Electronic News Journal on Reasoning about Actions and Change 72
Vol. 1: *72-74*, 31.10.1997
http://www.ida.liu.se/ext/etai/rac/notes/1997/05/

Position Statement for NRAC Panel on Theory Evaluation

Leora Morgenstern
IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA

Why have a panel on theory evaluation? Nonmonotonic reasoning, action, and change have been studied by the AI community for the past 2 - 3 decades. There has been much churning out of new theories, but limited attempt at analysis of these theories or at introspection. We tend to have little perspective about our work. There's been very little discussion of what makes a theory good, what makes a theory last, how much progress we've really made, and what are good ways to encourage progress in the future. This panel is intended to jump start a discussion on these issues.

Questions and issues to be discussed are divided into 2 broad categories:

1. By which criteria do we evaluate theories?
2. Can we understand the history of research on nonmon, action, and change in broader historical terms, as suggested by Kuhn, Lakatos, and Laudan?

Criteria for evaluation of theories

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"?

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 ;-))

Understanding research in a broader, historical perspective

Thirty-five years ago, Thomas Kuhn suggested that the history of science is best understood as a cycle of periods of "normal science" followed by "revolutionary science." It works as follows: A theory is developed which solves some problems. The theory is associated with a "paradigm," which is, to quote Kuhn, "the entire constellation of beliefs, values, techniques, and so on shared by the members of a given community." As time goes on, new problems are discovered which the theory doesn't solve; the theory is modified slightly, and the process continues, until all of a sudden, it becomes apparent to some that the old paradigm just doesn't work. Then comes the "revolutionary" phase, in which a new paradigm is suggested and refined, and the "normal" phase starts again. (The classic example of this is the geocentric theory of the universe, which explained certain phenomena; as new phenomena were discovered, this theory had to be modified (epicycles and deferents), until it became clear (to Copernicus, Galileo, Kepler, etc.) that the geocentric theory just wouldn't work. The revolutionary phase supplanted the geocentric paradigm with the heliocentric paradigm, which then became normal science.)

Questions: can we understand the history of our field in this way? If so, are we in a "normal" phase or a "revolutionary" phase? Can we identify any such phases? Or are we still in one of the prehistoric phases?

Or – perhaps we are better off viewing our history from another perspective. Lakatos suggests that there's no one "normal" paradigm at any one time, but a number of competing research programmes. What unites these programmes is a core set of assumptions; however, there are different auxiliary assumptions. What research programmes can we identify? Do we subscribe to a core set of beliefs? Which programmes, to use Lakatos's terms, are progressive? Which are degenerative? Have any become degenerative and then popped back to being progressive?

Or should we subscribe to Laudan's description of "research traditions" which deny a core set of beliefs, but assert a common set of ontological assumptions and a common methodology for revising old theories and developing new ones?

Any other suggestions?

Is it worthwhile going through this exercise at all? It could be argued that the major developments of physics, astronomy, biology, occurred without much introspection at all, and this is perhaps valueless. On the other hand, we could argue that given the miserable state of research in nonmonotonic reasoning and action today, we need all the analysis and introspection we can get.

Any more ideas?

Finally, if you want to get into the swing of theory evaluation, you may want to look at: Erik's book (Features and Fluents) My article "The Problems with Solutions to the Frame Problem" available at `http://www-formal.stanford.edu/leora` (available also in the collection of papers "The Robot's Dilemma Revisited", Ablex, 1996, but the web is more accessible).

Electronic News Journal on Reasoning about Actions and Change 75
Vol. 1: *75-76*, 31.10.1997
http://www.ida.liu.se/ext/etai/rac/notes/1997/06/

Modelling Language vs. Repository of Common-Sense Facts

David Poole
University of British Columbia, Vancouver, Canada

My guess is that we are in a phase of normal science. The revolution is coming. When we have to explicitly consider uncertainty much of what we think we understand now will have to be thrown out.

In order to go about evalation, we have to make our goals clear. (If it doesn't matter where you want to get to, it doesn't matter much which way you go, to paraphrase Lewis Carrol). There are two quite different goals people have in building KR system; there is much confusion generated by not making it clear what you are doing (so much so that the researchers who take one view often don't understand what the others are doing and why). These are:

1. A knowledge representation as a modelling language. If you have a domain in your head you can use the KR to represent that domain. The builder of a KR is expected to give a user manual on how to axiomatize the domain. There are right ways of saying something and there may be wrong ways of saying it. Missing knowledge may mean something. Prolog and Bayesian networks are examples of such knowledge representations.

2. A knowledge representation as a repository of facts for commonsense reasoning. Under this scenario, you assume you are given a knowledge base and you are to make as much sense out of it as possible. It isn't OK for the designer of the KR to prescribe how a domain should be axiomatized. The KR should be able to get by with whatever knowledge it has. Much of the nonmon work assumes this (as far as I can see).

If you goal is the first, you probably want a very lean language which doesn't provide multiple ways of doing the same thing. You want to provide a recipe book about how to go about modelling a domain. It should be judged by whether someone can go from an informal problem (not a representation of a problem) to a solution efficiently. Does it provide a good way to think about the world? Can it exploit any structure of the domain for efficiency?

If your goal is the second, you probably want a rich language that lets you state as much as possible. It should be some free-form language that doesn't constrain you very much. Here we need to go from a representation of a problem into a solution. Does it provide reasonable answers? Can the user debug the knowledge base if an aswer is wrong?

I have two ways of judging a representation:

1. Can I teach it to cynical undergraduates without squirming? Can I make a case that this is the obvious answer?

2. How well does it work in practice? What is the range of practical problems for which it provides a solution?

Electronic News Journal on Reasoning about Actions and Change 77
Vol. 1: *77-78*, 31.10.1997
http://www.ida.liu.se/ext/etai/rac/notes/1997/07/

What Should Count as a Research Result in our Area?

Erik Sandewall
Linköping University, Sweden

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/...

Research results presumably come in several colors and shapes; I am thinking of categories such as the following:

- a formalism (sitcalc, Allen interval algebra, Yoav's explicit time logic; the A language, GOLOG, and so on)
- a semantics for a formalism (maybe formalisms without semantics shouldn't count, but then there may be multiple semantics for the same formalism, so I put this as a separate category)
- a nonmonotonic entailment method (using my own term), for example chronological minimization of change, chron min of ignorance, causal minimization
- a theorem (with proof) about the validity or range of applicability of an entailment method. This kind of result of course is obtained relative to a validation criterium, and for that we need ontologies (next panel)
- a disqualification of a proposed formalism/semantics combination in terms of counterexample(s), e.g. the original Hanks-McDermott paper
- equivalence results in various dimensions (reexpressibility of one formalism/semantics in another one, for example)
- a metastructure, such as a classification scheme, an impossibility result (do we have any of those?)

- a computational property of an algorithm, e.g., a complexity result

To the extent that we have this kind of solid results, we can evaluate proposed theories (= formalism + semantics + entailment method ??) with respect to their range of applicability and their computational properties.

With respect to David's distinction between knowledge representations that are modelling languages and those that are intended for repositories of common-sense facts, my heart is with the former kind. Among the above categories of results, those that concern or make use of a formal semantics probably only make sense in the context of a modelling language, since the notion of common sense is so vague and inherently difficult to capture.

Electronic News Journal on Reasoning about Actions and Change 79
Vol. 1: *79-83*, 31.10.1997
http://www.ida.liu.se/ext/etai/rac/notes/1997/11/

Comments to a Panel on Theory Evaluation

Pat Hayes
University of West Florida, FL, USA

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?

Electronic News Journal on Reasoning about Actions and Change 84
Vol. 1: *84-90*, 30.9.1997
http://www.ida.liu.se/ext/etai/actions/nj/9710-1/

Protocol of a Panel Debate About Theory Evaluation

Edited by: Erik Sandewall
Linköping University, Sweden

This on-line panel debate is a continuation of the workshop panel on Theory Evaluation that was held at the NRAC workshop during the IJCAI 1997 conferece, and chaired by Leora Morgenstern. The present session started with position statements by the three panelists, namely Leora Morgenstern, David Poole, and Erik Sandewall. After them followed the additional position statement by Pat Hayes; these four statements are on previous pages in this News Journal issue. The subsequent discussion up to the end of the month was as follows.

Some of the contributions were actually more directed to the topic of "Ontologies for actions and change", which was the subject of another NRAC panel. We therefore opened an additional on-line panel with that topic; the position statements and discussion protocol for it follow after the present protocol.

Murray Shanahan on 23.10.1997

Pat wrote:

> *Suppose someone in this field were to announce that they had the whole thing finished, all the problems solved, etc.. What tests would be ask them to pass before we believed them?*
>
> *We need some characterisation of what ... proper reasoning is, or at least some examples of where it can be found. ... 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.*

This is where building robots based on logic-based KR formalisms comes into its own. When we construct a logical representation of the effects of a robot's actions and use that theory to decide the actions the robot then actually performs, we have one clear criterion for judging the formalisation. Does the robot do what it's supposed to? There are other criteria for judging the formalisation too, of course, such as its mathematical elegance. But when our formalisations are used to build something that actually does something, we're given an acid test. Furthermore, when the "something that actually does something" is a robot, we're forced to tackle issues to do with action, space, shape, and so on, which I think are crucial to common sense.

> *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.*

Absolutely! More papers in the Naive Physics Manifesto vein, please. However, I did manage to "actually publish a knowledge representation itself" in ECAI-96, and won the best paper prize for it. The paper supplies axioms describing the relationship between a mobile robot and the world, specifically the effect of the robot's actions on the world and the impact of the world on the robot's sensors. Two papers on the same theme appear in AAAI-96 and AAAI-97. (See

`http://www.dcs.qmw.ac.uk/~mps/pubs.html`

under the Robotics heading.)

Erik Sandewall on 23.10.1997

Pat,

I am puzzled by your remarks, because while I agree with most of your points, I think they have already been answered by research especially during the last five years. With respect to your second point, concerning the situation calculus as an example of a theory with staying power but considerable weaknesses, exactly those observations have led to the work on reasoning about actions using first-order logic with explicit metric time (integers and reals, in particular). This approach was introduced in systematic fashion by Yoav Shoham. It has been continued under the banners of "features and fluents" (in my own group) and "event calculus" (Shanahan, Miller, and others). The "motivated action theory" (Morgenstern, Stein) arguable belongs to this family as well.

To check off your points, we do model the world with successive and (if applicable) continuous change within the duration of an action; we are able to reason about exogenous events, and of course we can combine prediction, postdiction, planning, and so on in the same formal system. Also, we do use pairs of numbers to characterize intervals. It is true that the classical Kowalski-Sergot paper from 1986 about the event calculus is formulated in terms of intervals and does not mention metric properties, but the more recent event-calculus literature uses timepoints and defines intervals as pairs of timepoints.

With respect to worlds where there is change that's the result of actions, see my KR 1989 paper which proposes how to embed differential equations in a nonmonotonic logic, and to generalize minimization of change to minimization of discontinuities for dealing with mode changes in a hybrid world. See also the IJCAI 1989 paper which shows how to reason about actions in the presence of such external events, under uncertainty about their exact timing. The same general approach has been pursued by Dean, Shanahan, Miller, and others, and Murray Shanahan's award paper last year shows that this is now a very productive line of research. (Situation calculus has also assimilated some of this more recently).

We can certainly discuss whether the shortcomings in the basic sitcalc can be fixed by add-ons, or whether a metric-time approach is more fruitful, and this discussion is likely to go on for a while (see also Ray Reiter's comments, next contribution). However, since we agree about the shortcomings of sitcalc, it might also be interesting to discuss why *it* has such remarkable inertia. Does the frame assumption apply to theories, and what actions affect the research community's choice of theoretical framework?

Also, with respect to your first observation:

> *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...*

this is true, of course, but the remedy exists and has been published: it is the *systematic methodology* which I introduced in (the book) "Features and Fluents". In brief, the systematic methodology program proposes to work in the following steps:

- Define an *underlying semantics* for a suitable range of problems. The definition must be strictly formal, and should as far as possible capture our intuitions wrt inertia, ramification, etc. As usual, the underlying semantics shall specify entailment, that is, what are the *intended conclusions* from given scenario descriptions.

- Define a *taxonomy of scenario descriptions* using the underlying semantics. The taxonomy identifies key properties of the scenarios, such as whether they allow for concurrent actions, nondeterministic actions, delayed causation, etc. All distinctions in the taxonomy are defined using the semantics.

- Analyse the *range of applicability* of proposed entailment methods (for example invovling chronological minimization, or occlusion, and/or filtering). This analysis shall consist of a proof that under certain conditions, a proposed entailment method obtains the same conclusions as are specified by the underlying semantics.

In this way, we don't have to validate the logics against the ill-defined notion of common sense; validation is performed and range of applicability is defined from perfectly precise concepts.

And how do these formal structures relate to *real* common sense? Well, an additional step may also be appropriate, namely, that of comparing the intended conclusions (as specified by the underlying semantics) with the conclusions that people would actually tend to make by common sense. However, that would be a task for psychologists, and not for computer scientists.

With respect to your final point:

> *... 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'.*

it seems to me that what really counts in the long run is things like proven range of applicability results, proven methods for transforming logic formalizations to effectively computable forms, etc. However, we can't avoid the fact that whoever writes a paper using formalization F is well advised to include the standard references to where the formalization F was first introduced and defended. Again, Leora's question about staying power becomes significant: if introducing a new formalism can give you a high Citation Index rating for very little work, what are the factors that dictate success and failure for formalizations? Does a formalization win because it solves problems that previously proposed formalizations didn't - or is it more like in the world of commercial software, where people tend to go for the de facto standard?

Pat Hayes on 31.10.1997

I wrote and Murray answered as follows:

> We need some characterisation of what ... proper reasoning is, or at least some examples of where it can be found. ... 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.

This is where building robots based on logic-based KR formalisms comes into its own. When we construct a logical representation of the effects of a robot's actions and use that theory to decide the actions the robot then actually performs, we have one clear criterion for judging the formalisation. Does the robot do what it's supposed to? There are other criteria for judging the formalisation too, of course, such as its mathematical elegance. But when our formalisations are used to build something that actually does something, we're given an acid test. Furthermore, when the "something that actually does something" is a robot, we're forced to tackle issues to do with action, space, shape, and so on, which I think are crucial to common sense.

I'm sympathetic to the fact that robot-testing forces one into the gritty realities of the actual world, and admire Murray's work in this direction. However, I think that to use this as a paradigm for testing formalizations gets us even deeper into the other problem I worry about, which is how to separate the formalism itself from all the rest of the machine it is embedded in. With robots there are even more things that stand between the formalization and the test: all the architectural details of the robot itself, the ways it sensors work, etc., are likely to influence the success or otherwise of the robot's performance; and perhaps a better performance can be achieved by altering these aspects rather than doing anything to the logic it uses or the ontology expressed in that logic.

The same kind of problem comes up in cognitive psychology. It is very hard to design experiments to test *any* theories of cognitive functioning in humans. Noun meanings in psycholinguistics is about as far into the mind as any empirical tests have been able to penetrate; other, non-cognitive, factors interfere so much with anything measurable that hard data is virtually unobtainable.

(On the other hand, maybe this is something to be celebrated rather than to worry about! On this view, influenced by 'situatedness', one shouldnt expect to be able to divorce an abstract level of logical representation completely separated from the computational architecture it is supposed to be implemented on. I expect this view is not acceptable to most subscribers to this newsletter, however, on general methodological grounds. :-)

Erik wrote:

Pat,

I am puzzled by your remarks, because while I agree with most of your points, I think they have already been answered by research especially during the last five years.....

Even if I were to agree, just cast my remarks entirely in the past tense and only point to the fact that sitcalc exercised a remarkably strong hold on everyone's imaginations for a very long time in spite of its shortcomings. It still provides an example for Leora's query. As you say:

> *... since we agree about the shortcomings of sitcalc, it might also be interesting to discuss why* it *has such remarkable inertia. Does the frame assumption apply to theories, and what actions affect the research community's choice of theoretical framework?*

Yes, I think that there was (and still is) a tendency for the field to go through the following loop. We start with a genuine research problem; make some initial progress by inventing a formalism; the formalism fails to fit the original goals, but itself becomes the subject of investigation, and its failings themselves the subject of research; and then this research effort detaches itself completely from the original goal and becomes an end in itself. You provide a very elegant example of this with the methodology you suggest for evaluating formalisations:

> *.... the remedy exists and has been published: it is the* systematic methodology *which I introduced in (the book) "Features and Fluents". In brief, the systematic methodology program proposes to work in the following steps:*
>
> - *Define an* underlying semantics *for a suitable range of problems. The definition must be strictly formal, and should as far as possible capture our intuitions wrt inertia, ramification, etc. ...*
> - *Define a* taxonomy of scenario descriptions *using the underlying semantics. ...*
> - *Analyse the* range of applicability *of proposed entailment methods (for example involving chronological minimization, or occlusion, and/or filtering). ...*
>
> *In this way, we don't have to validate the logics against the ill-defined notion of common sense; validation is performed and range of applicability is defined from perfectly precise concepts.*

Yes, but to what end? The things you characterise as 'ill-defined' are the very subject-matter which defines our field. There is no objective account of 'action', 'state', etc. to be found in physics, or indeed any other science; our intuitions about hese things is the only unltimate test we have for the correctness or appropriateness of our formalisms. Theres no way for us to escape from philosophical logic into the clean, pure halls of JSL. For example, your first step requires a formal semantics which captures our intuitions regarding "inertia, ramification, etc.". But these are technical terms arising within the theory whose validity we are trying to test. People don't *have* intuitions about such things: they have intuitions about space and time, tables and chairs, liquids and solids, truth and lies; about the stuff their worlds are made of. Even if people did have intuitions about inertia and ramification, those intuitions wouldnt be worth a damn, because they would be intuitions about their own reasoning, and one thing that psychology can demonstrate very clearly is that that our intuitions about ourselves are often wildly mistaken.

> *And how do these formal structures relate to* real *common sense? Well, an additional step may also be appropriate, namely, that of comparing the intended conclusions (as specified by the underlying semantics) with the conclusions that people would actually tend to make by common sense. However, that would be a task for psychologists, and not for computer scientists.*

Surely this must be done first (if we are to pretend to be still pursuing the original research goals which gave rise to this field.) Until the 'psychologists', or somebody, has told us what it is that our formalisms are supposed to be doing, speculation about their properties is just an exercise in pure mathematics.