Electronic News Journal on Reasoning about Actions and Change 131
Vol. 1: *131-136*, 30.11.1997
http://www.ida.liu.se/ext/etai/actions/nj/9711-1/

Protocol of the discussion during November, 1997 on

Methodology of Research on Reasoning about Actions and Change

Edited by: Erik Sandewall
Linköping University, Sweden

From: Michael Gelfond on 3.11.1997

I found parts of it difficult to follow since I am not sure what the participants mean by the word "theory". Are you referring to theory as organized body of knowledge about some subject matter, to theory in mathematical sense (like theory of probability), or to logical theory - collection of formulae in some language with precisely defined entailment relation? (This is of course a very incomplete list of possibilities).

It is important to be somewhat more precise here because in the AI community "theory" is sometimes identified with an "idea" and I am not sure that it is very useful to publicly judge ideas until they develop into theories. Sometimes this process takes much more than 25 years especially if the idea is prevented from its natural development by premature judgments or if development of a theory requires more than one basic idea.

From: Hector Geffner on 5.11.1997

1. I think the goal in KR/Non-Mon is *modeling*, not logic. A formalism may be interesting from a logical point of view, and yet useless as a modeling language.

   A "solution" is thus a good modeling language:

   declarative, general, *meaningful*, concise, that non-experts can understand and *use*, etc. (I agree with David's remark on teaching the stuff to "cynical" undergrads)

   The analogy to Bayesian Networks and Logic Programs that David makes is very good. We want to develop modeling languages that are like Bayesian networks, but that, on the one hand, are more *qualitative* (assumptions in place of probabilities), and on the other, more expressive (domain constraints, time, first-order extensions, etc).

2. For many years, it was believed that the problem was mathematical (which device to add to FOL to make it non-monotonic). That, however, turned out to be only part of the problem; a part that has actually been *solved*: we have a number of formal devices that yield non-mon behavior (model preference, kappa functions, fixed points,

etc.); the question is how to use them to define good modeling languages

3. The remaning problem, that we can call the *semantic* problem, involves things like the frame problem, causality, etc.

   To a large extent, I think the most basic of these problems have also been *solved*:

   Basically, thanks to Michael and Vladimir, Erik, Ray, and others we know that a rule like:

   if A, then B

   where A is a formula that refers to time i or situation s, and B is a literal that refers to the next time point of situation, is just a constraint on the possible transitions from the the states at i or s, and the following states.

   Or put in another way, *temporal rules are nothing else but a convenient way for specifying a dynamic system (or transition function)*

   Actually, for causal rules, the solution (due to Moises, Judea, and others) is very similar: *causal default rules are just a convenient way for specifying (qualitative) Bayesian Networks*

4. These solutions (that appear in different dresses) are limited (e.g., in neither case B can be an aribtrary formula) but are *meaningful*: not only the work, we can also understand why.

   We also understand now a number of things we didn't understand before.

   e.g., 1. a formula can have different "meanings" according to whether it represents a causal rule, an observation or a domain constraint.

   (this is not surprising from a Bayesian Net or Dynamic systems point of view, but is somewhat surprising from a logical point of view)

   2. reasoning forward (causally or in time) is often but not always sound and/or complete; i.e., in many cases, forward chaining and sleeping dog strategies will be ok, in other cases, they won't.

5. It's not difficult to change the basic solutions to accommodate additional features (e.g., non-deterministic transition functions, unlikely initial conditions, concurrent actions, etc.) in a principled way.

   So, I think, quite a few problems have been solved and default languages, in many cases, are ripe for use by *non* non-mon people.

6. We have to make a better job in packaging the available theory for the outside world, and in delineating the solved problems, the unsolved problems and the non-problems, for the inner community and students.

   Actually I have been doing some of this myself, giving a number of tutorials in the last couple of years at a number of places (I invite you to look at the slides in `http://wwww.ldc.usb.ve/~hector`)

From: Pat Hayes on 5.11.1997

I think this meta discussion, though at times confused (mea culpa, of course), has been useful in revealing a clear divergence between two methodologies, giving different answers to the original question about how we should evaluate work in the field. ("NRAC panel on theory evaluation", ENRAC 21.10).

One view appeals to our human intuitions, one way or another. In this it is reminiscent of linguistics, where the basic data against which a theory is tested are human judgements of grammaticality. We might call this a 'cognitive' approach to theory testing. Talk of 'common sense' is rife in this methodology. Based on the views expressed in these messages, I would place myself, Erik Sandewall, Michael Gelfond in this category. The other, exemplified by the responses of Ray Reiter, Mikhail Soutchanski and Murray Shanahan, emphasises instead the ability of the formalism to produce successful behavior in a robot; let me call this the 'behavioral' approach.

This distinction lies orthogonal to the extent to which people find formality more or less congenial. Both Ray and Erik dislike 'vague claims', and Erik's suggested methodology (Newsletter 23.10) meticulously avoids all contact with psychology, as he emphasises; yet he ultimately appeals to capturing our intuition, rather than any successful application in a robot, to tell us which kinds of model-theoretic structures are more acceptable than others. It also lies orthogonal to the extent to which people see their ultimate goal as that of creating a full-blown artificial intelligence (as both Wolfgang Bibel and Mikhail Soutchanski seem to, for example, along with our founder, John McCarthy), or might be satisfied with something less ambitious. This distinction in approaches - start with insects and work 'up', or start with human common sense and work 'down' - is also a methodological split within AI in general, and seems to be largely independent of whether one feels oneself to be really working towards a kind of ultimate HAL.

Do people find this distinction seriously incomplete or oversimplifying? (Why?) Or on the other hand if they find it useful, which side of the division they would place themselves? In a nutshell, is the immediate goal of the field to understand and accurately model human intuitions about actions, or is it to help produce artifacts which behave in useful or plausible ways? I think this is worth getting clear not to see which 'side' wins, but to acknowledge that this difference is real, and likely to produce divergent pressures on research.

From: Erik Sandewall on 6.11.1997

Pat,

I agree with you that it's time to sort out the different perspectives, goals, and methods for reaching the goals that have confronted each other here. You write of two dimensions; in the first one you make the following distinction:

> *One view appeals to our human intuitions, one way or another. In this it is reminiscent of linguistics, where the basic data against which a theory is tested are human judgements of grammaticality. We might call this a 'cognitive' approach to theory testing. Talk of 'common sense' is rife in this methodology. Based on the views expressed in these messages, I would place myself, Erik Sandewall, Michael Gelfond in this category. The other, exemplified by the responses of Ray Reiter, Mikhail Soutchanski and Murray Shanahan, emphasises instead the ability of the formalism to produce successful behavior in a robot; let me call this the 'behavioral' approach.*

I agree with this, except that the term 'behavioral' is maybe not the best one, and also you put me in the wrong category; more about that later. Anyway, the distinction you make here seems to coincide with the one that David Poole made in his position statement:

> *There are two quite different goals people have in building KR system; — 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. —*
>
> *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. —*

If you are going to *design* a robot in a good engineering sense, you are going to need to model both the robot itself and its environment. That's why what you call the 'behavioral' approach coincides with the use of KR for modelling physical systems. Since 'modelling' can mean many things, I'll further qualify it with the term 'design goal'.

As for the other dimension, you propose

> *— the extent to which people find formality more or less congenial. Both Ray and Erik dislike 'vague claims' —*

This distinction I find less informative, since all the work in this area is formal in one way or another. Even the kludgiest of programs exhibits 'formality'. However, different researchers do take different stands wrt how we choose and motivate our theories. One approach is what you described in your first response to the panel (ENRAC Newsletter on 22.10):

> *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.*

The key word here is *examples*. In this example-based methodology, proposed logics are treated like hypotheses in a pure empirical paradigm: they are accepted until a counterexample is found; then one has to find another logic that deals correctly at least with that example. Ernie Davis characterized this approach in his book, Representation of Commonsense Knowledge [mb-Davis-90]. See also the discussion of this approach in my book, Features and Fluents [mb-Sandewall-94], p. 63).

The example-based methodology has several problems:

- It does not prove anything, just like you can not prove the correctness of a program by test examples, and particularly not by trying it on a small number of toy tests.
- The process does not seem to converge; it does not even allow us to believe in any one hypothesis for a very long time. Here's why the good old positivist methodology doesn't work in this case: the empirical data are notoriously unreliable. We may be reasonably clear

> about what are the commonsense conclusions for a simple scenario that is presented to us, but when more complex and sophisticated scenarios are admitted, we get more and more cases where common sense is not held in common. Therefore, people will *always* come up with proposed new counterexamples to any given theory.

The choice of methodology is indeed orthogonal to your first distinction, since the example-based methodology can be used both in the pursuit of theories of common sense, and in the development of intelligent robots by design iteration (try a design, see how it works, revise the design).

The alternative to this is to use a systematic methodology where, instead of searching for the "right" theory of actions and change, we identify a few plausible theories and investigate their properties. For this, we need to use an underlying semantics and a taxonomy of scenario descriptions; we can then proceed to analyse the *range of applicability* of proposed theories (entailment methods).

Your answer to this was (31.10):

> *Yes, but to what end? The things you characterize 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 in any other science; intuitions about these things is the only ultimate test we have for the correctness or appropriateness of our formalisms.—*

This would be true if the 'cognitive' (in your terms) goal were the only one. From the point of view of modelling and design, on the other hand, these are perfectly valid concepts. The concept of state is used extensively in control engineering (yes, control theory does deal with discrete states, not only with differential equations!), and I am sure our colleagues in that area would be most surprised to hear that our intuitions is "the only ultimate test we have" for the correctness or appropriateness of the formalisms that they share with us.

Now, when you placed me in the cognitive category, you got me wrong. As I wrote in my position statement for this panel, my heart is with the use of knowledge representations as modelling languages. The present major project in our group is concerned with intelligent UAV:s (unmanned aircraft), and in this enterprise we need a lot of modelling for design purposes; we have currently no plans to pursue the 'cognitive' goal.

However, just as the example-driven methodology can serve both the cognitive goal and the design goal, I do believe that the systematic methodology can *also* be relevant as one part of a strategy to achieve the 'cognitive' goal. More precisely, for the reasons that both you and I have expressed, it's not easy to find any credible methodology for research on understanding the principles of commonsense, and in fact I did not see any concrete proposal *for* such a methodology in your contributions. However, to the extent that people continue to pursue that goal, my suggestion was to divide the problem into two parts: one where our discipline can say something substantial, and one which is clearly in the domain of the psychologists.

Therefore, the contradiction that you believed having seen when writing

> *... and Erik's suggested methodology (Newsletter 23.10) meticulously avoids all contact with psychology, as he emphasises; yet he ultimately*

> *appeals to capturing our intuition, rather than any successful application in a robot, to tell us which kinds of model-theoretic structures are more acceptable than others.*

is not a real one; it only arises because your perception that

> *... this distinction in approaches - start with insects and work 'up', or start with human common sense and work 'down' - is also a methodological split within AI in general, and seems to be largely independent of whether one feels oneself to be really working towards a kind of ultimate HAL.*

which I also do not share. After all, the behavioral/ commonsense view and the modelling/ design view represent *goals*, not methodologies, and both choices of methodology (the example-based and the systematic one) can be applied towards both the goals.

References:

[mb-Davis-90] Ernest Davis. *Representation of Commonsense Knowledge.* Morgan Kaufmann Publishers, Inc., 1990.

[mb-Sandewall-94] Erik Sandewall. *Features and Fluents. The Representation of Knowledge about Dynamical Systems.* Oxford University Press, 1994.