To be published in:
Electronic News Journal on Reasoning about Actions and Change
Vol. 1: [*pages tbd*], 31.10.1997
http://www.ida.liu.se/ext/etai/rac/notes/1997/11/

Contribution to the Panel Debate on Theory Evaluation

Pat Hayes
University of West Florida

First, let me urge caution about getting too tied up in the Kuhnian vocabulary which Leora has introduced us to, for several reasons. First, Kuhn was talking about rather larger changes in scientific view than our entire field can realistically aspire to: things like the Newtonian revolution in physics. Second, Kuhn's story is easily distorted by being too quickly summarised, as Kuhn himself complained on several occasions; and third, because Kuhn himself later rejected it as overly simple and potentially misleading. The last thing we need is broad, historical discussion by amateur historians using an over-simplified theoretical vocabulary which is already out of date.

So, to turn to more practical matters:

Leora writes:

> *What makes a theory of nonmonotonic reasoning, action, and/or change a good theory? (These may be the same things that make any AI theory good.) Do we judge a theory by*
>
> - *the set of problems it can solve?*
> - *whether its ontology and axioms "make sense", i.e., are true in some sense?*
> - *it is easily accessible or "naive" as Pat Hayes would call it?*
> - *it can be integrated with existing "good theories"?*

Well surely we need to first focus on what problems we are *expecting* it to solve. Suppose someone in this field were to announce that they had the whole thing finished, all the problems solved, etc. What tests would we ask them to pass before we believed them? What do we expect these nonmonotonic logics to be able to DO, exactly?

Its not enough to just say, 'to reason properly'. We need some characterisation of what that proper reasoning is, or at least some examples of where it can be found. For 25 years we have been appealing to a vague sense of intuitive reasonableness, but this is a very weak basis to test theories on. Even linguists, whose empirical methods are treated with ridicule by experimental psychologists, have some statistical data to back up their 'seems-grammatical-to-native-speaker' criteria, but we don't have any hard data about 'common sense' at all, and the intuitions we appeal to often confuse linguistic, psychological and pragmatic issues.

One worry I have is that it seems to be impossible to test a logic or formalism as such, since the intuitiveness or otherwise of any example depends as much on the way the intuition is encoded in that logic as on the logic itself. Logics seem to require a kind of two-stage evaluation. Knowledge-hackers try to formalise an intuition using logic A and find it hard to match formal inference against intuition no matter how ingenious they are with their ontologies and axioms; so they turn to logic B, which enables them to hack the examples to fit intuition rather better. But the intuitive test is always of the axioms/ontologies, not of the logics themselves: there is always the possibility that a more ingenious hacker could have gotten things right with logic A, if she had only thought of the right ontological framework. For example, it has become almost accepted as revealed truth in this field that common sense reasoning isnt compatible with monotonic logic, because of examples such as if you are told that an automobile exists then you infer that its in working order, but if you later hear its out of gas you change your mind. (Or if you hear its only a toy bear, or a penguin, etc.) All of these examples assume that the new knowledge is simply conjoined onto the previous knowledge: you know some stuff, new stuff arrives, and you just chuck it into the set of mental clauses and go on running the mental inference engine. But maybe the updating process is more complicated than that. Maybe when you hear that the car tank is empty, you don't just add some new information, but also *remove* some previous assumptions; and maybe this is not part of the reasoning process but of the linguistic comprehension process. If so, then the representation may, perhaps, be able to use monotonic logic perfectly happily. Maybe not; but my point is only that the argument that it must be nonmonotonic makes assumptions about other mental processes - specifically, those involving the integration of new information - which have not been examined critically.

> *What gives a theory staying power? What are some examples of theories with staying power? Are these always the good ones? Specifically, are there examples of good theories which didn't last very long in the AI community? Examples of bad theories which did last long? (And who will be brave enough to identify these ;-))*

I'll take on that onerous task. At the risk of treading on almost everyone's toes, let me propose the situation calculus; or more properly the idea behind it, of describing change in terms of functions on world-states. "Bad theory" isnt really right: it was a really neat theory for a while, and better than anything going, and its still useful. But it has some pretty dreadful properties; and yet not only has it lasted a long time, but its almost considered to be inviolable by many people in the field. And even its critics - for example, Wolfgang Bibel's recent IJCAI survey gives an alternative approach based on limited-resource logics - seem to me to miss the essential things that are wrong with it.

This deserves a much longer treatment, but here are a few of the things that are wrong with sitcalc. First, its based on an overly simplistic view of the way things happen in the everyday world, one obviously inspired by reasoning about what happens inside computers. The everyday world just doesnt consist of static states and functions between them: its not organised like a series of snapshots. Sitcalc belongs with SHAKEY, in a world where only the robot can move and nothing else is happening.

Second, sitcalc only works properly if we are careful only to mention processes which can be acted upon; that is, it confuses change with action. (Consider how to describe the growth of a plant in sitcalc. It seems easy enough: something like this might be a beginning:

$$(Alive(p,s) \wedge Height(p,s) = h \wedge Watered(p,s)) \rightarrow$$

$$(Alive(p, grow(s)) \wedge Height(p, grow(s)) = h)$$

But in the sitcalc this would mean that there was an *action* called 'grow'. (All gardeners would find this action very useful, no doubt.)

Third, it confuses action with inference. The way that actions are described in the sitcalc involves asserting conditions on the past and inferring conclusions about the future: axioms have the general form ...(s) => ...(action(s)). But common-sense reasoning often involves reasoning from the present to the past (as when we infer an explanation of something we see) or more generally, can move around in time quite freely, or may have nothing particularly to do with time or action. We are able not just to say that if the trigger is pulled then the target will be dead, but also, given the corpse, that someone must have pulled the trigger. In the sitcalc this would require giving necessary and sufficient conditions for every action description, and Reiter's recent attempt to rejuvenate it does. (This conception of intuitive thought as being a progressive inferential progress in a past-to-future direction has been responsible for many other blind alleys, such as much of the work on principles for 'maintaining' truth as long as possible.)

Most intuitive reasoning done by humans lies entirely outside the purview of the situation calculus. Yet so firm has been the grip of the sitcalc ontology on people's thinking that examples which do not immediately fit into it are routinely ignored, while entire libraries are devoted to overcoming artificial problems, such as the frame problem and the YSP, which only arise in the sitcalc framework. Which brings us to the fourth thing wrong with sitcalc: it has many fatal, or at any rate very intractable, technical problems. Why is it that the only people who feel at all bothered by the frame/ramification/qualification problems are philosophers (who mostly dont even understand what they are) and people working in this rather isolated part of KR? Why hasnt the FP become a central difficulty in, say, natural language work, or qualitative physics, or planning (as used in industrial applications)? Because those fields typically dont use this clumsy ontology, that's why. These problems are all artifacts of the sitcalc; they are all concerned with how to keep track of what is true in what 'state' .

Then, Erik writes:

> *What Should Count as a Research Result in our Area?.*
>
> *I want to focus on Leora's first issue - criteria for the evaluation of theories, and I think the first thing to discuss is what could or should reasonably count as a* research result *in our area, that is, what things ought to be citable in the literature. "Citable" means that they are crisp, have a lasting value, that later researchers can build on earlier results, etc. Then, presumably, some of the respectable research results are those which tell us something about the qualities of a particular theory/approach/formalization/...*

Yes, but 'theory' is crucially ambiguous here. One of the biggest failures of the KR community generally is that it is virtually impossible to actually publish a knowledge representation itself! One can talk about formalisms and semantics and equivalences etc. etc. (the stuff in Erik's list), but this is all part of the *metatheory* of knowledge representation. But when it comes to actually getting any representing done, we hardly hear about that at all. Examples of actual formalizing are usually given as counterexamples to some conjectured technique rather than as things to be studied and compared in their own right.

There's nothing wrong with metatheory, provided there is something there for it to be the metatheory of. Right now, the chief problem with this field is that we've run out of subjectmatter. McCarthy set out in a pioneering direction, but instead of continuing his movement, we've set camp and are arguing interminably about what kind of compass to use. Let's get some actual knowledge represented, and only then study how it works and fit our theories to the things we find.

For example, here's an issue which might have some meat on it. Erik mentions Allen's time-interval algebra. Now, timepoints and intervals are a pretty simple structure, mathematically speaking, but nevertheless Allen's algebra has its problems. In particular, its not really compatible with the usual view of intervals as sets of points on a line – for details, see

```
http://www.coginst.uwf.edu/~phayes/TimeCatalog1.ps
http://www.coginst.uwf.edu/~phayes/TimeCatalog2.ps
```

I used to be convinced, all the same, that having intervals as a basic ontological category was fundamentally important, and spent a lot of time finding ways to show that certain interval theories were reducible to others and that points were definable in terms of intervals, etc.. But when I try to actually *use* these concepts to write axioms about clocks, timedurations, calendars and dates, I find that in fact the concept of interval is almost useless. One can think of an interval as just a fancy way to talk about a pair of points; and when one does so, the entire Allen apparatus just dissolves away into a simple theory of total linear order, all the axioms become simpler (for example, instead of writing 'duration (between(p,q))' one simply writes 'duration(p,q)'; there is no need to refer to the interval defined by the endpoints) and everything becomes clearer and more intuitive (for example, many quite natural relations on intervals become awkward disjunctions in the Allen framework, such as before, meet, overlap, start, equal, which is p1 => p2). So maybe there isnt much use to the concept of 'interval' at all: or, more exactly, since Allen intervals can't be thought of as sets of points but are uniquely specified by their endpoints, maybe thats really *all* they are, and the elaborate Allen theory is like the Wizard of Oz.

So, two points. First, in response again to Erik, when do we decide that something warrants the title of "theory/ approach/ formalisation.."? The sit. calc. is just a style of writing axioms, and the Allen algebra is just a complicated way to arrange order relationships. These seem to be little more than what Apple tried to sue IBM for, ie something like a 'look-and-feel'.

Second, more substantially: this is all because time is one-dimensional. I bet the story for spatial reasoning will be quite different, as there is no way there to encode the topology into an ordering. Now, what kinds of action and change make essential reference to two- or -three-dimensional things, and how can we formalise these? For example, consider the verbs 'spread', 'cover', 'surround', 'embed', 'emerge','penetrate' and similar actions that refer to a change in some spatially extended relation. Any ideas on this? Has anyone in this area even *considered* such actions/changes?