Neural Circuits Trained with Standard Reinforcement Learning Can Accumulate Probabilistic Information during Decision Making

Nils Kurzawa
n.kurzawa@stud.uni-heidelberg.de
Medical Research Council Brain Network Dynamics Unit, University of Oxford,
Oxford, OX1 3QT, U.K., and Institute of Pharmacy and Molecular Biotechnology,
University of Heidelberg, D-69120 Heidelberg, Germany

Christopher Summerfield
christopher.summerfield@psy.ox.ac.uk
Department of Experimental Psychology, University of Oxford,
Oxford OX1 3UD, U.K.

Rafal Bogacz
rafal.bogacz@ndcn.ox.ac.uk
Medical Research Council Brain Network Dynamics Unit, University of Oxford,
Oxford OX1 3UD, U.K., and Nuffield Department of Clinical Neurosciences,
University of Oxford, Oxford OX1 3UD, U.K.

Much experimental evidence suggests that during decision making, neural circuits accumulate evidence supporting alternative options. A computational model well describing this accumulation for choices between two options assumes that the brain integrates the log ratios of the likelihoods of the sensory inputs given the two options. Several models have been proposed for how neural circuits can learn these log-likelihood ratios from experience, but all of these models introduced novel and specially dedicated synaptic plasticity rules. Here we show that for a certain wide class of tasks, the log-likelihood ratios are approximately linearly proportional to the expected rewards for selecting actions. Therefore, a simple model based on standard reinforcement learning rules is able to estimate the log-likelihood ratios from experience and on each trial accumulate the log-likelihood ratios associated with presented stimuli while selecting an action. The simulations of the model replicate experimental data on both behavior and neural activity in tasks requiring accumulation of probabilistic cues. Our results suggest that there is no need for the brain to support dedicated plasticity rules, as the standard mechanisms proposed to describe reinforcement learning can enable the neural circuits to perform efficient probabilistic inference.

1 Introduction

Humans and other animals often have to choose a course of action based on multiple pieces of information. Consider a cat deciding whether to chase a bird in your back garden. Her decision will depend on multiple factors: how far away the bird is, how tasty it looks, and how long it is until humans will provide a bowl of kibble. To make the best decisions in novel or unfamiliar settings, animals have to learn by trial and error how to weigh information appropriately. In this letter, we study this question in the context of laboratory tasks in which multiple cues signal which response is most likely to be rewarded. We focus on how the weight associated with each cue is learned over time, in situations where multiple stimuli are present at each trial, and participants need to learn to appropriately assign credit to each cue for successes or failures. This, in turn, will facilitate subsequent decision making.

Two broad classes of theory have been developed to describe learning and decision making in such situations. The classical theory of reinforcement learning (RL) suggests that animals learn to predict scalar reward outcomes for each cue or combination of cues (Rescorla & Wagner, 1972). When multiple cues are presented, the animal's total expected reward is a sum of rewards associated with the stimuli presented. Following feedback, the individual reward expectations are updated proportionally to the reward prediction error, defined as the difference between reward obtained and expected. This model naturally generalizes to learning about expected rewards following actions (Sutton & Barto, 1998). The model also captures essential aspects of learning in basal ganglia; much evidence suggests that reward prediction error is encoded in phasic activity of dopaminergic neurons (Schultz, Dayan, & Montague, 1997; Fiorillo, Tobler, & Schultz, 2003), which modulates synaptic plasticity in the striatum (Reynolds, Hyland, & Wickens, 2001; Shen, Flajolet, Greengard, & Surmeier, 2008).

Another line of theoretical research, based on the sequential probability ratio test (SPRT), has focused on describing the integration of information by humans or animals during perceptual classification tasks (Gold & Shadlen, 2001; Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Yang & Shadlen, 2007; de Gardelle & Summerfield, 2011). In order to explain the SPRT, let us consider a probabilistic categorization task in which monkeys can choose either a red or a green target by fixating either of them with their gaze (Yang & Shadlen, 2007). Choices follow a combination of four sequentially displayed shapes, each of which has a different probability of appearing on trials when the green or red response was rewarding. Gold and Shadlen (2001) proposed that in such tasks, animals learn the log likelihood of each stimulus given the hypothesis of either action being correct:

$$w_s = \log\left(\frac{P(s|A_1)}{P(s|A_2)}\right). \tag{1.1}$$

In equation 1.1, A_1 and A_2 denote the hypotheses that a saccade to a red or green target, respectively, will result in a reward, and s is an index of the stimulus, which is in range $s \in \{1, \ldots, m\}$, where m is the number of different stimuli that can be presented during the task. $P(s|A_1)$ and, respectively, $P(s|A_2)$ describe the likelihood that stimulus s would be observed given either A_1 or A_2, and w_s is the weight of evidence (WOE) which is defined as the log ratio of both likelihoods. Stimulus s provides evidence for action A_1 if $w_s > 0$, and vice versa.

Representing log-likelihood ratios allows easy integration of information during decision making (Gold & Shadlen, 2001). While making a decision on the basis on n stimuli, we wish to choose an action with a higher posterior probability given the observed stimuli. However, instead of computing the posterior probabilities themselves, it is easier to compute a decision variable equal to the log ratio of posterior probabilities (Gold & Shadlen, 2007):

$$\begin{aligned}
&\log\left(\frac{P(A_1|s_1, \ldots, s_n)}{P(A_2|s_1, \ldots, s_n)}\right) \\
&= \log\left(\frac{P(A_1)P(s_1, \ldots, s_n|A_1)}{P(A_2)P(s_1, \ldots, s_n|A_2)}\right) \\
&= \log\left(\frac{P(A_1)}{P(A_2)}\right) + \log\left(\frac{P(s_1|A_1)}{P(s_1|A_2)}\right) + \ldots + \log\left(\frac{P(s_n|A_1)}{P(s_n|A_2)}\right) \\
&= \log\left(\frac{P(A_1)}{P(A_2)}\right) + \sum_{j=1}^{n} w_{s_j}. \tag{1.2}
\end{aligned}$$

In the transition from the first to the second line, we used Bayes' theorem, and in the transition from the second to the third line, we assumed conditional independence of stimuli. Thus, the ratio of posterior probabilities can be simply computed by adding the WOEs associated with presented stimuli to a term representing the initial, prior probabilities (this term is equal to zero when the prior probabilities of the two actions are equal). Choosing action A_1 or A_2 when the sign of the above decision variable is positive or negative, respectively, is equivalent to choosing the action with a higher posterior probability.

These theories (RL and SPRT) have largely been developed in parallel. Several models have attempted to combine the two approaches and describe how animals learn WOEs of stimuli using RL (Soltani & Wang, 2010; Coulthard et al., 2012; Berthet, Hellgren-Kotaleski, & Lansner, 2012; Soltani,

Khorsand, Guo, Farashahi, & Liu, 2016). However, these models employed novel synaptic plasticity rules, which for some of the models were relatively complex, and there is no evidence that synapses can implement these rules. Building on the ideas from these earlier models, this letter shows that WOEs can also be learned with a standard and simple plasticity rule. In particular, we show that for a certain class of tasks, the expected reward for selecting action A_1 with a stimulus s present is approximately linearly proportional to w_s. Therefore, a simple model based on the standard Rescorla-Wagner rule is able to learn WOEs and accumulate the WOEs associated with presented stimuli while selecting an action. The main novel contribution of this letter is showing that the learning of WOEs (previously demonstrated only with unconventional learning rules) can also be achieved with the standard RL plasticity rules, which are thought to be implemented in the basal ganglia circuits.

In the next section, we describe the class of tasks under consideration and present a model of learning in these tasks. In section 3, we show that this model approximates decision making through accumulation of WOEs, analyze how the WOEs estimated by the model depend on task and model parameters, and compare the model with the data. Finally, in section 4, we compare the proposed model with previous models and discuss further experimental predictions.

2 Model

We consider a class of tasks often used to investigate the neural bases of probabilistic decision making (Knowlton, Mangels, & Squire, 1996; Yang & Shadlen, 2007; Philiastides, Biele, & Heekeren, 2010; de Gardelle & Summerfield, 2011; Coulthard et al., 2012). On each trial, participants choose between two actions on the basis of multiple sensory cues, and on a given trial, only one of the actions is rewarded. At a start of each trial, n cues are presented, which are sampled with replacement from a set of m. The probability of each cue appearing depends on which action is rewarded on a given trial, so the subject can deduce from the cues which action is more likely to be rewarded. After the decision, a reward of $r = 1$ is received if the correct action was selected and no reward $r = 0$ if the incorrect action is selected.

2.1 Reinforcement Learning Model.
To capture learning in such tasks we employ a very simple model (single-layer perceptron) that learns based on the standard Rescorla-Wagner rule for learning in tasks with multiple stimuli. This model is schematically illustrated in Figure 1. It is composed of an input layer (e.g., putative cortical sensory neurons) x_s selective for the cues s and an output layer (e.g., putative striatal neurons) y_i selective for actions i. A similar two-layer structure has been used in other models of
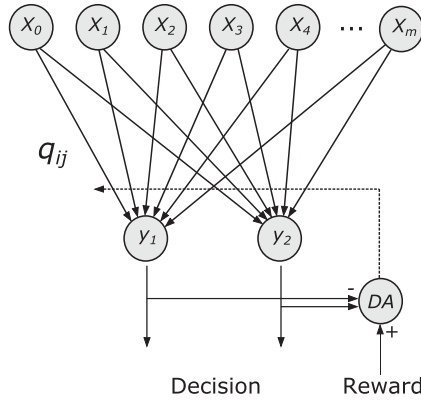
Figure 1: Schematic illustration of the model. The input layer is composed of $m + 1$ nodes $x_0, x_1, \ldots, x_m$ representing stimuli, and each of them is connected to nodes selective for actions y_1 and y_2 via connections with weights q_{ij}. Dopaminergic neurons (DA) receive input encoding reward, and inhibition encoding the value of chosen action, and thus compute reward prediction error. The dopaminergic neurons modulate changes of synaptic weights q_{ij}.

learning stimulus-response associations (Law & Gold, 2009; Gluck & Bower, 1998).

The nodes x_s have activity equal to the number of stimuli s present on a given trial (i.e., $x_s = 0$ if stimulus s is not present, $x_s = 1$ if stimulus s is present, $x_s = 2$ if two copies of stimulus s are present). Additionally, node x_0 is always set to $x_0 = 1$; we will refer to it as a bias node. The reward prediction for an action A_i is defined simply as the synaptic inputs to nodes y_i:

$$y_i = \sum_{j=0}^{m} q_{ij} x_j. \tag{2.1}$$

In equation 2.1, q_{ij} denote the synaptic weights from a neuron selective for stimulus j to the neuron selective for action i. Thus, for $j > 0$, the weights q_{ij} describe by how much the expected reward for selecting action i increases after observing stimulus j, while q_{i0} describes the expected reward for selecting action i irrespective of stimuli presented.

After computing y_i, the action is chosen stochastically such that the probability of selecting A_i follows the softmax distribution,

$$P_i = \frac{e^{\beta y_i}}{\sum_{u=1}^{2} e^{\beta y_u}}, \tag{2.2}$$

where β is a parameter that controls whether the models choose actions with the highest expected reward (high β) or explore different actions (low β).

After the choice, the weights q_{ij} to the neuron selective for the chosen action A_i are updated with

$$q_{ij} = q_{ij} + \alpha(r - y_i)x_j, \tag{2.3}$$

where α represents the learning rate (which was set to $\alpha = 0.05$ for all simulations). According to this rule, the weights between sensory nodes representing presented stimuli and the node representing the chosen action are modified proportionally to the reward prediction error $(r - y_i)$. So these weights are increased if the reward was higher than predicted and decreased if the reward was lower than predicted.

To implement such learning, at the time of choice, a memory trace, known as the eligibility trace, needs to form in synapses between neurons selective for presented stimuli and the neurons selective for the chosen action (Sutton & Barto, 1998). Subsequently, when the feedback is provided, the eligible synapses should be modified proportionally to the reward prediction error. The reward prediction error is thought to be encoded in the phasic activity dopaminergic neurons (Schultz et al., 1997). They receive inhibitory input from striatal neurons (Watabe-Uchida, Zhu, Ogawa, Vamanrao, & Uchida, 2012), which in our model encode y_i. Assuming that the dopaminergic neurons also receive an input encoding reward, they could subtract these two inputs and compute $r - y_i$. The dopaminergic neurons send dense projections to striatum and modulate plasticty of cortico-striatal synapses (Shen et al., 2008).

It has been also proposed how the weights q_{ij} are physically represented in strengths of cortico-striatal connections. The striatal projection neurons can be divided in two groups: those whose activity can facilitate movements (Go neurons, expressing D1 receptors) and those inhibiting movements (NoGo neurons, expressing D2 receptors) (Kravitz et al., 2010). Computational models have been proposed in which the weights of Go neurons increase, while the weights of NoGo neurons decrease when the prediction error is positive, and vice versa when prediction error is negative (Frank, Seeberger, & O'Reilly, 2004; Collins & Frank, 2014; Mikhael & Bogacz, 2016). It has been shown that for a certain class of plasticity rules, the difference between the weights of Go and NoGo neurons encodes q_{ij}, that is, this difference evolves according to equation 2.3 (Mikhael & Bogacz, 2016).

At the start of each simulated experiment, weights are initialized to $q_{ij} = 0$ for $j > 0$, while the weights from the bias node are set to $q_{i0} = 0.5$ and kept constant in all simulations except for those in section 3.3.

2.2 Generating Stimuli. Before each simulated trial, it was decided randomly which action would be rewarded, according to prior probabilities

$P(A_i)$, which were set to $P(A_1) = P(A_2) = 0.5$ in all simulations except when indicated otherwise. Depending on which action i was rewarded, n cues were drawn randomly with replacement according to probabilities $P(s|A_i)$.

In the experimental studies considered here, the WOEs for individual stimuli are reported (rather than $P(s|A_i)$). Here, for consistency, we also assume that the stimuli are assigned unique weights w_j, and in most simulations we use $\{w_1, w_2, \ldots, w_8\} = \{-2, -1.5, -1, -0.5, 0.5, 1, 1.5, 2\}$. We compute the probabilities $P(s|A_i)$ from w_s using

$$P(s|A_1) = \frac{2 \cdot \sigma(w_s)}{m}, \quad P(s|A_2) = \frac{2 \cdot \sigma(-w_s)}{m}, \tag{2.4}$$

where $\sigma(w)$ is a sigmoid function:

$$\sigma(w) = \frac{1}{1 + e^{-w}}. \tag{2.5}$$

Equations 2.4 satisfy the desired constraints, as the logarithm of ratio of probabilities defined in this ways is w_s, and the probabilities $P(s|A_i)$ add up to 1 across stimuli for sets of WEOs we consider, which contain pairs of stimuli with opposite WOEs.

3 Results

First, we investigate the values to which the weights converge in the model. We first derive a general condition that the weights need to satisfy at the stochastic fixed point and then analyze its implications for different variants of the task.

At the stochastic fixed point, the expected change in weights in equation 2.3 must be 0, that is, $E(r - y_i) = 0$, which implies that the weights at the stochastic fixed point must satisfy

$$\sum_{j=0}^{m} q_{ij}^* x_j = E(r|A_i, s_1, \ldots, s_n). \tag{3.1}$$

Since we assumed that only one action is rewarded, the expected reward for choosing action A_1 is equal to

$$E(r|A_1, s_1, \ldots, s_n) = 1 \cdot P(A_1|s_1, \ldots, s_n) + 0 \cdot P(A_2|s_1, \ldots, s_n)$$
$$= P(A_1|s_1, \ldots, s_n) \tag{3.2}$$

(which can analogously be derived for A_2). Because we wish to relate the expected reward to a ratio of probabilities, we note that for two alternatives

$P(A_1|s_1, \ldots, s_n) + P(A_2|s_1, \ldots, s_n) = 1$, the following relationship holds:

$$\frac{P(A_1|s_1, \ldots, s_n)}{P(A_2|s_1, \ldots, s_n)} = \frac{P(A_1|s_1, \ldots, s_n)}{1 - P(A_1|s_1, \ldots, s_n)}. \tag{3.3}$$

Rearranging terms, we obtain

$$P(A_1|s_1, \ldots, s_n) = \sigma\left(\log\frac{P(A_1|s_1, \ldots, s_n)}{P(A_2|s_1, \ldots, s_n)}\right), \tag{3.4}$$

where σ is a sigmoid function defined in equation 2.5. Combining equations 3.1, 3.2, 3.4, and using Bayes' theorem, we obtain the relationship between the synaptic weights learned by RL and WEOs:

$$\sum_{j=0}^{m} q_{1j}^* x_j = \sigma\left(\log\left(\frac{P(A_1)}{P(A_2)}\right) + \sum_{j=1}^{n} w_{s_j}\right). \tag{3.5}$$

To make it easier to understand what weights q_{ij}^* satisfy the above condition, we start with a very simple version of the task and progress through analysis of more complex versions.

3.1 Learning with a Single Stimulus. We first consider a simple case when only $n = 1$ stimulus is presented on a trial and prior probabilities of two actions are equal. When a single stimulus j is presented, only sensory nodes x_j and x_0 are equal to 1, while other sensory nodes are 0. Since we assumed equal prior probabilities of action, we also fix $q_{i0} = \frac{1}{2}$, so then equation 3.5 becomes

$$q_{1j}^* = \sigma(w_j) - \frac{1}{2}. \tag{3.6}$$

Figure 2A shows the values of weights at the end of the simulation, and indeed for one stimulus presented per trial ($n = 1$), the weights after learning q_{1j} are close to function $\sigma(w_j) - \frac{1}{2}$.

Furthermore, it is important to consider that the sigmoid function $\sigma(w)$ features an approximately linear region for $w \in [-1, 1]$. The linear approximation of sigmoid can be found by a Taylor expansion of $\sigma(w)$ around 0:

$$\sigma(w) \approx \sigma(0) + \sigma'(0)w$$

$$= \frac{1}{2} + \frac{e^{-0}}{(1 + e^{-0})^2}w$$
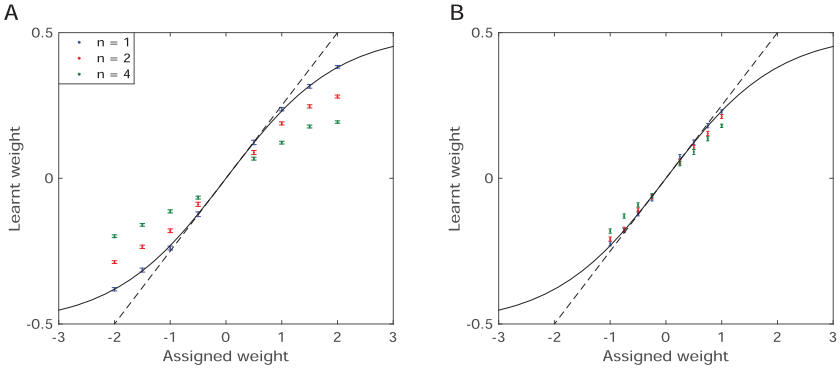
$$= \frac{w}{4} + \frac{1}{2}. \tag{3.7}$$

Figure 2: Learned weights for different ranges of assigned weights. (A) Final learned weights q_{1j} for different numbers of stimuli presented per trial after 100 repetitions of 5000 learning iterations with exploration parameter $\beta = 0$ were plotted over assigned weights: $\{w_1, w_2, \ldots, w_m\} = \{-2, -1.5, -1, -0.5, 0.5, 1, 1.5, 2\}$. Standard errors are indicated by error bars. The solid line represents $\sigma(w_j)$ and the dashed line $\frac{w_j}{4}$. (B) Final learned weights q_{1j} after analogous simulations with assigned weights: $\{w_1, w_2, \ldots, w_m\} = \{-1, -0.75, -0.5, -0.25, 0.25, 0.5, 0.75, 1\}$.

Thus, when we simulated a task with WOE $w_j \in [-1, 1]$, we observed that the weights learned by the model could be well approximated by $q_{1j} \approx \frac{w_j}{4}$ (see the dashed line in Figure 2B).

3.2 Learning with Multiple Stimuli. The analysis from section 3.1 can be naturally extended to the case when n stimuli $s_1, \ldots, s_n$ are presented. Then equation 3.5 becomes

$$\sum_{j=1}^{n} q_{1s_j}^* = \sigma(w_{cum}) - \frac{1}{2}, \tag{3.8}$$

where $w_{cum} = \sum_{j=1}^{n} w_{s_j}$. When WOEs are chosen such that on a majority of trials, $w_{cum} \in [-1, 1]$, then the sigmoid function in equation 3.8 can be approximated by the linear function, and weights $q_{ij} = \frac{w_j}{4}$ approximately satisfy that equation, which implies that the weights converge to similar values as for the case of single stimulus ($n = 1$). This is illustrated in Figure 2B, which shows results of a simulation with WOEs relatively close to 0. One can see that the weights for $n = 2$ and even $n = 4$ are relatively close to those for $n = 1$.

When more extreme WOEs are used and $w_{cum} \notin [-1, 1]$, the linear approximation does not hold. The simulations of this case are shown in

Figure 2A, where symbols of different colors indicate how the weights in the model depend on the number n of stimuli presented within a trial. The weights converge to less extreme values when more stimuli are presented. In this simulation, when $n > 1$, the value of $\sigma(w_{cum})$ is more likely to exceed the linear range, and therefore the final weighting by the model for each individual q_{ij} will be damped. Let us, for example, consider the case in which the model is presented with two stimuli in a trial: with $w_{s_1} = 1$ and $w_{s_2} = 1.5$. If the weights were equal to the values as for $n = 1$ (i.e., $q_{1s_1} \approx 0.25$, and $q_{1s_2} \approx 0.375$), then the expected reward would be $y_1 = 1.175$, so even if the reward $r = 1$ is received, the prediction error is negative and the weights are decreased. The more stimuli are presented per trial, the more the weights learned by the model will be damped.

Nevertheless, it is remarkable in Figure 2A that the weights q_{1j} learned by the model remain an approximately linear function of w_j, even for $n > 1$ when they are damped. This happens because all weights are damped, as even the stimuli with w_j closer to 0 may co-occur on the same trial with stimuli with high w_j, and so w_{cum} may exceed the linear range of the sigmoid, and weights of all stimuli on that trial will be damped. We will see in section 3.5 that for highly extreme weights, this linear relationship breaks, but the relationship between q_{1j} and w_j remains approximately linear for a wide range of weights used in Figure 2A, which is similar to those used typically in experimental studies (Yang & Shadlen, 2007; Philiastides et al., 2010).

Since $q_{1j} \approx c w_j$ where c is a proportionality constant, the activity of neurons selective for actions can be approximated by

$$y_1 \approx c w_{cum} + \frac{1}{2}, \quad y_2 \approx -c w_{cum} + \frac{1}{2}. \tag{3.9}$$

Thus, the activity of the action-selective nodes is proportional to the accumulated WOE of stimuli presented so far (i.e., to the decision variable of equation 1.2).

3.3 Learning Prior Probabilities. In all simulations so far, we assumed for simplicity that the two actions were correct equally often, and we fixed the weights from the bias node to $q_{i0} = 0.5$. Here we analyze to what values these weights converge when the probabilities of two actions are no longer the same.

Since q_{i0} encode the expected reward for selecting action i regardless of the stimuli, we would expect them to converge to $q_{i0} = P(A_i)$. In simulations where the number n of stimuli per trial was fixed, q_{i0} converged to a value between 0.5 and $P(A_i)$ but closer to 0.5 (see Figure 3A). So although the q_{i0} moved slightly toward $P(A_i)$, they never reached it. This happened because such simulated trials did not sufficiently constrain learning. For example,
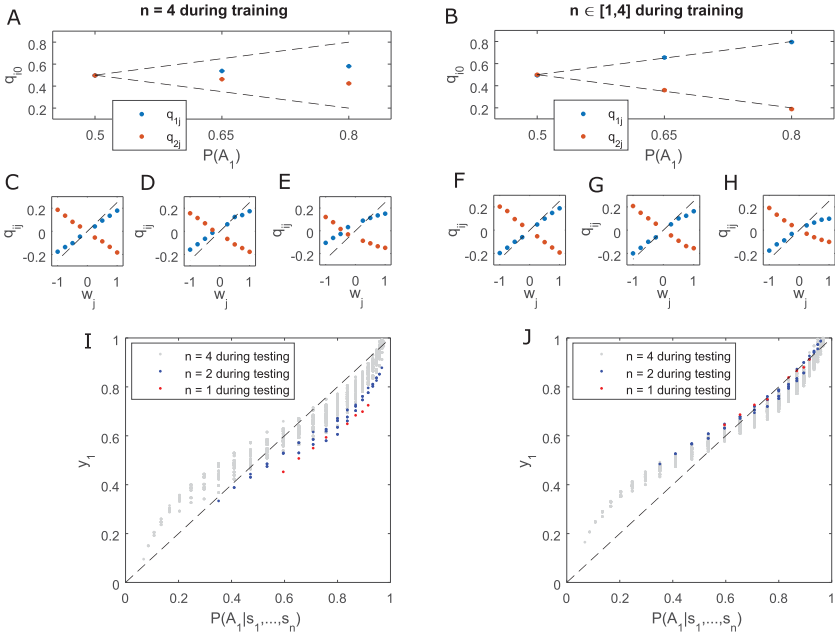
Figure 3: Learning prior probabilities of actions. The left panels (A, C–E, I) were obtained in simulations in which on each training trial, $n = 4$ stimuli were presented, while the right panels (B, F–H, J) came from simulations in which the number n of stimuli presented on each trial was randomly chosen between 1 and 4. (A, B) The synaptic weights from the bias node obtained in simulation with different prior probabilities of A_1 (dashed lines indicate the prior probabilities). (C–H) The synaptic weights from neurons representing stimuli in simulations with prior probability of A_1 indicated above the panels on the horizontal axis of panels A and B. Dashed lines indicate $w_j/4$. The dots represent the weights after 15,000 simulated trials, averages over 20 repetitions of the simulation. Simulations were performed using weights w_j with $\{w_1, w_2, \ldots, w_m\} = \{-1, -0.75, -0.5, -0.25, 0.25, 0.5, 0.75, 1.\}$. In the simulations, the exploration parameter was set to $\beta = 0$. (I, J) The activity of node y_1 during decision making in the models trained on a task where $P(A_1) = 0.8$. Each dot corresponds to a possible set of stimuli. Dashed lines indicate identity.

if we consider $n = 1$, then the condition, which the weights need to satisfy in the stochastic fixed point, given in equation 3.5, becomes

$$q^*_{1j} + q^*_{10} = \sigma \left(\log \left(\frac{P(A_1)}{P(A_2)} \right) + w_j \right). \tag{3.10}$$

Note that when m stimuli are used in the task, $j \in 1, \ldots, m$, so there are m equations that need to be satisfied, but there are $m + 1$ unknowns (q_{10}^* to q_{1m}^*), so there are multiple sets of weight values that satisfy the above condition.

In order for the model to learn the prior probabilities, the trials with a different number n of stimuli per trial had to be intermixed. Figure 3B shows that q_{i0} indeed converged to the vicinity of $P(A_i)$.

To gain a sense of how learning of the prior probabilities affects subsequent decision behavior, Figures 3I and 3J show the activity of the decision node y_1 (selective for the more likely action) when the trained model is presented with a particular set of stimuli. Each dot corresponds to a set of stimuli, and the actual probability of action A_1 being correct for a given set is reflected by the position along the horizontal axis. Since the posterior probability of action A_1 being correct was equal to the expected reward, and y_1 is the reward predicted by the model, a perfectly trained model should produce the activity on the identity line. Although model predictions are generally close to the true expected reward, there are systematic departures that are worth analyzing as similar misestimations of expected reward were observed by Soltani et al. (2016).

When the model was trained with a fixed number of stimuli per trial (as in the Soltani et al., 2016, study), the activity depended on the number of stimuli present on a given testing trial (indicated by color in Figure 3I). In particular, if only $n = 1$ stimulus was presented, the model underestimated the reward for choosing action A_1. This happened because the bias weight q_{10} underestimated the prior probability (see Figure 3A). For a larger number of stimuli n, the activity y_1 became closer to the expected reward. This happened because the model incorporated information about prior probabilities into learned weights, such that the weights for the more likely action were increased. Note in Figures 3C to 3E that the majority of blue points shift upward as the probability of action A_1 increases. Therefore, when more stimuli n were presented, these increased weights cumulated, raising the activity y_1. A similar dependence of expected reward on the number of presented stimuli has been observed in an analogous task by Soltani et al. (2016), and we will come back to it in section 4.

Figure 3J shows the activity of node y_1, when the model has been trained with the variable number of stimuli. Here, the activity y_1 was closer to the posterior probability of action A_1 being correct for trials with $n = 1$ and $n = 2$ stimuli than in Figure 3I, because the model has experienced such trials during training. To help understand why the prediction is not perfect, we need to analyze under what conditions the model is able to closely approximate the decision variable of equation 1.2.

Let us consider to what values the other weights q_{ij} converge when priors are unequal. In the simulation of Figure 3B, $q_{10}^* = P(A_1)$; hence, the condition of equation 3.5 that the weights need to satisfy becomes

$$P(A_1) + \sum_{j=1}^{n} q_{1s_j}^* = \sigma \left(\log \left(\frac{P(A_1)}{P(A_2)} \right) + w_{cum} \right). \tag{3.11}$$

The prior probability in equation 3.11 can be reexpressed using analysis analogous to equations 3.3 and 3.4:

$$P(A_1) = \sigma \left(\log \left(\frac{P(A_1)}{P(A_2)} \right) \right) \tag{3.12}$$

Combining the above two equations, we obtain the following condition the weights need to satisfy at the stochastic fixed point:

$$\sigma \left(\log \left(\frac{P(A_1)}{P(A_2)} \right) \right) + \sum_{j=1}^{n} q_{1s_j}^* = \sigma \left(\log \left(\frac{P(A_1)}{P(A_2)} \right) + w_{cum} \right). \tag{3.13}$$

If the prior probabilities are sufficiently close to 0.5 and w_{cum} is sufficiently close to 0 so that the sigmoid on the right-hand side of equation 3.13 can be approximated as in equation 3.7, then the weight $q_{ij}^* = \frac{w_j}{4}$ will approximately satisfy equation 3.13. Figures 3F and 3G show that the weights indeed converge in the vicinity of $q_{ij} \approx \frac{w_j}{4}$, for priors close to 0.5, but not in the case of the more extreme priors in Figure 3H. Equation 3.13 also implies that the higher the prior probability, the closer w_{cum} needs to be to 0 for the weights to converge to $q_{ij}^* \approx \frac{w_j}{4}$.

Let us now consider whether the model can incorporate learned priors into the decision variable described in equation 1.2. If the prior probabilities are sufficiently close to 0.5 and w_{cum} is sufficiently close to 0 so that we can approximate $q_{ij} \approx \frac{w_j}{4}$, then the activity of the unit selective for the first action is approximately proportional to the decision variable of equation 1.2:

$$y_1 = q_{10} + \sum_{j=1}^{n} q_{1s_j} \approx \frac{1}{4} \left(\log \left(\frac{P(A_1)}{P(A_2)} \right) + w_{cum} \right) + \frac{1}{2}. \tag{3.14}$$

When the conditions described in the previous paragraph are not closely satisfied, as in Figure 3H, then the accumulation of evidence will not be fully accurate and the expected reward will not be closely estimated, as seen in Figure 3J.

3.4 Properties of the Model. This section characterizes different aspects of learning in the model in different variants of the task.
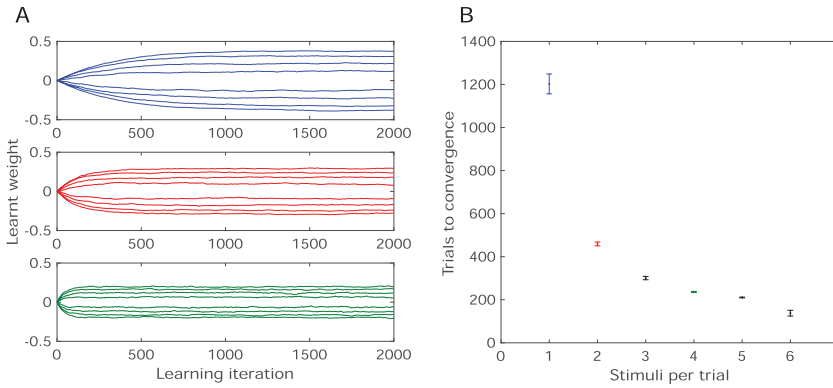
Figure 4: Speed of learning. (A) The values of weights q_{1j} as a function of learning iteration for different numbers $n = 1, 2, 4$ (blue, red, and green) stimuli presented at once. The weights are averaged over 100 repetitions of 5000 learning iterations with exploration parameter $\beta = 0$. (B) The number of trials to convergence as a function of the number of stimuli presented per trial. The number of trials to convergence was defined as the earliest trial number t in which the difference between the value of weight $q_{1,1}$ averaged over trials $([t - 200]_+, t + 200)$ of the 100 repetitions, and the average value on trials $(2000, 4000)$ was smaller than 0.01. For each n, the number of trials to convergence was computed 20 times, and its average is plotted together with error bars showing standard error.

3.4.1 *Speed of Learning.* Figure 4A compares how the weights changed during learning for the different numbers n of stimuli presented in the simulation shown in Figure 2A and reveals that the model converged faster when more stimuli were presented at a time. This effect is illustrated in Figure 4B, which shows the number of trials required for convergence as a function of the number of stimuli presented per trial. The model's weights are able to converge faster due to the fact that more information was presented at each trial and the final weights were less extreme when more stimuli were present at a time.

3.4.2 *Effect of Exploration Parameter.* For simplicity, in all simulations so far, we set the parameter β controlling how deterministic the choice is to $\beta = 0$, which corresponds to random action selection. In order to test learning in the model with more deterministic action selection, we performed simulations with different values for β. Figure 5A shows results when only $n = 1$ stimulus was presented per trial. We found that in cases of high β, which made the model choose only reward-promising actions, the neuron selective for action A_1 did not properly learn the weights of stimuli that predicted a low reward for this action. This happened because for such
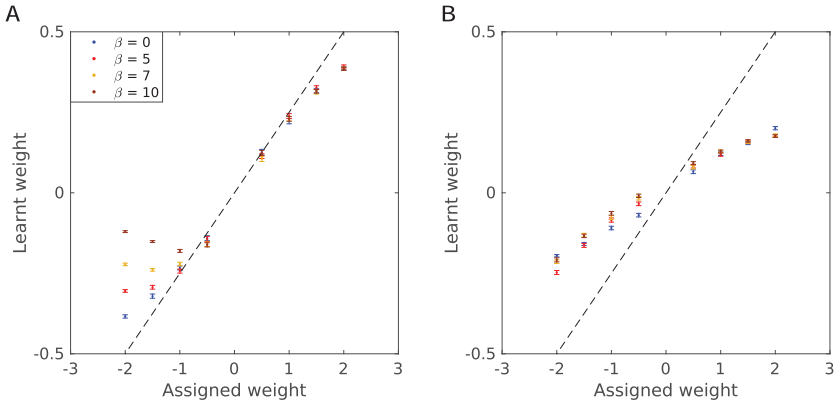
Figure 5: Effect of exploration parameter on learning. Results of 100 repetitions of 5000 learning iterations with (A) one stimulus presented per trial and (B) four stimuli presented per trial.

stimuli, action A_1 was rarely chosen. Nevertheless, we point out that the neuron selective for action A_2 did learn the weights of these stimuli (data not shown), so the network as a whole was able to preferentially select actions with higher expected value (note that the softmax equation 2.2 can be rewritten as $P_1 = \sigma(\beta(y_1 - y_2))$ and $P_2 = \sigma(\beta(y_2 - y_1))$, so the model makes a choice on the basis of the difference in activity of the two action-selective units).

The difficulty with learning weights for stimuli supporting the other action vanished as we performed simulations featuring $n = 4$ stimuli per trial (see Figure 5B). In this case, after extensive training, the model learned weights of all stimuli as it inevitably had to choose actions on the basis of four stimuli that may have included those predicting the nonchosen action.

3.4.3. Effect of Stimulus Frequency. It has been reported that humans weight stimuli that are unlikely to occur in learning tasks less strongly compared to more frequently appearing ones, as subjects are more uncertain about their influence and can update corresponding weights only infrequently (de Gardelle & Summerfield, 2011). To evaluate whether our model was able to reproduce this behavior, it was confronted with the same type of task as described above, but in this case, it featured pairs of stimuli with the same weights: $\{w_1, w_2, \ldots, w_8\} = \{-2, -2-1, -1, 1, 1, 2, 2\}$. For each pair of stimuli of the same weight, one was taken to appear more frequently ($P(s) = \frac{4}{20}$) than the other ($P(s) = \frac{1}{20}$). While computing the likelihoods of these stimuli, we used a formula analogous to equations 2.4 but scaled by the actual $P(s)$ rather than $\frac{1}{m}$.
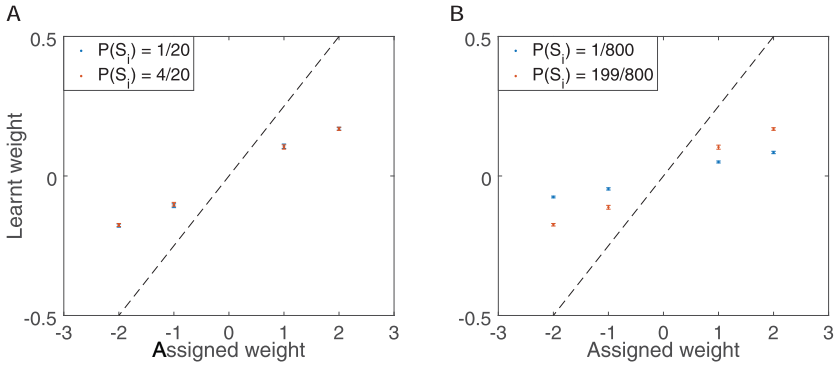
A

B

Figure 6: Effect of stimulus frequency on learning. (A) Simulations with a low-frequency difference ($\frac{4}{20}, \frac{1}{20}$) between stimuli of the same weight. (B) Simulations with a high-frequency difference ($\frac{199}{800}, \frac{1}{800}$) between stimuli of the same weight. One hundred repetitions of simulations with 5000 learning iterations were performed with $\beta = 0$.

Figure 6A shows that given enough training, the weights for stimuli with different frequency converge to similar values. In a second simulation, we set the differences in frequencies to be more extreme among stimuli with the same weight—$P(s_j) = \frac{199}{800}$ and $P(s_j) = \frac{1}{800}$ respectively—so that the infrequent stimuli are presented only rarely. Figure 6B illustrates that in this case, the weights for stimuli with lower frequency were closer to 0. This occurred because the infrequent stimuli were shown so rarely that their weights had not converged in the course of the simulation.

3.5 Simulation of Primate Learning Behavior. Yang and Shadlen (2007) conducted an experiment in which monkeys had to perform a probabilistic decision task similar to those in our simulations. In the experiment, they presented monkeys on each trial with $n = 4$ out of a total of $m = 10$ stimuli with the following WOEs: $\{w_1^{\log_{10}}, w_2^{\log_{10}}, \ldots, w_{10}^{\log_{10}}\} = \{-\infty, -0.9 - 0.7, -0.5, -0.3, 0.3, 0.5, 0.7, 0.9, \infty\}$. These WOEs were defined using $\log_{10}$ so are related to the WOE used so far according to

$$w_s^{\log_{10}} = \log_{10}\left(\frac{P(s|A_1)}{P(s|A_2)}\right) = \frac{\log\left(\frac{P(s|A_1)}{P(s|A_2)}\right)}{\log(10)} = \frac{w_s}{\log(10)} \tag{3.15}$$

In the Yang and Shadlen (2007) experiment, the stimuli presented on each trial were generated randomly with replacement, and then the probability
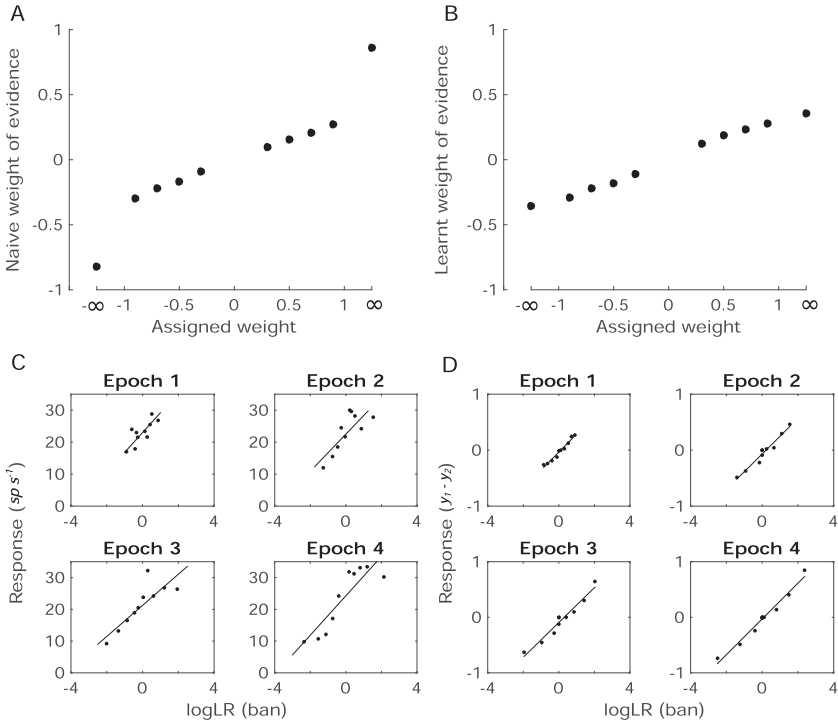
Figure 7: Model simulations reflect different characteristics of primate learning behavior. (A) Naive weight of evidence against assigned weights from the supplementary materials of Yang and Shadlen, 2007, replotted from Figure S3a. (B) Learned weight of evidence after 100 repetitions of 1000 learning iterations with $\beta = 0$. (C) Firing rates after presentation of the four stimuli within a trial in the experiment of Yang and Shadlen (2007), replotted from Figure 2c. (D) Difference in the activity of nodes selective for actions $y_1 - y_2$ of the model in different epochs of stimuli presentation.

of two targets being rewarded was computed from

$$P(r = 1 | A_1, s_1, s_2, s_3, s_4) = \frac{10^{\sum_{j=1}^{4} w_{s_j}}}{1 + 10^{\sum_{j=1}^{4} w_{s_j}}}, \tag{3.16}$$

$$P(r = 1 | A_2, s_1, s_2, s_3, s_4) = 1 - P(r = 1 | A_1, s_1, s_2, s_3, s_4). \tag{3.17}$$

Yang and Shadlen (2007) estimated WOE represented by the animals from behavioral data under different assumptions about the independence of the stimuli. Figure 7A replots the "naive" WOE estimated under the assumption

of stimuli being conditionally independent given action, which is typically assumed in the models of decision making, and used in equation 1.2. We found that our model simulated in the same paradigm was able to learn weights similar to the ones the two monkeys learned in their experiments (see Figure 7B). In order to compare our simulated final learned weights with the naive WOE, which Yang and Shadlen (2007) defined in their work, we defined the learned weight of evidence (LWOE) for our model data by considering the linear approximation:

$$LWOE_j = \frac{4\,q_{1j}}{\log(10)}.$$
(3.18)

Yang and Shadlen (2007) also recorded the activity of neurons in a decision-making area and observed it reflected integrated evidence for the action the neuron was selective for, as shown in Figure 7C. The four displays show the activity after the presentation of four consecutive stimuli. Within each display, the trials were sorted by the cumulative WOE of stimuli presented so far and binned into 10 groups. Each dot shows the average firing rate for trials in a given group.

Figure 7D shows analogous analysis of difference in activity in the action units in the model $y_1 - y_2$ after the presentation of consecutive stimuli. The trials were binned, excluding the trials in which stimuli with infinite WOE were present. We observed that the nodes in the model had activity proportional to cumulative WOE, similar to the primate data by Yang and Shadlen (2007).

4 Discussion

This letter has analyzed the relationship between computational accounts of learning and decision making based on reinforcement learning (RL) and the sequential probability ratio test (SPRT). We demonstrated that synaptic weights learned by RL rules in a certain class of tasks are proportional to WOEs, and hence allow information to be integrated from multiple cues to form a decision. Simulations of the model in the task of Yang and Shadlen (2007) replicated the key features of animal behavior and neural activity. In this section, we relate the model we have presented to other models and experimental data, and we discuss further experimental predictions.

4.1 Relationship to the Soltani and Wang Model.
In a closely related study, Soltani and Wang (2010) proposed a model that can also learn weights of synaptic connections allowing probabilistic inference and can also replicate the observations of Yang and Shadlen (2007). We briefly review their model, discuss in what ways it differs from the model we have proposed in

this letter, and suggest experiments that can differentiate between the two models.

The Soltani and Wang (2010) model describes a network that includes neurons selective for different stimuli and neurons selective for different actions. It assumes that the weights of connections between these neurons are binary, so that an individual synapse can have a weight of zero or one. After each trial, the weights between the neurons selective for presented stimuli and chosen action are modified according to the reward received. If a reward was received, the synapses equal to zero may increase with probability α_+, while if no reward was received, the synapses equal to one may decrease with probability α_-. We denote the average value of connections between neurons selective for stimulus s and neurons selective for action i by c_{is}, so:

$$\Delta c_{is} = \begin{cases} \alpha_+(1 - c_{is}), & \text{if the reward present} \\ -\alpha_- c_{is}, & \text{if no reward given} \end{cases}. \tag{4.1}$$

In equation 4.1, the weight increase after the reward depends on the fraction of inactive synapses (and analogously following the lack of reward). Despite seemingly different learning rules, the weights in the two models converge to closely related values. In particular, let us first consider the case when $\alpha_+ = \alpha_-$, the prior probabilities are equal, and only $n = 1$ stimulus is presented per trial. Under these conditions on trials when stimulus s is presented and action i is chosen, the expected change in the corresponding weights is

$$E\left(\Delta c_{is}\right) = \alpha(1 - c_{is})P(A_i|s) - \alpha c_{is}(1 - P(A_i|s)). \tag{4.2}$$

To find the value of c_{ij} at the stochastic fixed point, we set $E\left(\Delta c_{is}\right) = 0$ in equation 4.2 and find that $c_{ij}^* = P(A_i|s)$, which together with equations 3.1 and 3.2 implies the following relationship between the weights in the two models $c_{ij}^* = q_{ij}^* + \frac{1}{2}$. A linear relationship between c_{ij}^* and q_{ij}^* seems to also hold for $n > 1$, as can be seen by comparing the simulations of the two models in the Yang and Shadlen (2007) task (see Figure 7 in this letter and Figure 2b in Soltani and Wang, 2010).

Despite the similarities, the models differ in two key aspects: the plasticity rule and the presence of the bias node that is critical for learning prior probabilities. We now review these two differences, compare the models with experimental data, and suggest further experiments that can differentiate between the models.

In the Soltani and Wang (2010) model, the weight modification is modulated by reward, while in the model proposed here, it is modulated by reward prediction error. The model proposed here aims at capturing learning in the basal ganglia and assumes that such modulation of plasticity is

mediated by the neuromodulator dopamine, which is known to influence cortico-striatal plasticity (Reynolds et al., 2001; Shen et al., 2008), and encode the reward prediction error during learning tasks (Schultz et al., 1997; Fiorillo et al., 2003). By contrast, the Soltani and Wang (2010) model aimed at capturing learning in the cortex, where the effects of reward on synaptic plasticity are not as well understood. Importantly, the model proposed here uses the same synaptic plasticity rule, which is also known to well capture learning about reward magnitudes in reinforcement learning tasks. Therefore, we suggest that there is no need for the brain to have specialized plasticity rules dedicated to support probabilistic reasoning, as the standard synaptic mechanisms that learn expected rewards can fulfill this function.

The two models make differential predictions on how the learning about WEO should interact with reward. Consider an experiment in which some stimuli are presented on trials on which a reward of $r = 2$ is given for correct choices, while other stimuli are presented on trials where $r = 1$ is given for correct responses. Subsequently, on critical test trials, participants need to make a choice on the basis of stimuli from both groups presented together. In the model proposed here, the learned weights q_{ij} are proportional to the expected reward; thus, it would predict that the participants would be more influenced by the stimuli from the first group. Such increased influence is not predicted by the Soltani and Wang (2010) model, where the reward magnitude does not affect c_{ij}.

Recently Soltani et al. (2016) proposed an extended version of the model in which the weight changes depend on average reward rate $\bar{r}_i$ for selecting action i,

$$\Delta c_{is} = \begin{cases} \alpha_+ (1 - c_{is}) \times 2\sigma\left(\frac{\bar{r}_i - 0.5}{d}\right), & \text{if the reward present} \\ -\alpha_- c_{is}, & \text{if no reward given} \end{cases}, \qquad (4.3)$$

where d is an additional scaling parameter. In this extended model, the weight change depends on the overall average reward connected with an action, while in the Rescorla-Wagner rule, the weight change depends on the reward for a particular action after presentation of a particular stimulus. Consequently this extended model would still make the same prediction as the original Soltani and Wang (2010) model in the experiment suggested above if it is ensured that the average reward for both actions is the same. For example, consider a task in which during training, four stimuli A, B, C, D are interleaved. For stimuli A and B, the reward for the correct choice is $r = 2$, while for C and D, it is $r = 1$, and for stimuli A and C, the more rewarded response is left, while for B and D it is right. Since the average reward is the same for the left and right actions, the reward magnitude does not a effect weights c_{ij} of the extended Soltani et al. (2016) model, while it affects q_{ij} learned with the Rescorla-Wagner rule. Therefore, the model proposed here predicts that stimuli A and B will have higher magnitudes of learned

weights than C and D, while the Soltani et al. (2016) model predicts equal weight magnitudes.

It is also worth comparing the biological plausibility of the Rescorla-Wagner rule and the rule of equation 4.3. A nice property of equation 4.3 is that the change of a particular synaptic weight depends only on the value of this weight, not on other weights in the network. By contrast, in the Rescola-Wagner rule, the change in q_{ij} depends on the reward prediction error, which is a function of y_i, which in turn depends on other weights in the network. Nevertheless, we described in section 2.1 how this problem can be overcome in the basal ganglia circuit. Recall that the model assumes that the reward prediction error is computed by dopaminergic neurons that receive input from striatal neurons computing y_i. Consequently, the Rescorla-Wagner rule requires the eligible synapse to have information on only a single quantity: the reward prediction error that could be brought by a single neuromodulator, dopamine. By contrast, the rule of equation 4.3 requires the synapse to have information on two quantities: the presence of the reward on the current trial and the average reward rate. Thus, to implement such a rule, two separate neuromodulators would need to encode these quantities, and it is unclear which of the known neuromodulators could play this role.

The second difference between the models is that the one proposed here includes bias weights q_{i0} that allow it to learn about prior probabilities of the responses under certain conditions, while the Soltani and Wang (2010) model does not include the bias node and hence is unable to represent prior probabilities separately from likelihoods.

To test whether humans are able to learn prior probabilities separately from WEO, Soltani et al. (2016) trained participants with a fixed number ($n = 4$) of stimuli presented per trial. Then they presented the participants with one, two, or four stimuli and asked them to estimate how likely the two responses are to be correct. They found a pattern similar to that in Figure 3I (see Figure 2d in Soltani et al., 2016): participants underestimated the probability of the more likely option for $n = 1$ stimuli. Soltani et al. (2016) pointed out that these data can be explained only by the model that did not learn prior probabilities separately from WEO. Simulations in section 3.3 show that our model also produces this pattern of behavior, because our model also did not learn prior probabilities when a fixed number ($n = 4$) of stimuli was presented per trial during training.

The model presented here predicts that when the number of stimuli presented during learning is intermixed, the networks in the brain should be able to learn prior probabilities of responses. To test this prediction, one could modify an experiment from Soltani et al. (2016) such that trials with different n are intermixed and the model proposed here predicts that the participants then would be able to learn the prior probabilities and no longer underestimate the probability of a more likely option for small n (i.e., produce the pattern illustrated in Figure 3J).

It would be interesting to investigate whether a modified version of the Soltani and Wang (2010) model, including the bias node, could also learn the prior probabilities if the trials with different n are intermixed during training, but not when n is fixed during training.

The model presented in this letter describes learning in the basal ganglia, while the Soltani and Wang (2010) model focuses on learning in the neocortex. It is likely that both structures are involved in learning in tasks requiring evidence accumulation, so it is also possible that the two models describe complementary contributions of basal ganglia and cortex to probabilistic decision making.

4.2 Relationship to Other Models. A handful of other studies have linked the RL to the framework provided by the SPRT, or related sequential sampling models. Law and Gold (2009) have also used a standard RL model with an architecture and learning rule very similar to these considered here to capture learning and decision making in a motion discrimination task. Their model was able to learn weights allowing accumulation of information and reproduced many aspects of neural activity during motion discrimination tasks. Here we show that a similar model can also be used to describe decision tasks with discrete stimuli, and we explicitly demonstrate that in these tasks, the learned synaptic weights are approximately proportional to WOEs.

Two studies described models of the basal ganglia circuit that can learn probabilistic quantities, allowing the circuit to implement Bayesian decision making in a fashion equivalent to that described in equation 1.2 (Berthet et al., 2012; Coulthard et al., 2012). However, both of these models assume complex rules for the plasticity of cortico-striatal synapses, and it is not clear if such rules can be implemented by biological synapses. By contrast, here we show that weights allowing integration of information during decision making can also arise from the very simple plasticity rule of Rescorla and Wagner (1972).

In this letter, we have focused on decision making between two options, but it would be interesting to generalize our approach to choices with multiple alternatives. With more than two options, it is no longer possible to define a simple decision variable as in equation 1.2. Nevertheless, it has been proposed that the basal ganglia can compute posterior probabilities of actions given presented stimuli (Bogacz & Gurney, 2007; Bogacz & Larsen, 2011). In order to perform such computation, the neurons selective for an action in this model need to receive input proportional to the log likelihood of stimuli given the action (Bogacz & Gurney, 2007), or WOE in the case of a choice between two alternatives (Lepora & Gurney, 2012). Thus, for a choice between two alternatives, the cortico-striatal weights proportional to WOEs would allow this Bayesian model of basal ganglia to compute the posterior probabilities of actions. Future research may wish to investigate whether

the cortico-striatal weights learned with the Rescorla-Wagner rule can allow the model of the basal ganglia to approximate the posterior probabilities of actions for the choice between multiple alternatives.

4.3 Relationship to Experimental Data. The simulation of the model in the task of Yang and Shadlen (2007) showed that the model learned similar WOEs as the animals for cues with finite WOE. For cues with infinite WOE, the weights learned by the model were more dampened than those learned by the animals (see Figures 7A and 7B). Nevertheless, note that the animals also dampened the weights of these stimuli (i.e., the WOEs estimated by animals are not infinite). The difference in the extent to which these weights were dampened could arise from the fact that our model captured only model-free RL, while animals could have employed both model-free and model-based RL systems during their choices (Daw, Niv, & Dayan, 2005), and the model-based system could have learned simple deterministic rules for these stimuli (e.g., choose A_1 whenever stimulus 10 is presented). During decision making with such stimuli, the final choice could have been based on information brought by both the model-based and model-free system, resulting in a high but not fully deterministic influence of these stimuli on choice.

The simulations also showed that the model could replicate key features of neural responses to successive stimuli, the neural activity being proportional to the accumulated WOE of stimuli seen so far. Nevertheless, this relationship was more linear in the last "Epoch 4" in our model than in the experimental data, where it appeared more sigmoid (see Figures 7C and 7D). This difference may arise from the fact that after seeing the last stimulus, the animals knew that they had all available information, and the neural activity could have started to reflect the choice rather than the decision variable.

The neural activity in the experiment of Yang and Shadlen (2007) was recorded from the lateral intraparietal cortex, while our model described the activity in the striatum. Nevertheless, it has been observed that the neural activity in striatum also encodes information accumulated during decision making (Ding & Gold, 2010). This similarity in activity between decision-related cortical regions and striatum may arise from the prominent feedback connections from basal ganglia back to cortex via the thalamus (Alexander, DeLond, & Strick, 1986) and the fact that in highly practiced tasks, the stimulus-response mapping learned in the striatum becomes consolidated in the cortex (Ashby, Ennis, & Spiering, 2007).

The model presented here required fewer trials to converge when multiple stimuli were present per trial (see Figure 4). It seems unlikely that humans and animals would show such behavior, as humans often learn faster in tasks that start with training with a single stimulus per trial (Gould, Nobre, Wyart, & Rushworth, 2012). Our simulations (not shown here) indicate

that slower learning with multiple stimuli occurs in a modified version of the model in which on each trial, the weights q_{ij} are updated for only a single stimulus j (randomly chosen in the simulations). It is also possible that such slower learning arises because subjects have difficulty focusing on several stimuli presented to them and evolution has optimized perception in a way that only stimuli crucial to decision making are attended (Summerfield & Tsetsos, 2015).

4.4 Other Experimental Predictions. In addition to the predictions described in section 4.1, the model described in this letter makes a few more predictions. In the proposed model, the estimated WOEs often depend on the number n of stimuli presented within a trial (see Figure 2A; a similar prediction is also made by the Soltani and Wang, 2010, model). In particular, the model predicts that the estimated WOEs are closer to zero if the participants learn them in a task with multiple stimuli presented in a trial. This prediction could be tested in an experiment in which participants learn WOEs for one set of stimuli with $n = 4$ and WOEs for another set with $n = 1$, and then make decisions on the basis of multiple stimuli from both sets. The model predicts that participants would give more weight to stimuli learned with $n = 1$.

Experiments with human subjects by de Gardelle and Summerfield (2011) featuring shapes colored by stochastically drawn values of a two-color gradient showed that humans performed averaging among presented color values when they were asked to decide which of the two colors was predominant. It was also observed that outliers (extreme color values that appeared less frequently) were downweighted by the subjects, even though they should have had a strong influence on decision outcome (de Gardelle & Summerfield, 2011). We performed simulations featuring stimuli with the same weights but different frequencies of occurring. In cases where frequency of commonly and uncommonly occurring stimuli was sufficiently different, we were able to observe a downweighting by a constant factor. It would be interesting to perform experiments with human subjects in a similar scenario in order to see whether downweighting occurs only if the stimuli are sufficiently infrequent.

In summary, in this letter, we have shown that the same learning rule that allows estimating expected rewards associated with stimuli and actions can approximate WOEs in a class of tasks with binary rewards. Such WOEs can be efficiently integrated across different stimuli during decision making.

Acknowledgments

References

Alexander, G. E., DeLond, M. R., & Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience*, *9*, 357–381.

Ashby, F. G., Ennis, G. M., & Spiering, B. J. (2007). A neurobiological theory of automaticity in perceptual categorization. *Psychological Review*, *114*, 632–656.

Berthet, P., Hellgren-Kotaleski, J., & Lansner, A. (2012). Action selection performance of a reconfigurable basal ganglia inspired model with Hebbian-Bayesian Go-NoGo connectivity. *Frontiers in Behavioural Neuroscience*, *6*, 65.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*, 700–765.

Bogacz, R., & Gurney, K. (2007). The basal ganglia and cortex implement optimal decision making between alternative actions. *Neural Computation*, *19*, 442–477.

Bogacz, R., & Larsen, T. (2011). Integration of reinforcement learning and optimal decision-making theories of the basal ganglia. *Neural Computation*, *23*, 817–851.

Collins, A. G. E., & Frank, M. J. (2014). Opponent actor learning (OpAL): Modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. *Psychological Review*, *121*, 337–366.

Coulthard, E., Bogacz, R., Javed, S., Mooney, L. K., Murphy, G., Keeley, S., & Whone, A. L. (2012). Distinct roles of dopamine and subthalamic nucleus in learning and probabilistic decision making. *Brain*, *135*, 3721–3734.

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

de Gardelle, V., & Summerfield, C. (2011). Robust averaging during perceptual judgment. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 13341–13346.

Ding, L., & Gold, J. I. (2010). Caudate encodes multiple computations for perceptual decisions. *Journal of Neuroscience*, *30*, 15747–15759.

Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, *299*, 1898–1902.

Frank, M. J., Seeberger, L. C., & O'Reilly, R. C. (2004) By carrot or by stick: Cognitive reinforcement learning in Parkinsonism. *Science*, *306*, 1940–1943.

Gluck, M. A., & Bower, G. H. (1998). Evaluating an adaptive network model of human learning. *Journal of Memory and Language*, *27*, 166–195.

Gold, J. I., & Shadlen, M. N. (2001). Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Sciences*, *5*, 10–16.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, *30*, 535–574.

Gould, I. C., Nobre, A. C., Wyart, V., & Rushworth, M. F. S. (2012). Effects of decision variables and intraparietal stimulation on sensorimotor oscillatory activity in the human brain. *Journal of Neuroscience*, *32*, 13805–13818.

Knowlton, B. J., Mangels, J. A., & Squire, L. R. (1996). A neostriatal habit learning system in humans. *Science*, *273*, 1399–1402.

Kravitz, A. V., Freeze, B. S., Parker, P. R., Kay, K., Thwin, M. T., Deisseroth, K., & Kreitzer, A. C. (2010). Regulation of Parkinsonian motor behaviours by optogenetic control of basal ganglia circuitry. *Nature*, *446*, 622–626.

Law, C., & Gold, J. I. (2009). Reinforcement learning can account for associative and perceptual learning on a visual-decision task. *Nature Neuroscience*, *12*, 655–663.

Lepora, N. F., & Gurney, K. N. (2012). The basal ganglia optimize decision making over general perceptual hypotheses. *Neural Computation*, *24*, 2924–2945.

Mikhael, J. G., & Bogacz, R. (2016). Learning reward uncertainty in the basal ganglia. *PLoS Computational Biology*, *12*, e1005062.

Philiastides, M. G., Biele, G., & Heekeren, H. (2010). A mechanistic account of value computation in the human brain. *PNAS*, *107*, 9430–9435

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement or nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). East Norwalk, CT: Appleton-Century-Crofts.

Reynolds, J. N., Hyland, B. I., & Wickens, J. R. (2001). A cellular mechanism of reward-related learning. *Nature*, *413*, 67–70.

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

Shen, W., Flajolet, M., Greengard, P., & Surmeier, D. J. (2008). Dichotomous dopaminergic control of striatal synaptic plasticity. *Science*, *321*, 848–851.

Soltani, A., Khorsand, P., Guo, C., Farashahi, S., & Liu, J. (2016). Neural substrates of cognitive biases during probabilistic inference. *Nature Communications*, *7*, 11393.

Soltani, A., & Wang, X. J. (2010). Synaptic computation underlying probabilistic inference. *Nature Neuroscience*, *13*, 112–119.

Summerfield, C., & Tsetsos, K. (2015). Do humans make good decisions? *Trends in Cognitive Sciences*, *19*, 27–34.

Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning.* Cambridge, MA: MIT Press.

Watabe-Uchida, M., Zhu, L., Ogawa, S. K., Vamanrao, A., & Uchida, N. (2012). Whole-brain mapping of direct inputs to midbrain dopamine neurons. *Neuron*, *74*, 858–873.

Yang, T., & Shadlen, M. N. (2007). Probabilistic reasoning by neurons. *Nature*, *447*, 1075–1080.