Electronic News Journal on Reasoning about Actions and Change 37
Vol. 1: *37–44*, 31.10.1997
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 inttuitions 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.