ELECTRONIC NEWSLETTER ON
REASONING ABOUT ACTIONS AND CHANGE

Issue 98006 Editor: Erik Sandewall 22.1.1998

Today

The renewed debate about methodology (previously also called theory evaluation) continues with a contribution by Murray Shanahan, and an answer by Pat Hayes both to Murray and to Hector Geffner's contribution yesterday.

Debates

Research methodology

From: Murray Shanahan

Pat Hayes wrote,

> *I've never been very impressed by the famous Yale shooting problem, simply because it doesn't seem to me to be a problem.*

First I think we should distinguish between the Yale shooting scenario and the Hanks-McDermott problem. The Yale shooting is scenario is the one in which someone loads, waits and shoots, and in which those actions have certain prescribed effects. The Hanks-McDermott problem is a difficulty that arises when we take certain approaches to solving the frame problem, and is exemplified in the Yale shooting scenario. The frame problem is the problem of describing the effects of actions in logic without recourse to an excess of axioms describing their non-effects.

If you want to solve the frame problem, your solution had better be able to deal with the Yale shooting scenario. A number of early attempts at the frame problem couldn't, which is why the scenario is of interest.

Isn't that all pretty straightforward?

Murray

From: Pat Hayes

Hector Geffner:

> *... I think I understand the YSP. Here is the way I see it.*
>
> *In system/control theory there is a principle normally called the "causality principle" that basically says that "actions cannot affect the past". If a model of a dynamic system does not comply with this principle, it's considered "faulty".*
>
> *In any AI the same principle makes perfect sense when actions are* exogenous*; such actions, I think, we can agree, should never affect your beliefs about the past (indeed, as long as you cannot predict exogenous actions from your past beliefs, you shouldn't change your past beliefs when such actions occur).*

But actions - or external events - do change ones beliefs about the past. They do not change the past itself, of course: that is the causality principle. But consider for example coming into a room in an empty house and finding a hot cup of coffee resting on a table. One immediately infers that somone else has been present there recently. We constantly make inferences about the past on the basis of present knowledge, even in such technical areas as military planning.

I think there is a confusion here between (1) drawing conclusions about an external world and (2) simulating an external world by inferring a state from its previous state. The causality principle applies to the latter, but not the former; and even then, it is reasonable only when accompanied by a presumption of a certain kind of completeness in ones knowledge of the state. We often make predictions of the future by a kind of mental 'simulation' by inferring what is going to happen next from what is true now (as in the conventional situation calculus axiomatic approach); but in practice, such simulations are often unreliable precisely because we don't have sufficiently complete knowledge; and when this is so, we cannot cleave to the strict causality principle, but are obliged to use techniques such as nonmonotonic reasoning which allow us to recover gracefully from observed facts which contradict our predictions, which would otherwise enmesh us in contradictory beliefs. Nonmonotonicity is a good example of the need to revise ones beliefs about the past in the light of unexpected outcomes in the present, in fact, which gets us back to the YSP:

> *What Hanks and McDermott show is that* certain models of action in AI (like simple minimization of abnormality) violate the causality principle. *In particular they show that*
>
> *your beliefs at time 2, say, after LOAD AND WAIT (where you believe the gun is loaded)*

But why should you believe the gun is loaded at this time? Why is this considered so obvious? Remember, all the axioms say about WAIT is ...well, nothing at all. That's the point of the example: if you say nothing about an action, the logic is supposed to assume that nothing much happened. But if what we are talking about is a *description* of the world, saying nothing doesn't assert blankness: it just fails to give any information. If one has no substantial information about this action, the right conclusion should be that anything could happen. Maybe WAIT is one of those actions that routinely unloads guns, for all I know about it from an axiomatic description that fails to say anything about it. So the 'problem' interpretation about which all the fuss is made seems to me to be a perfectly reasonable one. If I see a gun loaded, then taken behind a curtain for a while, and then the trigger pulled and nothing happened, I would conclude that the gun had been unloaded behind the curtain. So would you, I suspect. If I am told that a gun is loaded, then something unspecified happens to it, I would be suspicious that maybe the 'something' had interfered with the gun; at the very least, that seems to be a possibility one should consider. This is a more accurate intuitive rendering of the YSS axioms than talking about 'waiting'.

We all know that waiting definitely does not alter loadedness, as a matter of fact: but this isn't dependent on some kind of universal background default 'normality' assumption, but follows from what we know about what 'waiting' means. It is about as secure a piece of positive commonsense knowledge as one could wish to find. Just imagine it: there's the gun, sitting on the table, in full view, and you can *see* that *nothing* happens to it. Of course it's still loaded. How could the bullet have gotten out all by itself? But this follows from knowledge that we have about the way things work - that solid objects can't just evaporate or pass through solid boundaries, that things don't move or change their physical constitution unless acted on somehow, that guns are made of metal, and so on. And the firmness of our intuition about the gun still being loaded depends on that knowledge. (To see why, imagine the gun is a cup and the loading is filling it with solid carbon dioxide, or that the gun is made of paper and the bullet is made of ice, and ask what the effects would be of 'waiting'.) So if we want to appeal to those intuitions, we ought to be prepared to try to represent that knowledge and use it in our reasoners, instead of looking for simplistic 'principles' of minimising changes or temporal extension, etc., which will magically solve our problems for us without needing to get down to the actual facts of the matter. (Part of my frustration with the sitcalc is that it seems to provide no way to express or use such knowledge.)

I know how the usual story goes, as Murray Shanahan deftly outlines it. Theres a proposed solution to the frame problem - minimising abnormality - which has the nice sideeffect that when you say nothing about an action, the default conclusion is that nothing happened. The Yale- shooting- scenario- Hanks- McDermott problem is that this gives this 'unintuitive' consequence, when we insert gratuitous 'waitings', that these blank actions might be the abnormal ones. My point is that this is not a problem: this is exactly what one would expect such a logic to say, given the semantic insights which motivated it in the first place; and moreover, it is a perfectly reasonable conclusion, one which a human thinker might also come up with, given that amount of information.

Murray says:

> *If you want to solve the frame problem, your solution had better be able to deal with the Yale shooting scenario.*

This is crucially ambiguous. The conclusion I drew from this example when it first appeared was that it showed very vividly that this style of axiomatisation simply couldnt be made to work properly. So if "the Yale shooting scenario" refers to some typical set of axioms, I disagree. If it refers to something involving guns, bullets and time, then I agree, but think that a lot more needs to be said about solidity, containment, velocity, impact, etc., before one can even begin to ask whether a formalisation is adequate to describing this business of slow murder at Yale. Certainly your solution had better be able to describe what it means to just wait, doing nothing, for a while, and maybe (at least here in the USA) it had better be capable of describing guns and the effects of loading and shooting them. But thats not the same as saying that it has to be able to deal with the way this is conventionally axiomatised in the situation calculus.

Imagine a gun which requires a wick to be freshly soaked in acetone, so that just waiting too long can cause it to become unready to shoot. This satisfies the usual YSS axioms perfectly: when loaded, it is (normally) ready to fire, when fired, it (normally) kills, etc.. But if you wait a while, this gun (normally) unloads itself. Now, what is missing from the usual axiomatisation which would rule out such a gun? Notice, one doesnt want to make such a device *logically* impossible, since it obviously could be constructed, and indeed some mechanisms are time-critical in this way (hand grenades, for example). So one wants to be able to write an axiom which would say that the gun in question isn't time-critical: it has what one might call non-evaporative loading. Maybe its something to do with the fact that the bullets are securely located inside the gun, and that they don't change their state until fired...or whatever. My point is only that there is no way to avoid getting into this level of detail; formalisations which try to get intuitive results with very sketchy information cannot hope to succeed except in domains which are severely restricted.

(Response to Erik Sandewall in later message.)

Pat Hayes

This newsletter is issued whenever there is new news, and is sent by automatic E-mail and without charge to a list of subscribers, in plaintext format. The present postscript issue is a bona fide rendering of the same information in alternative typography. Instruction for contributors and other additional information is found at the following URL:

`http://www.ida.liu.se/ext/etai/actions/njl/`