Multiple Mental Attitudes in Agents

Yoav Shoham
Stanford University

The TARK series of symposia has centered around the notions of knowledge and belief. The purpose of this panel is to ask whether additional categories should be brought into the picture, and if so how.

For many years AI has considered theories of mental state, whether that of humans or that attributed to machines. Motivation for doing so includes user modeling for intelligent human-computer interaction ("the user is querying about the next train to Boston; that means he has a goal of getting to Boston; he must not know that train service is down for the next 24 hours; I'd better let him know that"), planning ("I'll commit to doing this in five minutes, but not to doing the other job so as not to constrain my future plans") and especially multi-agent planning ("in order for him to be able to do this he must have the following information, so I'll find it out and inform him"), and more. As a result there's been much work on capturing various mental attitudes in formal systems. The work includes:

1. "Enabling theories" in the form of logics of time and action (much in line in theories within philosophy; that's true for the following as well).
2. "Enabling theories" in the form of nonmonotonic logics (those logics allowing tentative, default conclusions); the connection is explained below.
3. Logics for knowledge and belief (overshadowed by the later developments in distributed computation, as manifested in TARK), commitment, choice, obligation, decision, and to lesser success desire, goals, intention, and capability (the latter not really a mental quality).

(The connection to nonmonotonic reasoning is that beliefs, commmitments, and other qualities tend to persist over time, but not absolutely. For example, if I commit to doing something I must remain committed henceforth, but only if I don't discover that object of commitment is impossible. Similarly for belief: If I believe that it will rain tomorrow, and receive no new information, I should still believe it in five minutes. In either case the issue that comes up is that it is impractical to explicitly list all the conditions that might truncate the persistence at each point, and the alternative is to infer the persistence in the absence of information to the contrary. There are many subtle questions that come up in getting this to work right.)

In this panel I will set the stage by surveying (3) above, with detours into (1) and (2). I will then ask the panelists, experts in Philosophy, Linguistics, and Economics, to provide their perspectives on the role of mental state and its definition.