fourteen declarative principles of experience-oriented intelligence

Rich Sutton
April 24, 2008

1. all goals and purposes can be well thought of as the maximization of the expected value of the cumulative sum of a single externally received number (reward). "the reward hypothesis" thus life is a sequential decision-making problem, also known as a Markov decision process. "learning is adaptive optimal control"

2. a major thing that the mind does is learn a state representation and a process for updating it on a moment-by-moment basis. the input to the update process is the current sensation, action, and state (representation). "state is constructed"

3. all action is taken at the shortest possible time scale, by a reactive, moment-by-moment policy function mapping from state to action. anything higher or at longer time scales is for thinking about action, not for taking it. "all behavior is reactive"

4. all efficient methods for solving sequential decision-making problems compute, as an intermediate step, an estimate for each state of the long-term cumulative reward that follows that state (a value function). subgoals are high-value states. "values are more important than rewards"

5. a major thing that the mind does is learn a predictive model of the world's dynamics at multiple time scales. this model is used to anticipate the outcome (consequences) of different ways of behavior, and then learn from them as if they had actually happened (planning).

6. learning and planning are fundamentally the same process, operating in the one case on real experience, and in the other on simulated experience from a predictive model of the world. "thought is learning from imagined experience"

7. all world knowledge can be well thought of as predictions of experience. "knowledge is prediction" in particular, all knowledge can be thought of as predictions of the outcomes of temporally extended ways of behaving, that is, policies with termination conditions, also known as "options." these outcomes can be abstract state representations if those in turn are predictions of experience.

8. state representations, like all knowledge, should be tied to experience as much as possible. thus, the bayesian and POMDP conceptions of state estimation are mistaken.

9. temporal-difference learning is not just for rewards, but for learning about everything, for all world knowledge. any moment-by-moment signal (e.g., a sensation or a state variable) can substitute for the reward in a temporal-difference error. "TD learning is not just for rewards"

10. learning is continual, with the same processes operating at every moment, with only the content changing at different times and different levels of abstraction. "the one learning algorithm"

11. evidence adds and subtracts to get an overall prediction or action tendency. thus policy and prediction functions can be primarily linear in the state representation, with learning restricted to the linear parameters. this is possible because the state representation contains many state variables other than predictions and that are linearly independent of each other. these include immediate non-linear functions of the other state variables as well as variables with their own dynamics (e.g., to create internal "micro-stimuli").

12. a major thing that the mind does is to sculpt and manage its state representation. it discovers a) options and option models that induce useful abstract state variables and predictive world models, and b) useful non-linear, non-predictive state variables. it continually assesses all state variables for utility, relevance, and the extent to which they generalize. researching the process of discovery is difficult outside of the context of a complete agent.

13. learning itself is intrinsically rewarding. the tradeoff between exploration and exploitation always comes down to "learning feels good."

14. options are not data structures, and are not executed. they may exist only as abstractions.

some of these principles are stated in radical, absolutist, and reductionist terms. this is as it should be. in some cases, softer versions of the principles (for example, removing the word "all") are still interesting. moreover, the words "is" and "are" in the principles are a shorthand and simplification. they should be interpreted in the sense of Marr's "levels of explanation of a complex information-processing system." that is, "is" can be read as "is well thought of as" or "insight can be gained by thinking of it as."

——

a complete agent can be obtained from just two processes:

- a moment-by-moment state-update process, and
- a moment-by-moment action selection policy.

everything else has an effect only by changing these two. a lot can be done purely by learning processes (operating uniformly as in principle 10), before introducing planning. this can be done in the following stages:

(a) a policy and value function can be learned by conventional model-free reinforcement learning using the current state variables

(b) state variables with a predictive interpretation can learn to become more accurate predictors

(c) discovery processes can operate to find more useful predictive and non-predictive state variables

(d) prediction of outcomes, together with fast learning, can produce a simple form of foresight and behavior controlled by anticipated consequences

much of the learning above constitutes learning a predictive world model, but it is not yet planning. planning requires learning from anticipated experience at states other than the current one. the agent must disassociate himself from the current state and imagine absent others.