Behavioral Neuroscience
2007, Vol. 121, No. 1, 231–232

Copyright 2007 by the American Psychological Association
0735-7044/07/$12.00 DOI: 10.1037/0735-7044.121.1.231

COMMENTARIES

Biological Implementation of the Temporal Difference Algorithm for Reinforcement Learning: Theoretical Comment on O'Reilly et al. (2007)

James C. Houk
Northwestern University

The ability to survive in the world depends critically on the brain's capacity to detect earlier and earlier predictors of reward or punishment. The dominant theoretical perspective for understanding this capacity has been the temporal difference (TD) algorithm for reinforcement learning. In this issue of *Behavioral Neuroscience,* R. C. O'Reilly, M. J. Frank, T. E. Hazy, and B. Watz (2007) propose a new model dubbed primary value and learned value (PVLV) that is simpler than TD, and they claimed that it is biologically more realistic. In this commentary, the author suggests some slight modifications of a previous biological implementation of TD instead of adopting the new PVLV algorithm.

Keywords: dopamine, reinforcement learning, basal ganglia, temporal difference, credit assignment

Dopamine (DA) neurons have the special property of being able to predict future reward. The neuromodulatory signals they then send to medium spiny neurons in the striatum of the basal ganglia act on two time scales. The short-term modulatory action is an induction of bistability and nonlinear amplification in spiny neurons (Gruber, Solla, Surmeier, & Houk, 2003), which functions as an attentional factor (Nicola, Surmeier, & Malenka, 2000). The long-term modulation is a consolidation of synaptic strengths that occurs incrementally on a trial-by-trial basis (Charpier & Deniau, 1997), which functions in learning.

DA neuromodulation acts through intracellular second messengers. These are inherently slow pathways, although Gruber et al. (2003) concluded that they are sufficiently rapid to explain the short-term modulatory action mentioned above. The second messengers that mediate long-term modulation were postulated (Houk, Adams, & Barto, 1995) to play a decisive role in the actor–critic architecture that includes the temporal difference (TD) algorithm for reinforcement learning (Barto, Sutton, & Anderson, 1983). According to this model, DARPP-32 (Greengard, 2003) is instrumental in the creation of a time window of DA receptivity in the spines of medium spiny neurons. This was an elaboration of Sutton and Barto's (1990) trace concept for resolving the problem of temporal credit assignment in reinforcement learning.

In this issue, O'Reilly, Frank, Hazy, and Watz (2007) propose a new model that they have dubbed primary value and learned value (PVLV). The new model is simpler than TD, and it is claimed to be biologically more realistic. PVLV uses working memory instead of the trace mechanism postulated by TD to resolve the problem of temporal credit assignment. A point I come back to is that the trace mechanism has the advantage that it is local to the particular synapses that are important for acquiring the reward.

The problem of temporal credit assignment warrants some review (cf. Houk, 2005). Synaptic inputs act immediately to produce bursting discharges in spiny neurons. These bursts function to embody a small focus of activity in the area of cerebral cortex to which they project. The loop between that area of cerebral cortex and the cerebellum then amplifies and refines the spatiotemporal pattern of that activity to generate an output population vector that provides a precise representation of whatever action or thought is controlled by that area of cortex. Many of the synaptic actions that have occurred prior to the reward are crucial for obtaining the reward. The relevant ones need to be selectively reinforced, and the local aspect of the trace mechanism mentioned above could do that.

O'Reilly et al. (2007) rely on Joel, Niv, and Ruppin's (2002) review of actor–critic models to justify the introduction of their PVLV model. Joel et al. evaluated both anatomical and computational perspectives of actor–critic models. They reviewed several actor–critic models, and they introduced an alternative approach to modeling a critic network—one that uses evolutionary computational techniques to evolve an optimal reinforcement learning mechanism. They concluded that the previous anatomical models are not compatible with current data.

Here I suggest that correcting three problems with the biological implementation of the actor–critic architecture proposed by Houk et al. (1995) might be preferable to abandoning that model entirely. One problem with that earlier implementation is that the same spiny neuron was assumed to participate in both the direct and the indirect pathways to DA neurons. More recent anatomy does not support this assumption (Levesque & Parent, 2005), nor is it necessary. In fact, assuming separate spiny neurons in these two pathways would overcome a second problem noted by Joel et al. (2002), namely the failure to account for depression of DA neuron firing when reward is omitted. A third problem is that Houk et al. did not spell out the clear requirement of the TD algorithm for different learning rules in the actor and the critic.

Financial support for this article was provided by National Institute of Neurological Disorders and Stroke Grant P01 NS44383. I thank Andrew Barto, Enrico Mugnaini, and Jun Wang for helpful suggestions.

Correspondence concerning this article should be addressed to James C. Houk, Department of Physiology and Institute for Neuroscience, Feinberg School of Medicine, Northwestern University, 303 East Chicago Avenue, Chicago, IL 60611. E-mail: j-houk@northwestern.edu

The actor–critic architecture (Barto et al., 1983) has the interesting property that the actor and critic units differ in only a relatively minor way that is nevertheless critical. Both units use the same neuromodulatory signal (the TD error, which has been linked to the signaling of DA neurons) and almost the same learning rules in updating their synaptic weights. All the modifiable synapses require local memory to implement the necessary eligibility mechanism. Local memory ensures that the same synapses that were active at the time of important decisions are the ones that are reinforced. The only difference between these units is in what constitutes eligibility. For actors, synapses become eligible through a Hebbian-style correlation. That is, if a presynaptic signal participates in firing a unit, then that synapse becomes eligible for modification by a later modulatory signal, if one occurs. Thus, the activity of the postsynaptic unit is crucial to determining eligibility. In contrast, the eligibility mechanism of the critic unit remembers only past presynaptic activity, as opposed to past conjunctions of pre- and postsynaptic activity required of an actor unit. Consequently, the modifiable synapses of the critic unit must use a two-factor learning rule, whereas those of the actor units must use a three-factor learning rule.

Therefore, in a network implementation of the actor–critic architecture, both the actor units and the critic units use the same modulatory signal as a factor in updating their synaptic weights. All the modifiable synapses require local memory to implement the necessary eligibility mechanism. The eligibility mechanism of the critic units is one factor, meaning that there is no response contingency. This makes it similar to processes believed to be engaged in classical conditioning experiments in which there is no response contingency (Sutton & Barto, 1990). Therefore, when the modulatory signal is included, the critic units use a two-factor learning rule (one for eligibility, one for reward). However, the actor units use a two-factor eligibility trace, plus the reward, which is a three-factor learning rule.

Considerable progress in defining the distributed modular architecture of the brain has been made in recent years (cf. Houk, 2005). Strick and colleagues (cf. Kelly & Strick, 2004) continue to use viral tracers to define modules that interface the basal ganglia and the cerebellum with the neocortex. Strick's work on basal ganglia loops has focused on connectivity underlying the matrix modules of Houk et al. (1995), as opposed to the striosomal modules. Matrix modules were posited as actor models, whereas striosomal modules were posited as critic models. Although we still rely on earlier anatomy (Gerfen, Herkenham, & Thibault, 1987; Hedreen & DeLong, 1991; Jimenez-Castellanos & Graybiel, 1987; Selemon & Goldman-Rakic, 1990) for identifying reciprocity between DA neurons and spiny neurons, the agreement on this issue from four top laboratories is impressive. This is important because that reciprocity is critical for the ability of the model to replicate the bootstrapping property of the adaptive critic. This recursion-like operation allows DA neurons to find earlier and earlier predictions of reward. As pointed out by O'Reilly et al. (2007), there is a remarkable paucity of published work on these higher levels of conditioning even though they are crucial to the ability to survive in the world.

References

Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). Neuron-like elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man and Cybernetics, 13,* 835–846.

Charpier, S., & Deniau, J. M. (1997). In vivo activity-dependent plasticity at cortico-striatal connections: Evidence for physiological long-term potentiation. *Proceedings of the National Academy of Sciences, USA, 94,* 7036–7040.

Gerfen, C. R., Herkenham, M., & Thibault, J. (1987). The neostriatal mosaic: II. Patch- and matrix-directed mesostriatal dopaminergic and non-dopaminergic systems. *Journal of Neuroscience, 7,* 3915.

Greengard, P. (2003). The neurobiology of dopamine signaling. In H. Jörnvall (Ed.), *Nobel lectures, physiology or medicine 1996–2000* (pp. 328–347). Singapore: World Scientific Publishing.

Gruber, A. J., Solla, S. A., Surmeier, D. J., & Houk, J. C. (2003). Modulation of striatal single units by expected reward: A spiny neuron model displaying dopamine-induced bistability. *Journal of Neurophysiology, 90,* 1095–1114.

Hedreen, J. C., & DeLong, M. R. (1991). Organization of striatopallidal, striatonigral, and nigrostriatal projections in the Macaque. *Journal of Comparative Neurology, 304,* 569.

Houk, J. C. (2005). Agents of the mind. *Biological Cybernetics, 92,* 427–437.

Houk, J. C., Adams, J. L., & Barto, A. G. (1995). A model of how the basal ganglia generates and uses neural signals that predict reinforcement. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 249–274). Cambridge, MA: MIT Press.

Jimenez-Castellanos, J., & Graybiel, A. M. (1987). Subdivisions of the dopamine-containing A8–A9–A10 complex identified by their differential mesostriatal innervation of striosomes and extrastriosomal matrix. *Neuroscience, 23,* 223.

Joel, D., Niv, Y., & Ruppin, E. (2002). Actor–critic models of the basal ganglia: New anatomical and computational perspectives. *Neural Networks, 15,* 535–547.

Kelly, R. M., & Strick, P. L. (2004). Macro-architecture of basal ganglia loops with the cerebral cortex: Use of rabies virus to reveal multisynaptic circuits. *Progress in Brain Research, 143,* 449–459.

Levesque, M., & Parent, A. (2005). The striatofugal fiber system in primates: A reevaluation of its organization based on single-axon tracing studies. *Proceedings of the National Academy of Sciences, USA, 102,* 11888–11893.

Nicola, S. M., Surmeier, D. J., Malenka, R. C. (2000). Dopaminergic modulation of neuronal excitability in the striatum and nulceus accumbens. *Annual Review of Neuroscience, 23,* 185–215.

O'Reilly, R. C., Frank, M. J., Hazy, T. E., & Watz, B. (2007). PVLV: The primary value and learned value Pavlovian learning algorithm. *Behavioral Neuroscience, 120,* 31–49.

Selemon, L. D., & Goldman-Rakic, P. S. (1990). Topographic intermingling of striatonigral and striatopallidal neurons in the rhesus monkey. *Journal of Comparative Neurology, 297,* 359.

Sutton, R. S., & Barto, A. G. (1990). Time-derivative models of Pavlovian reinforcement. In M. Gabriel & J. Moore (Eds.), *Learning and computational neuroscience: Foundations of adaptive networks* (pp. 497–537). Cambridge, MA: MIT Press.

Received November 6, 2006
Revision received November 6, 2006
Accepted November 7, 2006 ■