Available online at www.sciencedirect.com
ScienceDirect
Neural Networks 20 (2007) 668–675
Neural Networks
www.elsevier.com/locate/neunet

Multiple model-based reinforcement learning explains dopamine neuronal activity

Mathieu Bertin[a,b,*], Nicolas Schweighofer[c], Kenji Doya[a,d]

[a] *ATR Computational Neuroscience Labs, 2-2-2 Hikaridai, "Keihanna Science City", Kyoto 619-0288, Japan*
[b] *Laboratoire d'Informatique de Paris 6, Universite Paris 6 Pierre et Marie Curie, 4 place Jussieu 75005, Paris, France*
[c] *Department of Biokinesiology and Physical Therapy, University of Southern California, 1540 E. Alcazar St. CHP 155, Los Angeles 90089-9006, USA*
[d] *Neural Computation Unit, Initial Research Project Laboratory, Okinawa Institute of Science and Technology, 12-22 Suzaki, Gushikawa, Okinawa, 904-2234, Japan*

Received 18 February 2005; accepted 11 April 2007

Abstract

A number of computational models have explained the behavior of dopamine neurons in terms of temporal difference learning. However, earlier models cannot account for recent results of conditioning experiments; specifically, the behavior of dopamine neurons in case of variation of the interval between a cue stimulus and a reward has not been satisfyingly accounted for. We address this problem by using a modular architecture, in which each module consists of a reward predictor and a value estimator. A "responsibility signal", computed from the accuracy of the predictions of the reward predictors, is used to weight the contributions and learning of the value estimators. This multiple-model architecture gives an accurate account of the behavior of dopamine neurons in two specific experiments: when the reward is delivered earlier than expected, and when the stimulus–reward interval varies uniformly over a fixed range.

Keywords: Dopamine; Reinforcement learning; Multiple model; Timing prediction; Classical conditioning

1. Introduction

Reacting correctly to its environment requires an animal to continuously anticipate the consequences of its observations and actions. Understanding how these predictions are constructed, through statistical inference of the current observations and memory of past experiences, is therefore critical. In the simple case of classical conditioning experiments, a reward is delivered shortly after a cue stimulus. Through repeated pairing of this conditioned stimulus and a reward, the animal learns to use the stimulus as a predictor of the occurrence and timing of the following reward. The issue we address in this paper is the prediction of the precise timing of the stimulus–reward interval (SRI). We propose a new model showing how animals can learn to associate a number of possible SRI to a single stimulus. Our model notably offers an explanation for two otherwise puzzling experimental results concerning the role of dopamine when the SRI varies.

Direct evidence from electrophysiological recordings in monkeys (Montague, Dayan, & Sejnowski, 1996; Schultz, 1998) and indirect evidence from fMRI studies in humans (Pagnoni, Zink, Montague, & Berns, 2002) during these simple conditioning experiments strongly suggest that the activity of dopamine (DA) neurons encode the error between predicted reward and actual reward. Early in training, a burst of DA neurons activity occurs at the time of the reward delivery. As training progresses, this burst disappears, and instead a burst of activity occurs at the time of the cue stimulus. If however, the reward is unexpectedly not delivered in one trial, there is a "dip" in DA activity, precisely at the time when the reward was supposed to be delivered.

Several early computational models (Moore et al., 1986; Sutton & Barto, 1990) use temporal difference (TD) methods (Sutton, 1988) to describe experimental results of the conditioned nictitating membrane response. Application of TD learning theory to DA measurements in later studies

* Corresponding author at: ATR Computational Neuroscience Labs, 2-2-2 Hikaridai, "Keihanna Science City", Kyoto 619-0288, Japan. Tel.: +81 774 95 1235; fax: +81 774 95 1259.
E-mail address: mbertin@atr.jp (M. Bertin).

0893-6080/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2007.04.028

(Daw & Touretzky, 2002; Montague et al., 1996; Schultz, Dayan, & Montague, 1997; Suri & Schultz, 1999), could accurately reproduce DA neuron activity during simple conditioning in terms of prediction error. TD learning is a real-time learning strategy aiming at building accurate predictions based on past experience. The predictions are computed as a "value" function, a sum of the expected future rewards. At each instant, predictions are compared to actual outcomes; the error in prediction (TD error) is then used to update the value function.

In these earlier implementations of the TD learning theory – for consistency, we will only refer to the tapped delay line model (Montague et al., 1996) – time is sequenced in steps. The current state is implemented as a row vector $s(t)$ with $s_i(t) = 1$ if i is the time steps elapsed since the stimulus, and $s_j(t) = 0$ otherwise. At each time step, the agent builds a value function, $V(t)$, prediction of future (discounted) rewards:

$$V(t_0) = \sum_{t=t_0}^{\infty} \gamma^t r(t) \tag{1}$$

where γ $(0 < \gamma < 1)$ is a discounting parameter. Note that in the typical simple conditioning experiment, there is only one reward per trial, and the value function simply equals one discounted reward.

In neural network implementations, the value function of the current state is computed by the inner product:

$$V(t) = s(t) * (w(t))^{\mathrm{T}} \tag{2}$$

with $w(t)$ a weight row vector, and $(w(t))^{\mathrm{T}}$ its transpose. Through learning, these weights are updated at each time step according to the current prediction error:

$$w(t) = w(t) + \eta s(t)\delta(t) \tag{3}$$

where η a learning rate and $\delta(t)$ the TD error (scalar), which models the DA neurons' activity, and is given by:

$$\delta(t) = r(t) + \gamma V(t+1) - V(t). \tag{4}$$

During learning, the agent gradually builds a value function that correctly predicts the incoming reward. After learning, if the reward is given, the TD error is null at all time, except at the time of the conditioned stimulus. If however a reward is not given, the TD error is negative at the time the reward was expected.

Thus, the TD error given by these earlier models reproduces the DA neurons' activity remarkably well in the simple conditioning experiments. These models however fail to account for two recent experimental results in which the intervals between the conditioning stimulus and the reward are varied. We now describe these two experimental conditions on temporal variability: (1) earlier reward delivery, and (2) uniform variation of the stimulus–reward interval.

2. Experiments on temporal variability

2.1. Earlier reward delivery

In experiments conducted on dopamine measurements (Hollerman & Schultz, 1998), a monkey is trained to expect a reward precisely one second after the conditioned stimulus. After training, the reward is suddenly presented 0.5 s early or late. Three different types of DA responses were found (see Fig. 1).

- if the reward is given when expected, no change in DA activity is visible.
- if the reward is given late, a "dip" in DA activity occurs at the time the reward was expected; then there is a burst of activity shortly after the reward is finally given.
- if the reward is given early, a burst marks the reward delivery; however, no significant dip is observed at the time the reward was expected.

These observations follow the previous conclusions on the experiment (Hollerman & Schultz, 1998). While we will adhere to the authors' claims, we would state a few precautions concerning the use of this figure. Some of the results are undisputable, such as the presence of a dopamine burst shortly after early and delayed rewards. The dip of activity in the case of a late reward is also described as "significant" by the initial authors. Other results, however, are suject to qualitative interpretation. For example, we could not quantitatively detemine if one of the two bursts (early or late) is higher than the other. More importantly, the authors state that no dip of activity is observed following earlier reward. Although we will adhere to this observation in the current paper, further experiment might be needed to ascertain this result in a more quantitative way.

The tapped delay line model (Montague et al., 1996) reproduces accurately the first two types of responses, but fails in the last: it predicts a pause in DA activity at the time when the reward is usually expected. This happens because the agent only reacts time-step by time-step, and thus cannot infer that the reward it received earlier is the one it was expecting. In order to explain this experimental data, a new model based on a semi-Markov architecture has recently been proposed (Courville, Daw, & Touretzky, 2004). In this model, only two states are considered (ISI and ITI, inter stimulus and inter trial interval), and the agent tries through learning to predict the duration and probability of these two states. This model reproduces the above data accurately, because once the reward has been given, the agent does not expect any more rewards—thus, in the third condition, no dip of activity is created at the time the reward was given during training.

2.2. Uniformly varying stimulus–reward interval

In another experiment on DA measurement on monkeys (Fiorillo & Schultz, 2001), the interval between the stimulus and the reward varies uniformly over a fixed range (1–3 s) throughout training. Fig. 2 shows the response of a DA neuron during this experiment. When the SRI is short (lower part of the figure), a strong burst of activity marks the time of reward; for longer SRI (higher part of the figure), the observed burst of activity is lower. For the longest intervals, it appears impossible to state if there is actually a positive response.

Earlier TD models do not account for these results. In the Montague model, for instance, the value function is reorganized

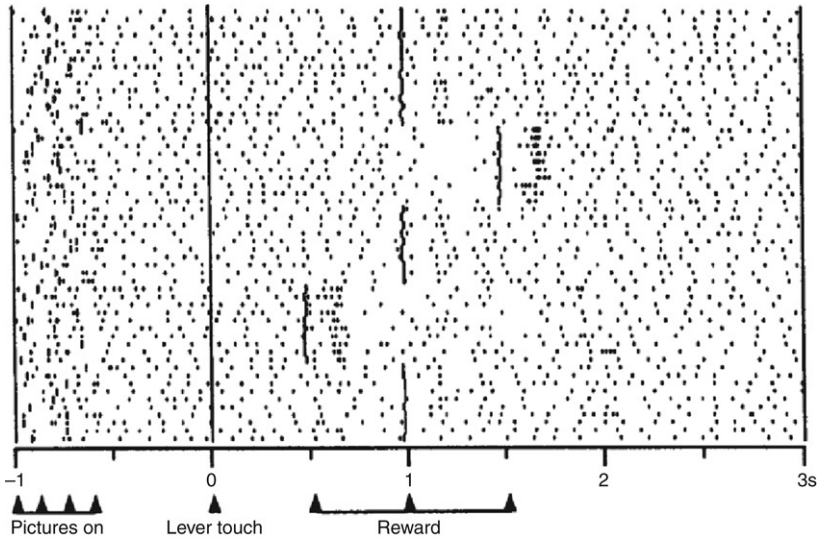

Fig. 1. Behavior of a dopamine neuron when a monkey expects a reward 1 s after the lever touch ((Hollerman & Schultz, 1998), with permission). Responding to the reward is seen in test trials when it is delivered half-second early or late. Note that following early reward, no subsequent dip of activity is observed at the time reward had been expected.

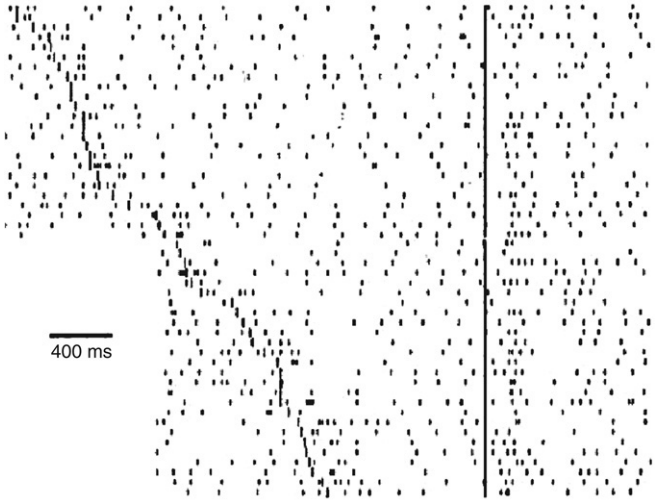

Fig. 2. Response of a dopamine neuron when the stimulus–reward interval varies uniformly over a 2 s range ((Fiorillo & Schultz, 2001) with permission). Rasters are sorted by stimulus–reward delay, with shortest delays at the bottom. The vertical line marks the time of the reward. More firing is seen after the reward for smaller delays.

according to the last SRI at each trial; because this interval changes at each trial, the agent cannot predict the next interval. Thus, the burst of dopamine neuron activity is similar whether the SRI is short or long (see Fig. 6(a)). The semi-Markov model (Courville et al., 2004) satisfyingly reproduces dopamine activity (Fig. 6(b)), with a TD error progressively decreasing with longer SRIs. However, for longer-than-average SRI, the semi-Markov model also predicts a negative TD error at the time of the reward. No below baseline activity is observable from the presented data (Fig. 2); but this conclusion could depend on the length of the time window over which spike counts is considered (Daw, Courville, & Touretzky, 2006).

In the present paper, we use a modular architecture to address these problems of interval prediction. Such a modular learning architecture decomposes a complex task into multiple domains of space and time (Wolpert & Kawato, 1998) and has been applied to reinforcement learning problems as Multiple-Model Reinforcement Learning (Doya, Samejima, Katagiri, & Kawato, 2002). Here, we propose an implementation of Multiple-Model Reinforcement Learning in the case of simple conditioning to model dopamine neuron activity.

3. Multiple Model-based Reinforcement Learning

The key property of a modular learning architecture is the capacity to learn distinct possible outcomes of a same cue stimulus. In the case of simple conditioning, this means the capacity of learning different possible reward distributions. For example, if a reward is presented repeatedly after an SRI t_1 or t_2 following a cue stimulus, a prediction model will be trained for each of the 2 possible delays, and the two prediction models will compete in the overall prediction.

The capacity to learn distinct outcomes of a cue stimulus is notably supported by a recent experiment of eyelid conditioning on rabbits (Ohyama & Mauk, 2001). In this work, rabbits are first subject to a training using a long SRI, until a low-criterion level of conditioned responding is attained. They are then trained longer to respond to a shorter SRI, to robust levels of conditioned responding. Subsequent exposure to a long cue stimulus reveals double-peak responses whose peaks are appropriately timed to the two SRIs. Thus, conditioning with a shorter SRI did not erase previous training of a longer interval.

In the Multiple Model-Based Reinforcement Learning (MMRL) (Doya et al., 2002) a complex task is decomposed into multiple domains in space and time, based on the predictability of the environmental dynamics. Using predictive models, each reinforcement learning module tries to predict the future states. A responsibility signal λ_i is assigned to each predictive model i depending on the likelihood of the current observed state

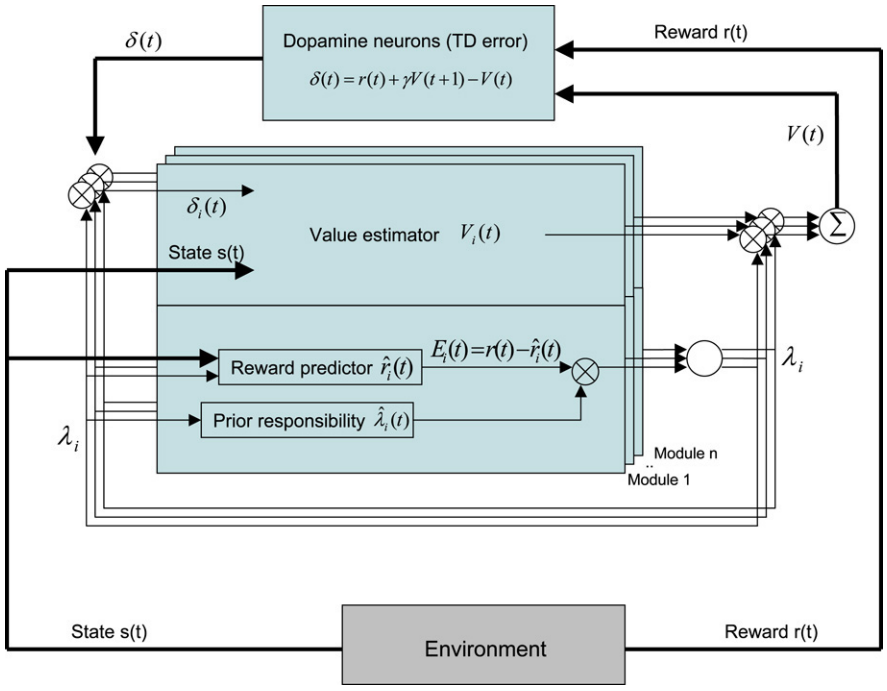

Fig. 3. Architecture of the MMRL (Multiple Model Reinforcement Learning) model. The activity of dopamine neurons is given by the global TD error $\delta(t)$. See text for abbreviations.

and the reliability of the past predictions. This responsibility signal is then used for four different purposes: weighting the outputs of predictive models, gating the learning of predictive models, weighting the action outputs, and gating the update of reinforcement learning controller.

Here we propose an implementation of MMRL for classical conditioning (see Fig. 3). Each module is composed of a reward predictor and a value estimator. The output of the model is the prediction of the presence or absence of a reward at each time step following the cue stimulus (as in the tap delay line model (Montague et al., 1996), we describe time as a sequence of steps). Therefore, each reward predictor simply gives a vector of the predicted amount of reward for each time step following the cue stimulus. At each step, the prediction errors of the reward predictors are used to compute the responsibility signal λ_i of the corresponding modules. This responsibility signal is then used to gate the update of the reward predictors and the update of the value estimators, and to weight the output of the value estimators.

To illustrate the functioning of our model, let us take the following example. Suppose that an agent, using $N > 2$ modules has previously trained two reward predictors 1 and 2, predicting an SRI of t_1 and t_2 respectively, with $t_1 < t_2$. This means each of those 2 modules predicts a low reward at each time step, except at the step t_1 or t_2 respectively, where they predict an amount of reward depending on the previous training. If the current time t elapsed since the stimulus occurred is small ($t < t_1 < t_2$), the two predictors propose a likely prediction, and they will both have high responsibility: the overall prediction will take the two predictors into account. At the step $t = t_1$, if no reward occurs, the first predictor makes a large prediction error, and its responsibility is downgraded. Therefore, while $t_1 < t < t_2$, the overall prediction will mostly depend on predictor 2. If at $t = t_2$ no reward occurs, the predictor 2 in turn makes a large prediction error, and its responsibility is also downgraded. If a reward occurs afterwards, the most likely of the remaining predictors will receive the highest responsibility.

We now describe in detail the role for each module of the reward predictor and the value estimator.

3.1. Reward predictor

The reward predictor $\hat{r}_i(t)$ of each module i gives a vector of the predicted reward for each time step following the cue stimulus. The prediction errors are used to compute the responsibility (Wolpert & Kawato, 1998):

$$\lambda_i(t) = \frac{\hat{\lambda}_i(t) e^{-\frac{E_i(t)^2}{2\sigma^2}}}{\sum_j \hat{\lambda}_j(t) e^{-\frac{E_j(t)^2}{2\sigma^2}}} \tag{5}$$

where σ is a constant, and $E_i(t) = r(t) - \hat{r}_i(t)$ the prediction error of each reward predictor, with $r(t)$ the actual amount of reward at time step t.

$\hat{\lambda}_i(t)$ is a "responsibility predictor" (Wolpert & Kawato, 1998), encoding some prior knowledge or belief about module selection. Here we consider as prior knowledge the temporal continuity of module selection (Doya et al., 2002). Thus, we compute $\hat{\lambda}_i(t)$ using the previous responsibility:

$$\hat{\lambda}_i(t) = \lambda_i(t-1)^{\alpha} \tag{6}$$

where $0 < \alpha < 1$ is a parameter that controls the strength of the memory effect.

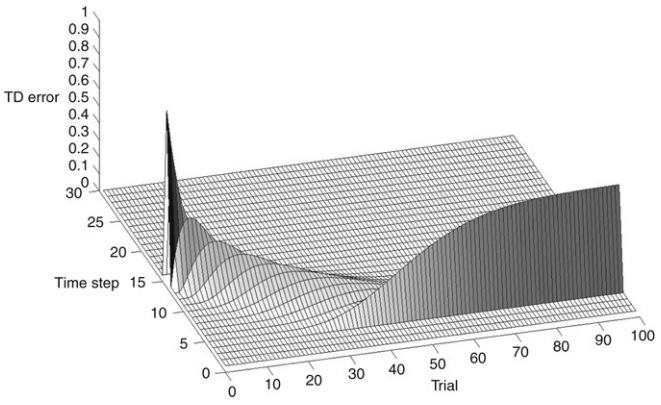

Fig. 4. MMRL model, evolution of the TD error through learning when the stimulus–reward interval is constant.

Each reward predictor $\hat{r}_i(t)$ is initialized with random values, and is updated at each time step by gating the error with the responsibility signal:

$$\hat{r}_i(t) = \hat{r}_i(t) + \mu\lambda_i E_i(t) \tag{7}$$

with $0 < \mu < 1$ an update rate.

3.2. Value estimator

The value estimator of each module compute and update values in a way similar than the tap delay line model (Eqs. (2) and (3)). Thus, for each module i:

$$V_i(t) = s(t) * (w_i(t))^{\mathrm{T}} \tag{8}$$

each w_i being a weight row vector. The global predicted value signal $V(t)$ is a weighted sum of the modules' values:

$$V(t) = \sum_i \lambda_i V_i(t). \tag{9}$$

The TD error $\delta(t)$ is computed as in the earlier TD model (Eq. (4)). We use the responsibility signal to gate the TD error for each module:

$$\delta_i(t) = \lambda_i \delta(t). \tag{10}$$

This gated TD error is finally used to update the weight vector of each value estimator:

$$w_i(t) = w_i(t) + \eta s(t)\delta_i(t) \tag{11}$$

with η a learning rate.

4. Results

We first tested MMRL in the simple conditioning experiment, i.e., when the SRI stays constant. Fig. 4 shows the evolution of TD error through learning. These results reproduce those of the tap delay line model (Montague et al., 1996). Further, the model also reproduced the "dip" seen in the reward omission case (result not shown). Note that in this case, only one module is eventually used and trained. If the agent has more than one module, the one offering the most reliable prediction will quickly reach a high responsibility, and the other modules will therefore have a negligible effect on the TD error course.

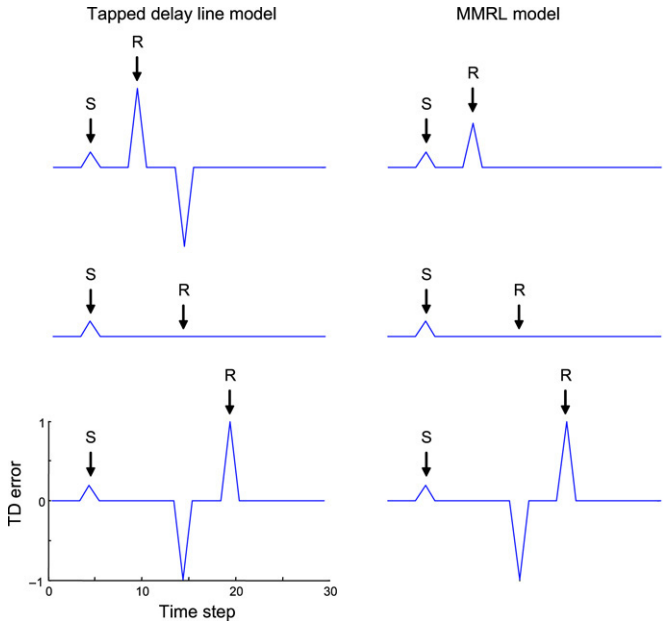

Fig. 5. Simulated dopamine response in early and delayed reward conditions, for the tapped delay line model (left) and MMRL model (right). S: cue stimulus. R: reward. After a 150 trials training with a 10 time steps ISI, the reward is presented 5 steps earlier (top), on time (middle), or 5 steps later (bottom). The tapped delay line model wrongly produces a negative error after an earlier reward (left, top), at the time the reward was expected. This negative error doesn't occur in our MMRL model (right, top), in accord with experimental data.

We then ran simulations in which we varied the SRI. After training with a given interval for 100 trials, the reward is suddenly presented earlier or later than expected. As shown in Fig. 5, when the reward is delivered on time (middle), or later (bottom), the results are identical to those of the tapped delay line model. A difference occurs when the reward is delivered earlier (top): in MMRL, no inhibition is seen at the time the reward is expected, which is consistent with the previous interpretation of the experimental data (Fig. 1). In that case, one module trained to the usual SRI had a higher responsibility than the other modules. When the unpredicted earlier reward occurs, this module makes a high prediction error, and its responsibility is downgraded. Thus, at the time of the usual reward, the prediction of this previously trained module is not taken into account, and no negative TD error is observed.

MMRL predicts a smaller response in the case of early reward (top) than for a late one (bottom), as does Daw's semi-Markov model (Daw, 2003).

We then ran simulation of the uniformly varying SRI experiment, first with the tap delay line model, then with the semi-Markov model, using the fully observable implementation (Daw, 2003), and finally with our MMRL implementation. At each trial of the simulation, an interval was randomly selected, length varying between 3 and 7 time steps, and the corresponding prediction error was recorded. Fig. 6 shows the means for the different possible intervals after 1000 trials (we omit the first 200 trials, during which the MMRL response

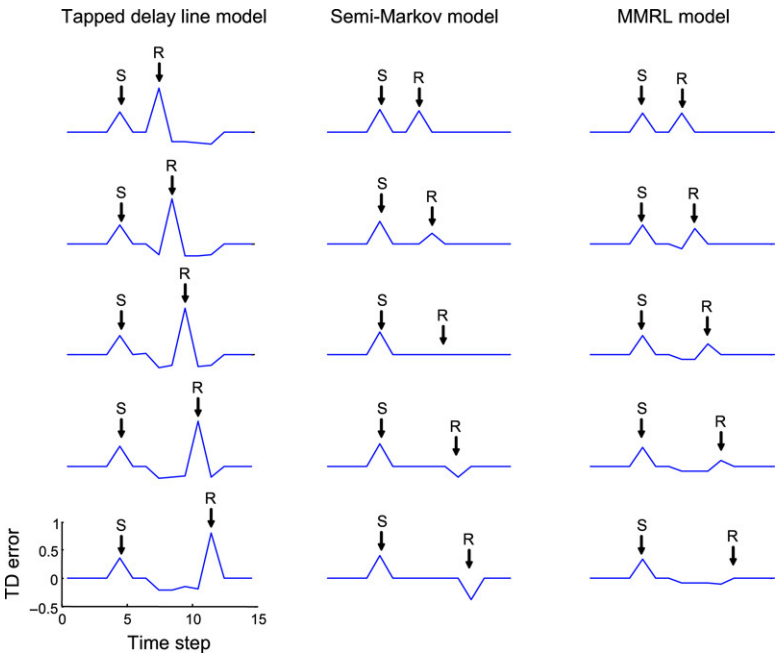

Fig. 6. Simulated dopamine responses when ISI varies uniformly over a range. The tapped delay line model (left) incorrectly predicts identical excitation for all possible reward locations. The semi-Markov model (center) and MMRL model (right) predict decreasing excitation as the interval gets longer, as observed in experimental data. The semi-Markov model also predicts a negative response for longer-than-average rewards, which is not obvious in experimental data.

looks erratic while the modules are not fully trained, see below). The MMRL model (right) predicts a TD error decreasing as the SRI becomes longer; nearing zero for the longest interval, which is consistent with experimental data (Fig. 2).

In the case of MMRL, we use a number of modules equal or higher than the number of possible SRIs (5 possible intervals in the preceding simulation). After a training period, whose length depends on the parameters of the simulation (200 trials for the results presented above), each possible SRI is predicted by one of the modules: each of these modules predicts a reward at the time of the corresponding SRI, and no reward otherwise.

The satisfactory behavior of MMRL, for the trials following this initial training period, can be explained by the definition of the global value function as a weighted sum of the modules' values. When the reward is delivered early, all the modules have similar responsibilities. The TD error then takes into account the prediction of all the modules, including the ones predicting a late reward. A large positive TD error at the time of the reward is thus observed. If the reward is delivered late, for each possible previous SRI, a module makes a prediction error, and sees its responsibility reduced. Thus, at the time of the reward, only a fraction of the modules still have a high responsibility, and the given reward is therefore more likely. As a result, a smaller positive TD error is observed at the time of reward.[1]

5. Discussion

We proposed a Multiple-Model Reinforcement Learning model that can accurately reproduce data of two previous classical conditioning experiments: earlier reward delivery (Hollerman & Schultz, 1998), and uniform variation of the stimulus–reward interval (Fiorillo & Schultz, 2001).

In its current implementation, our model presents, however, a notable limitation: the number of reward predictors-value estimator modules must be at least equal to the number of possible SRIs. A larger number of modules will have a negligible effect on the TD error course, but a fewer number will result in a highly unstable response. Therefore, a better precision, in terms of sharpness of time sequencing, can only be attained as the expense of higher model complexity. The issue of temporal discrimination has been addressed by the Spectral Timing model (Grossberg & Merrill, 1992; Grossberg & Schmajuk, 1989). This model, based on the integration of a multitude of differentially active timing signals occurring synchronously, produces through conditioning a bell-shape response signal centered on the learned SRI, its spread increasing with the length of the SRI. Integrating such a representation of time in the MMRL could provide more realistic results while setting a more satisfying balance between precision and model complexity.

What could be the neural substrate of our model? Parallelism is an important property of the basal ganglia—cerebral cortex learning circuitry (Hoover & Strick, 1993; Tanaka et al., 2004). Reinforcement learning models of the basal ganglia e.g., Doya (2000) and a recent imaging study (O'Doherty et al., 2004) suggest that the ventral striatum computes value functions. Thus, we speculate that multiple value functions $Vi(t)$ are computed in the ventral striatum, and summed in the ventral globus pallidus. Learning of the value function would occur via the dopaminergic reinforcement signals modulated by serotonergic inputs directly at the level of the striatum (Lucas, De Deurwaerdere, Porras, & Spampinato, 2000). The reward

[1] Note that in the Montague model, the agent has equal expectation for all the possible reward locations; thus, except at the time the reward is actually delivered, a slight negative TD error occurs.

predictors could possibly reside in the cerebellum, as it has been implicated in the prediction of the precise timing of events and in the associations of multiple distinct SRI to a single stimulus (Fiala, Grossberg, & Bullock, 1996; Millenson, Kehoe, & Gormenzano, 1977; Ohyama & Mauk, 2001). As serotonin appears to control the selection of different cortico-basal ganglia loops in reward prediction at different time scales (Schweighofer, Tanaka, et al., 2004; Tanaka et al., 2004), serotonergic projections could carry the responsibility signals necessary for learning and controlling appropriate reward predictors and value estimators (Schweighofer, Doya, & Kuroda, 2004).

6. Conclusion

We introduced a new model of dopamine neurons based on the MMRL architecture. This model, by augmenting the earlier TD dopamine models, could simulate experiments dealing either with earlier delivered reward or with uniformly varying intervals between conditioned stimulus and rewards.

In this MMRL implementation, the agent uses a number of representations, or predictors, of its current experience. Each reward predictor represents a possible reward distribution that the animal has inferred from past experience. To each is assigned a responsibility, which represents the credit the agent gives to each predictor. Each time the stimulus–reward association is presented, the agent updates those responsibilities according to the outputs of the predictors, and to what was actually observed.

7. Simulation procedures

7.1. Simple conditioning (Fig. 4)

We use a constant number of time steps for each trial, with a constant location for the delivery of the stimulus and the reward. No re-initialization is made at the end of the trials; a trial to trial transition is treated the same way than a within-trial time step transition.

Parameters: number of trials = 100; time steps by trial = 30: reward value = 1; Stimulus-reward interval = 10; number of modules = 2; $\eta = 0.2$; $\mu = 0.6$; $\gamma = 0.95$; $\alpha = 0.9$; $\sigma = 0.1$.

7.2. Earlier reward delivery (Fig. 5, right)

Each of the three parts (top, middle, and bottom) were simulated independently. In each case, we first train the model using a fixed stimulus–reward interval. At the last trial (again, without any re-initialization during trial transition) the reward timing is changed according to the situation simulated.

Parameters: number of trials = 150; time steps by trial = 30; reward value = 1; Stimulus-reward interval =10 (5, 10 or 15 at the last trial); number of modules = 2; $\eta = 0.2$; $\mu = 0.4$; $\gamma = 0.85$; $\alpha = 0.84$; $\sigma = 0.05$.

7.3. Uniformly varying stimulus–reward interval (Fig. 6, right)

The simulation is run as before, but the stimulus–reward interval is chosen randomly for each trial over a 5 time steps range, using a uniform law. For each possible reward location, we plot the average of the TD time course of all the corresponding trials (except the 200 first trials—see main text).

Parameters: number of trials = 1000 (including the 200 first trials); time steps by trial = 30; reward value = 1; Stimulus-reward interval = randomly chose from the [3 7] interval; number of modules = 5; $\eta = 0.2$; $\mu = 0.5$; $\gamma = 0.8$; $\alpha = 0.81$; $\sigma = 0.3$.

For all the simulations above, we tested different numbers of modules. A higher number of modules had only marginal effect on the results. However, a smaller number resulted in highly unstable results, especially in the case of Fig. 6.

We also tested the effect of a variation of the ITI (inter-trial-interval, time between the reward delivery of a trial and the cue stimulus of the following trial). In the simpler tapped delay line model, updates occur only after the cue stimulus, while reward occurs or is expected (Eq. (3)). Thus, a change of ITI has no effect on the learning behavior (as long as there can be no confusion between the cue stimuli of two different trials). MMRL is a bit more sensible, since the responsibility of the modules is updated at each time step (Eq. (6)), including the time steps of the ITI. However, test simulations in which the ITI varied randomly over a 5 time steps range showed no change in behavior.

Acknowledgements

Mathieu Bertin was supported by a Lavoisier grant of the French Foreign ministry, and Nicolas Schweighofer was supported in part by NSF grant IIS 0535282.

References

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2004). Similarity and discrimination in classical conditioning: A latent variable account. *Advances in Neural Information Processing Systems*, *17*, 313–320.

Daw, N. D. (2003). Reinforcement learning models of the dopamine system and their behavioral implications. *Ph.D. Thesis*. CMU Dept of Computer Science.

Daw, N. D., Courville, A. C., & Touretzky, D. S. (2006). Representation and timing in theories of the dopamine system. *Neural Computation*, *18*, 1637–1677.

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

Doya, K. (2000). Complementary roles of basal ganglia and cerebellum in learning and motor control. *Current Opinion of Neurobiology*, *10*, 732–739.

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

Fiala, J. C., Grossberg, S., & Bullock, D. (1996). Metabotropic glutamate receptor activation in cerebellar Purkinje cells as substrate for adaptive timing of the classically conditioned eye-blink response. *Journal of Neuroscience*, *16*, 3734–3760.

Fiorillo, C. D., & Schultz, W. (2001). The reward responses of dopamine neurons persist when prediction of reward is probabilistic with respect to time or occurrence. In Society for Neuroscience Abst. 27: 827.5.

Hollerman, J. R., & Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience*, *1*, 304–309.

Hoover, J. E., & Strick, P. L. (1993). Multiple output channels in the basal ganglia. *Science*, *259*, 819–821.

Grossberg, S., & Merrill, J. W. L. (1992). A neural network model of adaptively timed reinforcement learning and hippocampal dynamics. *Cognitive Brain Research*, *1*, 3–38.

Grossberg, S., & Schmajuk, N. A. (1989). Neural dynamics of adaptive timing and temporal discrimination during associative learning. *Neural Networks*, *2*, 79–102.

Lucas, G., De Deurwaerdere, P., Porras, G., & Spampinato, U. (2000). Endogenous serotonin enhances the release of dopamine in the striatum only when nigro-striatal dopaminergic transmission is activated. *Neuropharmacology*, *39*, 1984–1995.

Millenson, J. R., Kehoe, E. J., & Gormenzano, I. (1977). Classical conditioning of the rabbit's nictitating membrane response under fixed and mixed CS–US intervals. *Learning and Motivation*, *8*, 351–366.

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

Moore, J. W., Desmond, J. E., Berthier, N. E., Blazis, D. E. J., Sutton, R. S., & Barto, A. G. (1986). Simulation of the classically conditioned nictitating membrane response by a neuron-like adaptive element: Response topography, neuronal firing and inter-stimulus intervals. *Behavioral Brain Research*, *21*, 143–154.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, *304*, 452–454.

Ohyama, T., & Mauk, M. (2001). Latent acquisition of timed responses in cerebellar cortex. *Journal of Neuroscience*, *21*(2), 682–690.

Pagnoni, G., Zink, C. F., Montague, P. R., & Berns, G. S. (2002). Activity in human ventral striatum locked to errors of reward prediction. *Nature Neuroscience*, *5*, 97–98.

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

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

Schweighofer, N., Doya, K., & Kuroda, S. (2004). Cerebellar aminergic neuromodulation: Towards a functional understanding. *Brain Research Reviews*, *44*, 103–116.

Schweighofer, N., Tanaka, S. C., Asahi, S., Okamoto, Y., Doya, K., & Yamawaki, S. (2004). An fMRI study of the delay discounting of reward after tryptophan depletion and loading. 1. Decision-making Program No. 776.14. 2004 Abstract Viewer/Itinerary Planner. Washington, DC: Society for Neuroscience. Online.

Suri, R. E., & Schultz, W. (1999). A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. *Neuroscience*, *91*, 871–890.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, *3*, 9–44.

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). MIT Press.

Tanaka, S. C., Doya, K., Okada, G., Ueda, K., Okamoto, Y., & Yamawaki, S. (2004). Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nature Neuroscience*, *7*, 887–893.

Wolpert, D. M., & Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks*, *11*, 1317–1329.