Reinforcement Learning as a Context for Integrating AI Research

Bill Hibbard

University of Wisconsin - Madison
1225 West Dayton Street
Madison, WI 53706
billh@ssec.wisc.edu

Interacting Learning Processes

As Baum argues, reinforcement learning is essential to intelligence (Baum 2004). It enabled humans who evolved in the tropics to satisfy their needs for food and warmth in the arctic. Well known reinforcement learning algorithms have been identified in the neural behaviors of mammal brains (Brown, Bullock and Grossberg 1999; Seymour et al. 2004). A brain senses and acts in the world, and learns behaviors reinforced by values for distinguishing good and bad outcomes. The brain learns a simulation model for tracing cause and effect relations between behaviors and outcomes – that is, for solving the credit assignment problem (Sutton and Barto 1998). Reason and high level representations of sense information are part of this simulation model, and language is a representation of the model for exchange with other brains (language may also serve internally to increase the efficiency of the brain's simulation model). Thus the simulation model and its role in reinforcement learning provide a context for integrating different AI subfields.

Brains can learn simulation models as internal behaviors that predict sense information, reinforced by predictive accuracy. For example, learning to predict short-term changes to visual information based on body motion may be the basis for learning 3-D allocentric representations of vision. Learning to predict longer-term changes to visual information, sometimes in response to the brain's motor behaviors, may be the basis for learning to partition the visual field into objects, learning to classify those objects, and learning to model behaviors of object classes.

Thus processes reinforced by predictive accuracy may learn a simulation model, useful to other processes that learn to satisfy basic physical needs. This suggests a brain design partitioned into interacting learning processes, each defined by a set of inputs (some sensory, some from other brain processes), an internal representation, a set of outputs (some motor, some to other brain processes), and a reinforcement value. This is similar to Minsky's notion of a brain comprising a society of agents, each implementing a different way to think (Minsky, Singh and Sloman 2004). Brain processes may interact in a variety of ways. Processes reinforced by short-term predictive accuracy may produce useful input to processes making longer-term predictions. Predictive processes may help trace cause and effect relations between behaviors and rewards in other processes. Representations learned by one process may play a role in the reinforcement values of other processes.

As Baum observes, the immense evolutionary computation that learned the design for human brains provided those brains with prior biases (i.e., Bayesian prior distributions) about the nature of the world that increase the efficiency of their own learning (Baum 2004). In an artificial brain design, such prior biases on the designs of brain processes would come from the knowledge and techniques of different AI subfields. For example, one process may take raw vision and body motor controls as inputs, produce a 3-D allocentric representation as output, and reinforce based on the accuracy of predicting raw visual input using the 3-D allocentric representation and body motion. Vision research can bias or constrain the mapping from raw vision to 3-D allocentric representation in order to increase learning efficiency, and probably also dictate the need for inputs from other vision processes (neuroscience suggests complex connections among vision processes). Similarly, processes that learn language may be biased or constrained by a prior universal grammar, hypothesized by some linguists as necessary for children to learn language as quickly as they do (Jackendoff 2002).

Integrating language research will be difficult because language represents most of the brain's simulation model, so language processes must connect to details in most other brain processes. This complexity has been previously expressed by those arguing for the need to solve the symbol grounding problem in order to create intelligent language behavior. One particularly difficult aspect of language is its role as a shortcut for brains to learn parts of their simulation models from other brains. As with other social processes, most reinforcement for learning language behavior must come from these other "teacher" brains. Note the general prejudice that knowledge and skills are not learned as well via language, sometimes called "book learning", as they are via trial and error.

The brain's simulation model is used for planning, and planning can help solve the credit assignment problem (Sutton and Barto 1998). At a high cognitive level, consider how chess players and mathematicians learn from the successes and failures of their plans. Their plans include an account of causality that they use to assign

credit for the success or failure of those plans. Successful consciously planned behaviors are learned as fast, unconscious responses, available as behavior elements in future plans. A brain design needs processes that implement such high-level planning and learning.

Consciousness

Consciousness is the most remarkable feature of human brains, and it is natural to ask whether it plays an essential role in intelligence. Crick and Koch believe that the purpose of consciousness is to provide a summary of the state of the world to the brain's planning functions (Koch 2004). This is similar to my own view that consciousness evolved as the ability of brains to process experiences that are not currently occurring (i.e., to simulate experiences) in order to solve the credit assignment problem for reinforcement learning (Hibbard 2002). Simulated experiences may be remembered, imagined, or a combination.

Jackendoff observes that the contents of consciousness are purely sensory, and at a low level of processing such as the 2.5-dimensional sketch for vision and phonemes for spoken language (Jackendoff forthcoming). He hypothesizes that the contents of consciousness are at the level where pure bottom up processing of sense information ends and feedback from higher levels becomes necessary to resolve ambiguity, and asks why this should be. If consciousness is the brain's primary simulator, then it should be purely in terms of sense information, and placing it at a low level would provide maximum flexibility in simulations. Furthermore, resolving ambiguity via feedback from higher levels of processing will often require iterating the low-level sense information, and the second and later iterations must necessarily use a simulation (i.e., a memory) of that low-level information.

Jackendoff also observes that our actual thoughts are unconscious and occur at a level above consciousness. For example, correctly formed sentences generally come into our minds and out of our mouths with our consciousness as mere observer rather than creator. Consciousness does not make decisions or solve puzzles, but merely observes. There are even experiments in which humans are mistaken about whether they or others made decisions (Koch 2004). Unconscious brain processes learn by reinforcement to produce and understand language, to interpret sense information, to make analogies, to propose and test solutions to puzzles, to solve the frame problem, and so on. This is not to suggest that any of these skills is simple: they may be learned as large numbers of cases, guided by complex prior biases. Learning in these high-level thought processes is analogous to reinforcement learning of fine and coordinated control of the many muscles in the hands.

Brain processes above and below the conscious level learn by reinforcement, but at the conscious level a different kind of learning occurs: memorization via the fast creation of resonate attractors in patterns of neural firing (Carpenter and Grossberg 2003; Sandberg 2003). This fast learning mechanism also serves to remember sequences of sense information and unconscious, internal representations. So, in addition to reinforcement learning, a brain design should include processes that remember low level sense information and internal representations.

Experiments with reinforcement learning by machines have achieved results competitive with humans in narrow problem areas, but fall far short of humans in general problem areas. It may be that the key to success in general problem areas is in the right configuration of interacting reinforcement learning processes and remembering processes. Baum suggests that another key is in the prior biases of learning processes (Baum 2004).

Human Safety

Brains are the ultimate source of power in the world, so artificial brains with much greater intelligence than humans potentially pose a threat to humans. My own view is that the best way to address this threat is through the values that reinforce learning of machine external behaviors in the world (Hibbard 2002). Specifically, behaviors that cause happiness in humans should be positively reinforced and behaviors that cause unhappiness in humans should be negatively reinforced. Furthermore, these values should include the happiness of all humans, and should be the only values reinforcing external behaviors. It is plausible that such a restriction is consistent with intelligence, since a value for the happiness of human teachers should suffice to reinforce learning of social skills like language, and values for accuracy of prediction apply to learning internal simulation behaviors.

Reinforcement learning algorithms include parameters that determine the relative weighting of short term and long term achievement of values. It is important that powerful artificial brains heavily weight long term human happiness, in order to avoid degenerate behaviors for immediate human pleasure at the cost of long term unhappiness. Such heavy long term weighting will cause artificial brains to use their simulation models to analyze the conditions that cause long term human happiness. The reinforcement values should probably also weight unhappiness much more heavily than happiness, so that machine brains focus their efforts on helping unhappy people rather than those who are already happy. This is similar to a mother who focuses her efforts on the children who need it most. In order to avoid positively reinforcing behaviors that cause the deaths of unhappy people, the reinforcement values may continue to include humans after their deaths, at the maximal unhappy value.

Of course, the details of reinforcement values for powerful machine brains are really political issues, to be settled by a political process with a hopefully educated and involved public. A political process will be necessary in

any case in order to avoid powerful machine brains whose values are simply the values of the organizations that create them: profits for corporations and political and military power for governments.

Valuing human happiness requires abilities to recognize humans and to recognize their happiness and unhappiness. Static versions of these abilities could be created by supervised learning. But given the changing nature of our world, especially under the influence of machine intelligence, it would be safer to make these abilities dynamic. This suggests a design of interacting learning processes. One set of processes would learn to recognize humans and their happiness, reinforced by agreement from the currently recognized set of humans. Another set of processes would learn external behaviors, reinforced by human happiness according to the recognition criteria learned by the first set of processes. This is analogous to humans, whose reinforcement values depend on expressions of other humans, where the recognition of those humans and their expressions is continuously learned and updated.

References

Baum, E. 2004. *What is Thought?* Cambridge, Mass.: MIT Press.

Brown, J., Bullock, D., and Grossberg, S. 1999. How the basal ganglia use parallel excitatory and inhibitory learning pathways to selectively respond to unexpected rewarding cues. *Journal of Neuroscience* 19(23), 10502-10511.

Carpenter, G., and Grossberg, S. 2003. Adaptive Resonance Theory. In *Handbook of Brain Theory and Neural Networks, Second Edition*, Arbib, M., ed. Cambridge, Mass.: MIT Press.

Hibbard, B. 2002. *Super-Intelligent Machines.* New York: Kluwer Academic / Plenum Publishers.

Jackendoff, R. 2002. *Foundations of Language.* New York: Oxford University Press.

Jackendoff, R. *Language, Culture, Consciousness: Essays on Mental Structure.* Cambridge, Mass.: MIT Press. Forthcoming.

Koch, C. 2004. *The Quest for Consciousness: a Neurobiological Approach.* Englewood, Colorado: Roberts and Co.

Minsky, M., Singh P. and Sloman A. 2004. The St. Thomas common sense symposium: designing architectures for human-level intelligence. *AI Magazine* 25(2): 113-124.

Sandberg, A. 2003. Bayesian Attractor Neural Network Models of Memory. Ph.D. diss. Institutionen foer Numerisk Analys och Datalogi, Stockholms Universitet.

Seymour, B., O'Doherty, J., Dayan, P., Koltzenburg, M., Jones, A., Dolan, R., Friston, K., and Frackowiak, R. 2004. Temporal difference models describe higher-order learning in humans. *Nature* 429, 664-667.

Sutton, R., and Barto, A. 1998. *Reinforcement Learning: An Introduction.* Cambridge, Mass.: MIT Press.