EJN fens EUROPEAN JOURNAL OF NEUROSCIENCE

European Journal of Neuroscience, Vol. 35, pp. 987–990, 2012 doi:10.1111/j.1460-9568.2012.08074.x

EDITORIAL
Beyond simple reinforcement learning: the computational neurobiology of reward-learning and valuation

John P. O'Doherty
California Institute of Technology, Pasadena, CA 91125, USA

Keywords: basal ganglia, computational neuroscience, conditioning, decision-making, prefrontal cortex

Abstract

Neural computational accounts of reward-learning have been dominated by the hypothesis that dopamine neurons behave like a reward-prediction error and thus facilitate reinforcement learning in striatal target neurons. While this framework is consistent with a lot of behavioral and neural evidence, this theory fails to account for a number of behavioral and neurobiological observations. In this special issue of EJN we feature a combination of theoretical and experimental papers highlighting some of the explanatory challenges faced by simple reinforcement-learning models and describing some of the ways in which the framework is being extended in order to address these challenges.

Introduction

Our understanding of the neural mechanisms by which predictions about future reward are learned, and of the means by which such learned predictions are used to guide reward-related decisions and behavior, has progressed significantly within the past two decades. Underpinning such progress in large part has been the observation that neural signals resembling the features of a class of simple computational learning models collectively called 'reinforcement learning' (RL), that were originally developed to enable reward-guided decision-making in artificial systems, appear to be present in the mammalian brain (Sutton & Barto, 1998). Specifically, the phasic activity of dopamine neurons has been found to resemble a type of learning algorithm signal called a temporal difference prediction error rule, in which the difference between successive temporal predictions of future reward is used to update the expected value of particular stimuli or actions (Montague *et al.*, 1996; Schultz *et al.*, 1997; Schultz, 1998). Moreover, neuronal responses in target areas of such dopamine neurons, such as the ventral and dorsal striatum, amygdala and orbitofrontal cortex, have been found to resemble the type of value signals that might be acquired by means of such a prediction error signal (Schoenbaum *et al.*, 1998; Gottfried *et al.*, 2003; Samejima *et al.*, 2005; Paton *et al.*, 2006). However, while this approach has given rise to a blossoming research agenda in both human and animal models, it is becoming increasingly clear that the simple temporal difference reinforcement learning (TDRL) framework is likely to provide only a partial account of the computational mechanisms underlying learning and choice in the mammalian brain. In this issue of EJN we explore some of the challenges facing TDRL-based theories of reward learning, and highlight a number of the recent advances that have been made in this area.

Challenges to the TDRL framework

Let us briefly consider a number of the explanatory gaps in the TDRL framework. The finding that action selection can be differentiated on a behavioral level into at least two distinct mechanisms, a goal-directed mechanism in which actions are selected with reference to the incentive value of the associated outcome and a habit-driven stimulus–response (S-R) mechanism in which actions are selected reflexively in a stimulus-driven manner (Balleine & Dickinson, 1998), has created difficulties for a TDRL account because such models are unable to capture the outcome value-sensitive features of goal-directed control (Daw *et al.*, 2005), nor are they capable of capturing the contingency degradation insensitivity known to be prevalent in habits (Balleine & O'Doherty, 2010). More challenges for the TDRL framework, at least insofar as its hypothesized dopaminergic basis goes, have emerged from studies that have investigated the extent to which the presence of dopamine is essential for reward learning by manipulating dopaminergic activity via either pharmacological or genetic methods (Berridge & Robinson, 1998; Robinson *et al.*, 2005; Flagel *et al.*, 2011). Such studies have found that the presence of dopamine appears not to be critical for at least some aspects of reward learning, as well as indicating possible contributions for this neurotransmitter in modulating other aspects of reward processing such as the performance of reward-related behaviors (Berridge, 2007). Additionally, while dopamine neurons are capable of coding for positive reward-prediction errors with considerable fidelity, these neurons appear to possess a very limited dynamic range with which to encode negative prediction errors, thereby rendering it difficult to implicate such neurons in learning about aversive as opposed to rewarding events (Bayer & Glimcher, 2005). Furthermore, in spite of the biological plausibility of dopamine-mediated RL via dopaminergic afferents into the ventral and dorsal striatum, the precise mechanism by which a distal reward-prediction error signal can ultimately come to mediate neural plasticity between stimuli and response representations, elicited perhaps seconds before, poses considerable mechanistic challenges.

Correspondence: J. O'Doherty, as above.
E-mail: jdoherty@caltech.edu

Received 9 February 2012, revised 14 February 2012, accepted 15 February 2012

Overview of the special issue

The aim of the current Special Issue of European Journal of Neuroscience is to provide an overview into some of the ongoing research efforts among both theoreticians and experimentalists on these and related problems, as described below.

Model-based RL and other extensions to the RL framework

One of the most influential proposals to have recently emerged is the proposal that, rather than reward learning being mediated exclusively by a TDRL system, there are at least two different mechanisms underpinning RL described above, and additionally a 'model-based' mechanism in which values are learned not by means of a reward-prediction error but instead are computed online using a learned model or 'cognitive map' of the decision problem (Doya *et al.*, 2002; Daw *et al.*, 2005). The model-based system is proposed as being responsible for learning goal-directed actions, while the model-free system is presumed to underpin the learning of S-R habits. In the present issue, we feature a number of contributions centered around this proposal. McDannald *et al.* (2012) review evidence of a role for both the orbitofrontal cortex and the ventral striatum in encoding predictive representations of an upcoming event consistent with model-based as opposed to model-free RL. These authors argue that the orbitofrontal cortex is better conceived of as exclusively a contributor to model-based RL, whereas the ventral striatum is held to contain neural representations consistent with both a model-based and a model-free system. The contribution by Noonan *et al.* (2012) reviews evidence for the role of the primate orbitofrontal cortex in learning associations between stimuli and both rewarding and punishing events. While Noonan and colleagues do not specify whether they are envisaging a model-based or model-free mechanism, it is unlikely that dopamine release in the orbitofrontal cortex would support dopamine prediction-error learning due to the slower temporal dynamics of dopamine release in cortex relative to the striatum (Seamans & Yang, 2004), indicating that model-free RL mechanisms are unlikely to contribute to learning in this region.

It has also been argued that the hippocampus contributes to model-based RL (Johnson *et al.*, 2007), as the encoding of relationships between events in which the hippocampus has long been implicated (in essence by forming a cognitive map) is a core feature of the model-based system. Accordingly, Bornstein & Daw (2012) describe findings from an fMRI study implicating the hippocampus in encoding and learning of expectancies based on stimulus–stimulus relationships, consistent with a putative role for this brain region in model-based learning.

An additional cognitive mechanism that might play a role in RL is that of working memory, in which rewards received on recent trials are used to guide future behavior by virtue of their encoding in a dynamic but capacity-limited working memory. Based on this hypothesis, Collins & Frank (2012) propose a computational model that has a mix of working memory and TDRL components, and they provide behavioral and neurogenetic evidence to support their claims. Dezfouli & Balleine (2012) take the model-based RL proposal one step further by suggesting that in fact all RL is done in a model-based manner and that there is no need at all to posit the existence of a model-free RL system. They develop a computational model which they argue accounts for the devaluation insensitivity and contingency degradation insensitivity effects of habits by conceiving of them as chunked or sequenced goal-directed actions.

Another major behavioral challenge for TDRL models is that, in their simplest implementation, these models do not accurately capture the way in which animals and humans discount the value of delayed rewards compared to rewards available immediately (Daw & Touretzky, 2002), nor do they take into account the influence of cognitive factors in modulating such discounting. Kurth-Nelson *et al.* (2012) describe a model of temporal discounting in which discounting emerges through a model-based cognitive search process aimed at determining which rewards will come available in the future, thereby suggesting a role for model-based RL mechanisms in contributing to intertemporal choice.

Also related to temporal aspects of choice, Sokol-Hessner *et al.* consider the timecourse of the computation of value in two brain regions that have been implicated in model-based RL: the ventromedial prefrontal cortex and the dorsolateral prefrontal cortex. They show that the timing and duration of value signals in these regions are modulated by the amount of time available to make a decision, a finding that has implications for understanding how value signals computed through model-based RL might be used to guide choice behavior.

Yet another emerging theme is that of a possible role for hierarchies within the RL system, according to which one node of the hierarchy resolves higher level decisions, such as for instance which goal should be selected or which task structure or state space is currently applying in a particular decision problem, while a lower node in the hierarchy determines which specific actions should be selected to obtain rewards (Koechlin *et al.*, 2003; Botvinick *et al.*, 2009). Along these lines, Venkatraman & Huettel (2012) propose a role for anterior dorsomedial frontal cortex in exerting higher order control during decision-making, implicating this region as a possible top level node within a hierarchical RL framework.

Stimulus and action generalization

A key issue that has not yet been adequately addressed regarding RL mechanisms (whether model-free or model-based) concerns how such systems can facilitate generalization across stimulus or response categories. In other words, if I've learned about how one particular stimulus or action sequence can generate rewards, to what extent do I generalize my reward prediction to other similar stimuli or action sequences and, if so, how is this done? Pan & Sakagami (2012) review a series of careful neurophysiological studies in which they implicate the prefrontal cortex in the capacity to learn about higher-order categories for stimuli, and thus facilitate generalization between stimuli as a function of the superordinate category to which those stimuli are deemed to belong. Interestingly, they report that while prefrontal neurons are capable of rapidly performing inferences about category membership, neurons in the dorsal striatum are much slower to do so, suggesting they need more trial-and-error experience in order to learn about the properties of stimuli, in a manner reminiscent of the proposed role for parts of the striatum in model-free as opposed to model-based learning. Wimmer *et al.* (2012) also address the issue of stimulus generalization by exploring the extent to which value predictions could be generalized incidentally across stimuli on the basis of uninstructed but experimentally induced correlations between those stimuli in their reward-predictability across trials. They report a role for the hippocampal formation in mediating encoding of the type of relational structure required for such value generalization. On the other hand, Hilario *et al.* (2012) investigate the role of different sub-regions of the striatum in mediating generalization of instrumental responding across different situations. They find that a lesion to the region of the dorsal striatum associated with goal-directed or 'model-based' action selection in fact enhances generalization, whereas a lesion to the region of dorsal striatum implicated in habits has the reverse effect.

Dopamine, RL and performance

As alluded to earlier, continuing efforts have been directed at establishing the precise role that dopamine might play in reward learning. Solving this problem depends on more fully understanding, at the neurobiological level, how it is that dopamine release can mediate plasticity in corticostriatal circuits, and determining the extent to which this phenomenon can be captured by TDRL. In this issue, Aggarwal *et al.* (2012) review what is known about the neurophysiological mechanisms underpinning dopamine release and the role of striatal circuits in learning and plasticity. They identify a number of cases in which findings from neurobiology deviate from what would be expected on the basis of a faithful implementation of TDRL.

A related concern is that dopamine has been implicated not only in learning but also in other aspects of reward processing. Berridge (2012) summarizes experimental evidence to support his argument for a role of dopamine in regulating incentive salience. It is argued that dopamine is involved not in reward-learning at all but rather exclusively in regulating motivational processes, i.e. in performance but not in learning. Consistent with this proposal, Smittenaar *et al.* (2012) report evidence from a study in human patients with Parkinson's disease who are on or off levodopa and/or dopamine agonists. These patients showed evidence for modulation in their performance on a reward-learning task but not in learning *per se*. This evidence suggests that a more complete understanding of the role of dopamine in reward will require computational accounts that can explain effects of dopamine on performance as well as on learning. In this context, Dayan (2012) develops an extension of a previously proposed model in which it is suggested that tonic levels of dopamine (as opposed to the phasic dopamine associated with reward-prediction errors) influence the degree of vigor of action selection, by virtue of a hypothesized role for tonic dopamine in representing long-run average reward (Niv *et al.*, 2007). In the paper published in this issue, this theoretical framework is extended into the aversive domain, in which actions are selected in order to actively avoid an aversive outcome. This paper also considers the thorny issue of how learning is mediated in the aversive as opposed to the rewarding domain, and the putative role of dopamine and other neurotransmitters such as serotonin in this process.

Bayesian and other learning mechanisms

Yet another influential approach has been to use models which deploy a Bayesian formulation in which the availability of rewards following the presence of particular stimuli or the selection of particular actions are modeled as full probability distributions, and Bayes' rule is used to perform inference about the likely distribution of rewards on a particular option given the sample evidence and one's prior expectations (Hampton *et al.*, 2006; Behrens *et al.*, 2007). O'Reilly *et al.*, 2012) provide a very accessible overview of Bayesian models and their use in neuroscience, including in reward-learning and decision-making. These Bayesian models also implicitly include a representation of a model of the decision problem, and for this reason they can be considered to share many commonalities with model-based RL. However, they are quite distinct from such RL models in the sense that, by encoding full probability distributions, these Bayesian models also facilitate computation of the degree of uncertainty one has in the estimated predicted reward (in essence by calculating the variance of the distribution). Such representations of uncertainty can be used in a variety of important ways to influence learning and choice. For instance, the rate at which one updates one's expectations about an outcome on the basis of new information should, from an optimal learning perspective, depend on the degree of uncertainty one has about that outcome: the more uncertain one is the more rapidly one should change one's expectations on the basis of new evidence. Funamizu *et al.* (2012) use a combination of TDRL to learn values and a Bayesian formulation to compute uncertainty and show effects in rat behavior on both the rate of learning and on choice behavior, confirming a role for uncertainty in driving reward-related behavior. However, one doesn't necessarily need to use a Bayesian approach to estimate uncertainty: in many situations an approximation of uncertainty can be generated through model-free trial-and-error updating. An example of this is the Pearce–Hall learning rule (Pearce & Hall, 1980) whereby uncertainty is approximated by the overall degree of surprise experienced for a given outcome given an initial prediction, denoted by an unsigned (absolute valued) prediction-error signal. According to this model, the larger the unsigned prediction error the greater the surprise and hence the greater is one's uncertainty in a particular prediction. Roesch *et al.* (2012) review evidence for the existence of unsigned surprise signals in the rodent and human amygdala which they interpret as reflecting a role for neurons in this region in controlling the allocation of attention toward more uncertain events (and hence presumably influencing the rate of learning).

Conclusion

The papers in this special issue are a testament to the enormous influence that the reward-prediction error hypothesis of dopamine has exerted on the field since it was first proposed (Montague *et al.*, 1996; Schultz *et al.*, 1997). These papers also highlight many of the challenges and controversies in accounting for behavioral and neural evidence within the RL framework, as well as pointing the way toward the resolution of at least some of these challenges. It is hoped that you will find them rewarding to digest.

Abbreviations

RL, reinforcement learning; S-R, stimulus–response; TDRL, temporal difference reinforcement learning.

References

Aggarwal, M., Hyland, B.I. & Wickens, J.R. (2012) Neural control of dopamine neurotransmission: implications for reinforcement learning. *Euro. J. Neurosci.*, **35**, 1115–1123.

Balleine, B.W. & Dickinson, A. (1998) Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology*, **37**, 407–419.

Balleine, B.W. & O'Doherty, J.P. (2010) Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology*, **35**, 48–69.

Bayer, H.M. & Glimcher, P.W. (2005) Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, **47**, 129–141.

Behrens, T.E., Woolrich, M.W., Walton, M.E. & Rushworth, M.F. (2007) Learning the value of information in an uncertain world. *Nat. Neurosci.*, **10**, 1214–1221.

Berridge, K.C. (2007) The debate over dopamine's role in reward: the case for incentive salience. *Psychopharmacology*, **191**, 391–431.

Berridge, K.C. (2012) From prediction error to incentive salience: mesolimbic computation of reward motivation. *Euro. J. Neurosci.*, **35**, 1124–1143.

Berridge, K.C. & Robinson, T.E. (1998) What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain Res. Brain Res. Rev.*, **28**, 309–369.

Bornstein, A.M. & Daw, N. (2012) Dissociating hippocampal and striatal contributions to sequential prediction learning. *Euro. J. Neurosci.*, **35**, 1011–1023.

Botvinick, M.M., Niv, Y. & Barto, A.C. (2009) Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. *Cognition*, **113**, 262–280.

Collins, A. & Frank, M.J. (2012) How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *Euro. J. Neurosci.*, **35**, 1024–1035.

Daw, N.D. & Touretzky, D.S. (2002) Long-term reward prediction in TD models of the dopamine system. *Neural Comput.*, **14**, 2567–2583.

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

Dayan, P. (2012) Instrumental vigour in punishment and reward. *Euro. J. Neurosci.*, **35**, 1152–1167.

Dezfouli, A. & Balleine, B.W. (2012) Habits, action sequences, and reinforcement learning. *Euro. J. Neurosci.*, **35**, 1036–1051.

Doya, K., Samejima, K., Katagiri, K. & Kawato, M. (2002) Multiple model-based reinforcement learning. *Neural Comput.*, **14**, 1347–1369.

Flagel, S.B., Clark, J.J., Robinson, T.E., Mayo, L., Czuj, A., Willuhn, I., Akers, C.A., Clinton, S.M., Phillips, P.E. & Akil, H. (2011) A selective role for dopamine in stimulus-reward learning. *Nature*, **469**, 53–57.

Funamizu, A., Ito, M., Doya, K., Kanzaki, R. & Takahashi, H. (2012) Uncertainty in action-value estimation affects both action choice and learning rate of the choice behaviors of rats. *Euro. J. Neurosci.*, **35**, 1179–1188.

Gottfried, J.A., O'Doherty, J. & Dolan, R.J. (2003) Encoding predictive reward value in human amygdala and orbitofrontal cortex. *Science*, **301**, 1104–1107.

Hampton, A.N., Bossaerts, P. & O'Doherty, J.P. (2006) The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *J. Neurosci.*, **26**, 8360–8367.

Hilario, M., Holloway, T., Jin, X. & Costa, R. (2012) Different dorsal striatum circuits mediate action discrimination and action generalization. *Euro. J. Neurosci.*, **35**, 1105–1114.

Johnson, A., van der Meer, M.A. & Redish, A.D. (2007) Integrating hippocampus and striatum in decision-making. *Curr. Opin. Neurobiol.*, **17**, 692–697.

Koechlin, E., Ody, C. & Kouneiher, F. (2003) The architecture of cognitive control in the human prefrontal cortex. *Science*, **302**, 1181–1185.

Kurth-Nelson, Z., Bickel, W.K. & Redish, A.D. (2012) A Theoretical account of cognitive effects in delay discounting. *Euro. J. Neurosci.*, **35**, 1052–1064.

McDannald, M.A., Takahashi, Y., Lopatina, N., Pietras, B., Jones, J.L. & Schoenbaum, G. (2012) Model-based learning and the contribution of the orbitofrontal cortex to the model-free world. *Euro. J. Neurosci.*, **35**, 991–996.

Montague, P.R., Dayan, P. & Sejnowski, T.J. (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.*, **16**, 1936–1947.

Niv, Y., Daw, N.D., Joel, D. & Dayan, P. (2007) Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology (Berl)*, **191**, 507–520.

Noonan, M.P., Kolling, N., Walton, M. & Rushworth, M. (2012) Re-evaluating the role of orbitofrontal cortex in reward and reinforcement. *Euro. J. Neurosci.*, **35**, 997–1010.

O'Reilly, J.X., Jbabdi, S. & Behrens, T.E. (2012) How can a Bayesian approach inform neuroscience? *Euro. J. Neurosci.*, **35**, 1168–1178.

Pan, X. & Sakagami, M. (2012) Category representation and generalization in the prefrontal cortex. *Euro. J. Neurosci.*, **35**, 1083–1091.

Paton, J.J., Belova, M.A., Morrison, S.E. & Salzman, C.D. (2006) The primate amygdala represents the positive and negative value of visual stimuli during learning. *Nature*, **439**, 865–870.

Pearce, J.M. & Hall, G. (1980) A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychol. Rev.*, **87**, 532–552.

Robinson, S., Sandstrom, S.M., Denenberg, V.H. & Palmiter, R.D. (2005) Distinguishing whether dopamine regulates liking, wanting, and/or learning about rewards. *Behav. Neurosci.*, **119**, 5–15.

Roesch, M., Esber, G.R., Li, J., Daw, N. & Schoenbaum, G. (2012) Surprise! Neural correlates of Pearce-Hall and Rescorla-Wagner coexist within the brain. *Euro. J. Neurosci.*, **35**, 1189–1199.

Samejima, K., Ueda, Y., Doya, K. & Kimura, M. (2005) Representation of action-specific reward values in the striatum. *Science*, **310**, 1337–1340.

Schoenbaum, G., Chiba, A.A. & Gallagher, M. (1998) Orbitofrontal cortex and basolateral amygdala encode expected outcomes during learning. *Nat. Neurosci.*, **1**, 155–159.

Schultz, W. (1998) Predictive reward signal of dopamine neurons. *J. Neurophysiol.*, **80**, 1–27.

Schultz, W., Dayan, P. & Montague, P.R. (1997) A neural substrate of prediction and reward. *Science*, **275**, 1593–1599.

Seamans, J.K. & Yang, C.R. (2004) The principal features and mechanisms of dopamine modulation in the prefrontal cortex. *Prog. Neurobiol.*, **74**, 1–58.

Smittenaar, P., Chase, H.W., Aarts, E., Nusselein, B., Bloem, B.R. & Cools, R. (2012) Decomposing effects of dopaminergic medication in Parkinson's disease on probabilistic action selection: learning or performance? *Euro. J. Neurosci.*, **35**, 1144–1151.

Sutton, R.S. & Barto, A.G. (1998) *Reinforcement Learning*. MIT Press, Cambridge, MA.

Venkatraman, V. & Huettel, S. (2012) Strategic control in decision making under uncertainty. *Euro. J. Neurosci.*, **35**, 1075–1082.

Wimmer, G.E., Daw, N.D. & Shohamy, D. (2012) Generalization of value in reinforcement learning by humans. *Euro. J. Neurosci.*, **35**, 1092–1104.