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/1997/05/

NRAC Panel on Theory Evaluation

Issues and Questions

Leora Morgenstern
IBM Thomas J. Watson Research Center

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).