again proves helpful in enabling such inferences (though there are other options). An example of a desired conclusion is

appearance-wff-about(Bill,'dark(hair-of(Bill))'),

where the wording "*wff*" rather than "*proposition*" is intended to maintain neutrality about the nature of propositions. Such metapredications can be derived systematically if we assume metaknowledge like the following:

color-pred('dark')
($\forall_{pred}$ P: color-pred('P'))
($\forall_{term}$ x) appearance-wff-about(x,'P(x)')
($\forall$x)($\forall_{wff}$ w: appearance-wff-about(hair-of(x),'w'))
appearance-wff-about(x,'w').

In this way an agent's knowledge could be broadly categorized, providing a basis for question-answering when the answers required are descriptive. In fact, EPILOG has a version of such a capability, allowing use of topic hierarchies for selective access to knowledge about individuals and types of individuals, for topics and subtopics concerned with appearance, part-structure, physical, social or mental behavior, and other classes of properties. The small self-awareness demo at the end relies on this capability. However, though topical categories are computed for sentences (relative to certain of the arguments involved), this is done using topical markers on predicates and algorithms that are invisible to the reasoning mechanism.

3.6 Summarizing One's Own (Or Someone Else's) Experiences

Consider the following imaginary dialogue between our robot R and a human H:

R: *Please tell me how my programs were developed.*

H: Well, ... [imagine a detailed account here of who wrote what and when, what funds supported the work, what goals were set, what failures and successes occurred, etc.]

R: *So it was a struggle stretching over many years with many incremental advances and setbacks. I am surprised that it took several generations of graduate students, as well as multiple research associates and technicians, all working under your and Jane's leadership, to create all the modules and integrate them. I feel flattered to be the product of so much effort!*

I find it quite puzzling how people come up with summaries like that of the imaginary robot in the final turn above. The capacity it illustrates is perfectly ordinary and constantly employed, whenever people hear or read some story or gossip, assess or debate news or politics or movies, and most of all, when they chat about their own experiences and lives, at various levels of detail or generality. This seems to me a neglected area – and incidentally has little to do with the sorts of summarization techniques employed by current text summarization programs, i.e., ones that extract rather abstract from texts.

At least two sorts of things seem required for the example above. First, R must be able to abstract from the individual events of H's story, for instance inferring that certain successful tests reflected "incremental advances", while failures, unexpected departure of personnel, loss of funding, etc., comprised "setbacks" in the overall project. One could imagine relatively straightforward inference techniques generating these conclusions, event-by-event, in input-driven fashion. The second requirement is more difficult. R must be able to assess the reported events *collectively*, for instance recognizing that the continual goal-directed effort, with intermittent successes and setbacks, can be viewed as a "struggle" on the part of the protagonists; and R must be able to gather events and objects into similarity groups, in order to be able to talk about "many incremental advances and setbacks", "all the modules", or all the tasks carried out under someone's (or a team's) leadership; and R must be able to make judgements about the frequency and temporal extent of the various event types, for instance in noticing that the work "stretched over many years", or involved "generations of graduate students".

This capacity for abstraction and summarization seems part and parcel of our general capacity for discovering larger-scale properties and "regularities" in the world, both trivial and significant. Early on, we notice that squalling gets attention, some people are more often around us than others, people in the neighborhood seem to live in houses or apartments much like our own (with rooms, doors, windows, TVs, etc.); later we learn that most children go to school on weekdays till nearly adulthood or even beyond, there is quite often strife between nations, etc., etc. Not to belabor the obvious, we acquire many millions of such abstract generalizations. That this centrally important capacity also plays an important role in our assimilation of narratives should therefore not surprise us – but it should *interest* us more than it has done so far! And in particular, in the context of self-awareness, no system can be fully self-aware, in the human sense, without this capacity for abstracting from its own experiences, its own personal narrative, as stored, presumably, in its episodic memory.

4 CONCLUDING REMARKS

The notion of *explicit self-awareness* that I have discussed calls not only for a comprehensive declarative KR and deductive and metareasoning abilities, but also a general capacity for uncertain inference (exploiting presumed causal independence wherever possible) and continual goal-directed, utility-aware planning. I have said nothing substantive about the last point, but want to stress its importance. Note that the EPILOG transcript in the Appendix lacks the feel of a real conversation not only because of the logic-based I/O interface, but because the system is utterly bland and lifeless, simply reacting to the user's questions without displaying any interest or initiative of its own. Such a system is unlikely to be judged self-aware, in any human sense.

I envisage an explicitly self-aware system as being ultimately driven by a planning executive, that continually augments, modifies and partially executes a "life plan" that guides all of the system's deliberate actions, whether physical, verbal or mental. The next few steps are always directly executable (by procedure/process invocation) while those lying further in the future are coarser and generally not fully worked out or sequenced. The executive devotes part of its effort to utility estimation (locally and globally), where positive utility corresponds to knowledge gains and "vicarious satisfaction" in helping users, and ultimately the modifications of the life plan are guided by iterated attempts to modify the current plan to yield probably-higher net utility. Such a conception of a self-aware agent seems compatible with views of human consciousness like those of Baars and Franklin (Baars 1988; Baars & Franklin 2003), according to which "goal hierarchies" play a key role in the control of the agent via consciousness (mediated by the "global workspace").

Acknowledgements

Thanks to Don Perlis and Aaron Sloman for email exchanges that helped me to clarify my ideas, and to Michael Cox for detailed comments that helped to improve the paper. This work was supported in part by NSF grant IIS-0328849.

APPENDIX: Sample Session – A Preliminary Attempt

To provide a preliminary example of what we mean by explicit, overtly displayed self-awareness, Aaron Kaplan developed a simple EPILOG implementation of the sort of behavior we have in mind. The example was intended to show both the system's awareness of "what is currently going on" in the dialog, who is present (and their ontological categories), and topically directed, descriptive question answering. The example is very limited because of the system's very small KB, but it is genuine, in the sense that responses are produced purely by inference, not any kind of scripting. Inputs unfortunately still needed to be supplied in EL (for lack of a complete English/EL interface), and outputs are also in EL (supplemented in a few simple cases with automatic verbalizations generated by a rudimentary English generator). Some quick pointers concerning EL syntax: predications are in infix format, (x P y ...), e.g., (a1 action-type) means that a1 is an action type; (E x (...) (...)) is restricted existental quantification; similarly (wh x (...) (...)) is a wh-question formula; ((...) ** e) connects a sentence with the event e that it *characterizes*; (L x (...)) is a lambda abstract; (KA p) and (K p) are the nominalizations of an action predicate p and of a type predicate p respectively; and (THAT wff) is the nominalization of a sentence.

The only editing done to the transcript is the deletion of the verbose 'stream of thought' between inputs and outputs, and insertion of explanatory glosses and comments, signalled by '`;;`'.

```
EPI(9): (new-session)
       ;; The user initiates a new session

EPI(10): (query '(wh a1 (a1 action-type)
                        (E e1_ep (now during e1)
                                 ((epilog-system do a1) ** e1))))
       ;; The user asks, "What are you doing right now?", i.e.,
       ;; "What type of action are you doing such that 'now' is
       ;; during your doing it?"

Answer(s): ((KA (L VAR1 (VAR1 CONVERSE-WITH YOU))))
       ;; The system answers that it is conversing with the user
       ;; KA is an action-reifying operator and L is lambda

EPI(11): (query '(wh x (x = (system-knowledge-about me nil))))
       ;; The user asks, "What do you know about me?" (without
       ;; topical constraint)

Answer(s):
       ;; (EL representations omitted)
       ;; That I am conversing with you;
       ;; That you are conversing with me;
       ;; That you asked me what I know about you;
       ;; That you asked me what I was doing;
       ;; That you are human.

EPI(12): (query '(wh x (x = (system-knowledge-about
                                    (K train) nil))))
       ;; The user asks "What do you know about trains?"
       ;; (without topical constraint)

Answer(s): (($ 'SET (THAT ((K (PLUR TRAIN))
                           MEANS-OF-TRANSPORTATION))))
       ;; The system answers that trains are a means of trans-
       ;; portation

EPI(15): (query '(you human))
       ;; The user asks, "Are you human?"

Answer: NO
       No, EPILOG-SYSTEM is a computer-system.
       ;; The longer reiteration of the answer is produced by
       ;; the English generator, which also makes use of the
       ;; inference system's basis for the negative answer

EPI(16): (query '(you believe (that (not you human))))
       ;; The user asks, "Do you believe that you are not
       ;; human?"

Answer: YES
       Yes, EPILOG-SYSTEM is a computer-system.

EPI(17): (assertion '(me = aaron))
       ;; The user asserts "I am Aaron"
       ;; The system responds by printing inferences (omitted)

EPI(18): (query '(wh x (x = you)))
       ;; The user asks, "Who are you?"

Answer(s): (EPILOG-SYSTEM)
       ;; This is the list of answers satisfying the question,
       ;; i.e., the system answers "I am the EPILOG-system"

EPI(19): (query '(wh x (x = me)))
       ;; The user asks, "Who am I?"

Answer(s): (AARON)
       ;; The system answers, "You are Aaron"

EPI(20): (query '(wh x (you able x)))
       ;; The user asks, "What kinds of things can you do?"

Answer(s): ((KA (L VAR2 (E VAR1 (VAR1 HUMAN)
                                (VAR2 CONVERSE-WITH VAR1)))))
           ((KA
             (L VAR4
              (E VAR3 (VAR3 HUMAN)
                 (E VAR1 (VAR1 TRANSPORTATION-PROBLEM)
                         (VAR4 HELP VAR3
                                (KA (L VAR2
                                       (VAR2 SOLVE VAR1)))))))))
       ;; The system answers
       ;; "I can converse with humans"
       ;; "I can help solve transportation problems"
```

For the most part of the 'stream of thought' that we have suppressed consists of propositions posited by input-triggered generators and and inferences made from generated propositions. The input-triggered generators provide the system's "perceptions". For example, the session opener "`(new-session)`" causes the generation of propositions to the effect that a new episode wherein a human user converses with the system has just begun (at the clock time). The forward inference mechanism then uses a KB postulate stating that conversations go in both directions to infer that the system is concurrently conversing with the user. The subsequent question causes generation of a proposition that the user asked that question (at the new clock time), and so on.

The version of EPILOG we used contains enhancements due to A. Kaplan, which were presented and demonstrated at ICoS (Kaplan 1999). These enhancements provide a context mechanism that allows EPILOG to perform simulative inference about the beliefs of other agents.

In claiming to be able to help a user solve transportation problems, the system is currently mistaken. It believes this since we supplied the corresponding declarative information. We did so simply to demonstrate that the system can use such information, enabling it to answer questions about its abilities.

References

Aleksander, I.; Rasetti, M.; Bertoletti, T.; and R. Sanz (organizers). 2003. *Exystence Topical Workshop - Machine Consciousness: Complexity Aspects*.

Aleksander (organizer), I. 2004. *Workshop on Machine Models of Consciousness*.

Anderson, M. L., and Perlis, D. 2005. The roots of self-awareness. *Phenomenology and the Cognitive Sciences* 4:(forthcoming).

Baars, B., and Franklin, S. 2003. How conscious experience and working memory interact. *Trends in Cognitive Sciences* 7(4):166–172.

Baars, B. 1988. *A Cognitive Theory of Consciousness*. New York: Cambridge Univ. Press.

Barwise, J., and Perry, J., eds. 1983. *Situations and Attitudes*. Cambridge, MA: MIT Press, Bradford Books.

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

Carlson, G., and Spejewski, B. 1997. Generic passages. *Natural Language Semantics* 5(2):101–165.

Cox, M. T., and Ram, A. 1999. Introspective Multistrategy Learning: On the construction of learning strategies. *Artificial Intelligence* 112(1-2):1–55.

Forbus, K. D.; Gentner, D.; and Law, K. 1994. MAC/FAC: A model of similarity-based retrieval. *Cognitive Science* 19(2):141–205.

Franklin, S. 2003. IDA: A conscious artifact? *J. of Consciousness Studies* 10:47–66.

Halpern, J. 2003. *Reasoning About Uncertainty*. MIT Press.

Hobbs, J.; Stickel, M.; Appelt, D.; and Martin, P. 1993. Interpretation as abduction. *Artificial Intelligence* 63(12):69–142.

Holland (editor), O. 2003. *J. of Consciousness Studies, Special Issue on Machine Consciousness 10*(4-5), April-May.

Hwang, C. H., and Schubert, L. K. 1993. Episodic Logic: a situational logic for natural language processing. In Aczel, P.; Israel, D.; Katagiri, Y.; and Peters, S., eds., *Situation Theory and its Applications, vol. 3 (STA-3)*. Stanfors, CA: CSLI. 303–338.

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

Kaplan, A. N. 1999. Reason maintenance in a hybrid reasoning system. In *Proc. of the 1st Workshop in Computational Semantics (ICoS-1)*.

Koch *et al.* (organizers), C. 2001. *Workshop, Can a Machine be Conscious?* Cold Spring Harbor Laboratory, May 13-16: The Banbury Center.

Lespérance, Y., and Levesque, H. J. 1995. Indexical knowledge and robot action - a logical account. *Artificial Intelligence* 73:69–115.

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 4*. Edinburgh Univ. Press. 463–502.

McCarthy (chair), J., and Chaudhri (co-chair), V. 2004. *DARPA Workshop on Self Aware Computer Systems*.

McCarthy, J. 1999. Making robots conscious of their mental states. Available at http://www-formal.stanford.edu/jmc/.

Minsky, M. L. 1965. Matter, mind, and models. In Kalenich, W., ed., *Proc. of the Int. Federation of Information Processing Congress 65, New York*, volume 1, 45–49. Washington, D.C.: Spartan Books.

Ngo, L., and Haddawy, P. 1996. Probabilistic logic programming and Bayesian networks. In *Algorithms, Concurrency, and Knowledge (Proc. ACSC95), Lecture Notes in Computer Science 1023*, 286–300. New York: Springer-Verlag.

Norman, D. A. 2001. How might humans interact with robots? Human-robot interaction and the laws of robotology. Available at http://www.jnd.org/dn_mss/Humans_and_Robots.html.

Pasula, H., and Russell, S. 2001. Approximate inference for first-order probabilistic languages. In *Proc. of the 17th Int. Joint Conf. on Artificial Intelligence (IJCAI-01)*, 741–748.

Pfeffer, A. J. 2000. *Probabilistic Reasoning for Complex Systems*. Ph.D. Dissertation, Stanford Univ., Stanford, CA.

Pinto, J. A. 1994. *Temporal Reasoning in the Situation Calculus*. Ph.D. Dissertation, Dept. of Computer Science, University of Toronto.

Poole, D. 1993. Probabilistic Horn abduction. *Artificial Intelligence* 64(1):81–129.

Reichenbach, H. 1947. *Elements of Symbolic Logic*. New York, NY: Macmillan.

Sanz, R.; Sloman, A.; and Chrisley (organizers), R. 2003. *Models of Consciousness Workshop: The Search for a Unified Theory*. Univ. of Birmingham.

Schubert, L. K., and Hwang, C. H. 2000. Episodic Logic meets Little Red Riding Hood: A comprehensive, natural representation for language understanding. In Iwanska, L., and Shapiro, S. C., eds., *Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language*. Menlo Park, CA, and Cambridge, MA: MIT/AAAI Press. 111–174.

Schubert, L. K. 1990. Monotonic solution of the frame problem in the situation calculus: An efficient method for worlds with fully specified actions. In Kyburg, H. E.; Loui, R.; and Carlson, G. N., eds., *Knowledge Representation and Defeasible Reasoning*. Dortrecht: Kluwer. 23–67.

Schubert, L. K. 2000. The situations we talk about. In Minker, J., ed., *Logic-Based Artificial Intelligence (Kluwer Int. Series in Engineering and Computer Science, Vol. 597)*. Dortrecht: Kluwer. 407–439.

Schubert, L. K. 2004. A new characterization of probabilities in Bayesian networks. In *Uncertainty in AI: Proc. of the 20th Conf. (UAI 2004)*, 495–503.