Journal Club

Editor's Note: These short, critical reviews of recent papers in the *Journal*, written exclusively by graduate students or postdoctoral fellows, are intended to summarize the important findings of the paper and provide additional insight and commentary. For more information on the format and purpose of the Journal Club, please see http://www.jneurosci.org/misc/ifa_features.shtml.

Unblocking the Neural Substrates of Model-Based Value

Aaron M. Bornstein, Erik L. Nylen, and Sara A. Steele
New York University, New York, New York 10003
Review of McDannald et al.

A central premise of prominent reinforcement learning (RL) models is that actions and cues are considered valuable only insofar as justified by the rewards that actors learn to expect with them. This expectation can be adjusted with further experience, to the degree that new rewards received are greater or lower than expected. Learning driven by the difference between expected and experienced rewards, the reward prediction error (RPE), is a signature feature of RL systems. Importantly, when the experienced value exactly matches expectations, the RPE is zero and no learning occurs.

However, different stimuli may have similar values. In situations where only the identity of the reward (e.g., the color or flavor), and not the value (in, e.g., metabolic or hedonic units) changes, some RL models make the counterintuitive prediction that, because no RPE will be generated, no learning will occur. But, clearly, animals learn about new outcomes, even if their reward value is the same as was expected. To account for this phenomenon, a class of RL models also record the specific outcomes associated with cues and actions, regardless of their value. This record is called a world model and the algorithms are termed model-based, in contrast to model-free algorithms that simply cache estimated reward value. Though computationally expensive, model-based systems are uniquely able to inform goal-directed behaviors that depend on outcome identity.

As in value learning, identity associations may be updated via error-driven, incremental learning algorithms, leading to the suggestion of an identity prediction error (IPE). An open question is whether IPEs are reflected by the nigrostriatal dopamine system, as are RPEs (Gläscher et al., 2010). A significant literature treats the systems as parallel and at least partly independent (Daw et al., 2005). In the domain of instrumental learning, model-free and model-based processes have been linked to habitual (identity-insensitive) and goal-directed (identity-sensitive) actions, respectively (Redish et al., 2008). The independence of these systems has been supported by targeted lesions to subregions of dorsal striatum (Yin et al., 2008). However, it is unclear whether a similar parallelism exists in Pavlovian responding, thought to be largely supported by ventral striatum, and, if so, what the precise nature of this functional fractionation is.

In a recent issue of *The Journal of Neuroscience*, McDannald and colleagues (2011) evaluated the necessity of two brain structures thought to support identity-sensitive Pavlovian responses. Isolating behavioral and neural signatures of purely identity-based learning has proven notoriously tricky, in part because choice behavior necessarily reflects any previous response reinforcement. To circumvent this, the authors used an experimental manipulation to elicit the classical conditioning phenomenon of identity unblocking.

Traditionally, appetitive conditioning paradigms pair a cue (e.g., a light) with a reward (e.g., food). Blocking occurs when a second cue is paired with a cue–reward pair that has already been learned. Typically, the association between the new cue and the reward is not acquired (Kamin, 1969). However, if the cues are presented on trials in which the paired reward value changes, the second paired cue can acquire predictive power, i.e., it has become unblocked. The second cue can also become predictive if the outcome changes, even if the value of the new reward remains the same. Thus, identity unblocking reveals the influence of representations of a reward's features on responses. Conversely, a failure to observe identity unblocking could reflect a disruption of outcome-identity representations.

McDannald and colleagues (2011) performed chemotoxic lesions on rats, targeting two brain regions that have been strongly implicated in reward learning: ventral striatum (VS) and orbitofrontal cortex (OFC). OFC in particular has been identified as crucial to the flexible, putatively model-based decisions of interest in this study (Schoenbaum et al., 2009). OFC's role in representing both the sensory and motivational content of outcomes is thought to arise from its abundant input from sensory areas and reciprocal loops with limbic areas such as amygdala. For its part, VS is a primary target of midbrain dopaminergic nuclei (thought to signal RPE), sitting at the nexus of multiple networks by which the

Received April 13, 2011; revised May 13, 2011; accepted May 23, 2011.

This work was supported by MacCracken Graduate Fellowships from New York University. We thank Nathaniel D. Daw, Paul W. Glimcher, Franchesca Ramirez, and Jackie E. Reitzes for helpful conversations.

Correspondence should be addressed to Aaron M. Bornstein, Erik L. Nylen, or Sara A. Steele, New York University, 4 Washington Place, Suite 888, New York, NY 10003. E-mail: aaronb@nyu.edu, eln232@nyu.edu, or sas756@nyu.edu.

DOI:10.1523/JNEUROSCI.1883-11.2011
 0270-6474/11/3110117-02$15.00/0

values of cues and actions are computed and transmitted. After recovery, the rats were trained to associate three lights (A, B, and C) with food pellet rewards varying in amount and flavor. Next, in the unblocking procedures, each primary visual cue was presented together with a secondary auditory cue (X, Y, and Z; giving A + X, B + Y, C + Z). These compound cues were paired with either the same reward (the same flavor and quantity of pellets) or one that differed in either flavor or quantity. The authors took care to ensure that the (food-restricted) animals showed no preference for either flavor.

The authors then probed whether animals associated the new cues with new rewards, for example, whether sound Y signaled the presence of new food after identity unblocking, as light B signaled more food after value unblocking. Animals with OFC lesions responded to the new sound when it was paired with a reward with a different value, but not when it was paired with a reward that had the same value but a different flavor. Thus, the cues were value-unblocked but not identity-blocked, supporting the hypothesis that an intact OFC is necessary for identity-sensitive learning. Conversely, VS-lesioned animals showed reduced responding to cues associated with new values, and, surprisingly, also reduced responding to cues associated with new flavors. Thus, contrary to the stated hypothesis and a dominant theoretical framework, an intact VS appears necessary for the revision and/or expression of cue–outcome associations.

In demonstrating the necessity of VS for identity-based learning, this study strongly argues against the hypothesis of an isolated model-based reinforcement learning system operating in parallel with a model-free system. This result is congruent with recent data that point to involvement of VS in flexible decision-making (van der Meer and Redish, 2011). However, the study points to a single dissociation between systems in OFC, as ablating this structure did not disrupt value-based learning. Therefore, even if the systems converge in VS, the critical representations may still be distinct at some point.

The result highlights an ongoing debate about the representations supported by VS in reward learning. In particular, the VS consists of two subregions (shell and core) that have distinct connectivity to cortical and subcortical structures. Shell is emerging as a candidate for supporting value representations sensitive to fluctuations in outcome identity (Yin et al., 2008). Although the lesions performed by McDannald et al. (2011) predominantly affected core, a more targeted comparison of these subdivisions could help clarify hypothesized roles for VS. In Pavlovian conditioning terms, the behaviors examined here are what is called preparatory: reward-general, outcome-insensitive responses thought to be explicitly dependent on the core. A direct comparison of core and shell lesions on both preparatory and consummatory (responses whose physical manifestations are specific to aspects of outcome identity) behavior, in the context of this unblocking paradigm, could provide clarification of the representations encoded by each subregion.

Along these lines, the specific role of the OFC in representing identity-based information invites further study. In particular, the single dissociation by which OFC disrupts identity-based learning, but not value, is analogous to the disruption of goal-directed learning through lesions to dorsomedial striatum. However, without the complementary second dissociation identifying a region that exclusively disrupts value learning, sparing identity, the precise necessity of the OFC remains unclear. While the suggested shell/core distinction may yield further insight into this question, another tack would be to explore whether OFC is necessary and sufficient for another key feature of model-based learning: the flexible use of stimulus–stimulus contingencies during planning-based decisions for reward. These paradigms have been used successfully to elicit signatures of model-based behavior in humans (Daw et al., 2011) but have not yet been examined with the causal manipulations available to researchers of animal models.

The finding that VS is critical for model-based learning opens numerous exciting avenues for investigating the neural substrates of reinforcement learning. In particular, it suggests that a clean separation between model-based and model-free learning may not ultimately be a distinction wholly respected by the functional anatomy. For instance, the cascading, loop-like anatomical connectivity of the basal ganglia can be taken to support a partly serial model in which each striatal subregion is informed by (though perhaps not exclusively parasitic on) the representations maintained in others (Yin et al., 2008). This interpretation bears strong similarity to theoretical work in hierarchical reinforcement learning, though an explicit correspondence between these theories and model-based reinforcement learning is still in early stages of development (Botvinick and An, 2008). While normal animals can indeed exhibit identity-insensitive behavior after extensive training, this behavior may be uniquely expressed by habitual action chunks, potentially subserved by action chaining functions in the dorsolateral striatum, independent of VS cue evaluation. Thus, demonstrating parallel pathways in instrumental behavior cannot conclusively argue for a symmetric distinction in Pavlovian responding. The existence and role of a neural substrate of purely reward-guided value estimation remains an open question, worthy of extensive future evaluation.

References

Botvinick MM, An J (2008) Goal-directed decision making in prefrontal cortex: a computational framework. In: Advances in neural information processing systems (Koller D, Bengio YY, Schuurmans D, Bouttou L, Culotta A, eds), pp 169–176. New York: Curran Associates.

Daw ND, Niv Y, Dayan P (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci 8:1704–1711.

Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ (2011) Model-based influences on humans' choices and striatal prediction errors. Neuron 69:1204–1215.

Gläscher J, Daw N, Dayan P, O'Doherty JP (2010) States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron 66:585–595.

Kamin LJ (1969) Predictability, surprise, attention and conditioning. In: Punishment and aversive behavior (Campbell BA, Church RM, eds), pp 279–296. New York: Appleton-Century-Crofts.

McDannald MA, Lucantonio F, Burke KA, Niv Y, Schoenbaum G (2011) Ventral striatum and orbitofrontal cortex are both required for model-based, but not model-free, reinforcement learning. J Neurosci 31:2700–2705.

Redish AD, Jensen S, Johnson A (2008) A unified framework for addiction: vulnerabilities in the decision process. Behav Brain Sci 31: 415–437; discussion 437–487.

Schoenbaum G, Roesch MR, Stalnaker TA, Takahashi YK (2009) A new perspective on the role of the orbitofrontal cortex in adaptive behaviour. Nat Rev Neurosci 10:885–892.

van der Meer AA, Redish AD (2011) Ventral striatum: a critical look at models of learning and evaluation. Curr Opin Neurobiol. Advance online publication. Retrieved April 12, 2011. doi:10.1016/j.conb.2011.02.011.

Yin HH, Ostlund SB, Balleine BW (2008) Reward-guided learning beyond dopamine in the nucleus accumbens: the integrative functions of cortico-basal ganglia networks. Eur J Neurosci 28:1437–1448.