Decision Threshold Learning in the Basal Ganglia for Multiple Alternatives

Thom Griffith
Department of Engineering Mathematics
University of Bristol
thom.griffith@bristol.ac.uk

Sophie-Anne Baker
Department of Engineering Mathematics
University of Bristol
sophie.baker@bristol.ac.uk

Nathan F. Lepora
Department of Engineering Mathematics
University of Bristol
n.lepora@bristol.ac.uk

Abstract

In recent years, researchers have integrated the, historically separate, reinforcement learning (RL) and evidence-accumulation-to-bound approaches to decision modelling. A particular outcome of these efforts has been the RL-DDM, a model which combines value-learning through reinforcement with a diffusion decision model (DDM). While the RL-DDM is a conceptually elegant extension of the original DDM, it faces a similar problem to the DDM in that it does not scale well to decisions with more than two options. Furthermore, in its current form, the RL-DDM lacks flexibility when it comes to adapting to rapid, context-cued changes in the reward environment. The question of how to best extend combined RL and DDM models so they can handle multiple choices remains open. Moreover, it is currently unclear how these algorithmic solutions should map to neuro-physical processes in the brain, particularly in relation to so-called Go/No-go-type models of decision-making in the basal ganglia. Here, we propose a solution that addresses these issues by combining a previously proposed decision model, based on the multi-choice sequential probability ratio test (MSPRT), with a dual-pathway model of decision-threshold learning in the basal ganglia region of the brain. Our model learns decision thresholds to optimize the trade-off between time cost and the cost of errors and so efficiently allocates the amount of time for decision deliberation. In addition, the model is context-dependent and hence flexible to changes to the speed-accuracy trade-off (SAT) in the environment. Furthermore, the model reproduces the magnitude effect, a phenomenon seen experimentally in value-based decisions, and is agnostic to the types of evidence and so can be used on perceptual decisions, value-based decisions as well as other types of modelled evidence. The broader significance of the model is that it contributes to the active research area of how learning systems interact, by linking the previously separate models of RL-DDM to dopaminergic models of motivation and risk taking in the basal ganglia, as well as scaling to multiple alternatives.

1 Introduction

Decision-making models within the cognitive sciences have historically fallen into two distinct frameworks: sequential sampling models (SSM) and reinforcement learning (RL) models (Miletić et al., 2020; Collins, 2021). Sequential sampling models employ parallel accumulators to gather sampled evidence for each option and act as a dynamic choice rule that accounts for choice probabilities as well as, crucially, response times (Ratcliff, 1978; Gold and Shadlen, 2002, 2007; Ratcliff et al., 2016). On the other hand, reinforcement learning models focus on the learning and encoding of choice values (Collins and Frank, 2014; Moeller et al., 2022; Frank et al., 2007) and often neglect the deliberative process. These types of RL models are therefore only able to predict choice probability or, alternatively, use response time models that are dependent on the intrinsic choice values and do not explicitly factor in a time cost.

Sequential sampling models have been used to gain insight into neurophysical processes underlying decision-making (Gold and Shadlen, 2002, 2007; Churchland et al., 2008; Cisek and Kalaska, 2010) and in perceptual-decision making in humans and animals (Hanks and Summerfield, 2017). The drift diffusion model (DDM) is a classic example of an SSM (Ratcliff, 1978; Ratcliff et al., 2016), which has primarily been applied to binary choices although there are also extensions to multiple alternatives (Krajbich and Rangel, 2011). The DDM can be viewed as a specific implementation of the sequential probability ratio test (SPRT), a statistically optimal test between two hypotheses (Wald, 1945). The binary limitation of the original DDM has prompted the development of other models capable of handling multiple choices. Notable examples include the multi-alternative sequential probability ratio test (MSPRT) (Baum and Veeravalli, 1994; Veeravalli and Baum, 1995) an extension to the SPRT, which uses a Bayesian framework, and is the most sample efficient statistical test in the limit of vanishing error rate for multiple hypotheses. Other examples of multiple choice decision models include those that employ race models (Heathcote, 2022; van Ravenzwaaij et al., 2020) and parallel accumulators with global inhibition (McMillen and Holmes, 2006; Kriener et al., 2020). One critical element of all these models is the decision threshold on the evidence which influences the balance of speed and accuracy (Heitz, 2014) suggesting a method by which the cost of time in sampling evidence for a decision is a controllable factor (Drugowitsch et al., 2012; Khodadadi et al., 2017).

Recent efforts have aimed to bridge the gap between RL and evidence accumulation approaches (for reviews see Miletić et al. (2020) and Collins (2021)) resulting in hybrid models that use SSMs as the decision model but allow the parameters of the SSM (such as accumulation rate (Frank et al., 2015; Sewell et al., 2019) and decision threshold (Lepora, 2016; Khodadadi et al., 2017)) to adapt through reinforcement. One of the outcomes of these efforts has been a unified model for binary decisions, the RL-DDM, which combines value-learning with a diffusion decision model and seems to describe human binary choice data well (Pedersen et al., 2017; Fontanesi et al., 2019a,b; Miletić et al., 2021). The RL-DDM is a conceptually elegant extension to the original DDM decision model. However, it encounters similar problems as the DDM, such as poor scalability to decisions involving more than two options and inflexibility to context-dependent changes in the reward environment. Updated versions of the model should therefore account for multiple-choice decisions and also reflect changes to task demands in their outputs. For instance linear ballistic accumulator (LBA) modelling of a perceptual decision-making task has shown that participants adjust their decision thresholds in response to cued instructions manipulating the speed accuracy trade-off (SAT) (Forstmann et al., 2010).

Researchers using hybrid models such as the RL-DDM also contend with the challenge of connecting these algorithmic solutions to the biological implementation level, particularly how they correspond to the so-called Go/No-go decision models, which capture the supposed activity of signalling pathways in the basal ganglia – a region of the brain important for perceptual decision-making (Ding and Gold, 2010, 2013) and value-based decision-making (Frank, 2006; Frank and Claus, 2006; Cisek, 2007). The basal ganglia Go/No-go model has also been proposed to implement decision thresholds by multiplicative gain on choice evidence, the Go pathway effectively reducing the threshold on evidence for a decision and the No-go pathway effectively increasing the threshold (Lo and Wang, 2006; Lepora and Gurney, 2012). This model tallies with the proposal that striatal levels of dopamine control the effective decision threshold due to the complementary effect dopamine concentration has on Go and No-go pathway activation (Chakroun et al., 2023; Thura and Cisek, 2017).

In this study, our goal is to integrate SSM and RL approaches to modelling decisions with more than two options, solving the multiple choice issue in the RL-DDM. We also aim to align our approach with current understanding of the basal ganglia's role in the deliberative process. For our proposed solution we combine a dual-pathway model of decision-threshold learning through reinforcement with a multi-choice decision model that further unifies the previously separate models of decision making in the basal ganglia and RL-DDM type models. The RL model successfully learns to find decision thresholds that minimize the time cost relative to costs from errors. The model is also context-dependent and hence flexible to changes to the optimal speed-accuracy trade-off in the environment. The context switching can therefore take place over faster time scales without forgetting of previously learned SATs. In addition, the model reproduces the magnitude effect, a phenomenon observed experimentally during value-based decision tasks. Furthermore, our model is agnostic to the types of evidence and so can be used on perceptual decisions and value-based decisions, and could potentially be used for other types of modelled evidence including model-based imaginative deliberative planning or episodic recall. The broader significance of the model is to contribute to the active research area of how learning systems interact by linking the previously separate models of RL-DDM to dopaminergic models of motivation and risk taking in the basal ganglia in addition to scaling beyond binary choice decisions.

2 Model and Methods

2.1 Problem definition

We assume that the primary role of the basal ganglia in decision-making is to identify the predicted best course of action, and to do so in a way that balances the expected cost of errors with the expected cost of the decision time (Redgrave et al., 1999; Gurney et al., 2001). The basal ganglia inputs are assumed to be a time-series stream of action-salience values that are received in parallel and that the role of the basal ganglia is to select the index of the stream with the highest expected value (Cisek, 2007; Cisek and Kalaska, 2010).

Specifically, we assume N input channels to the basal ganglia, each carrying a time series of action-salience values. These values are assumed to be independent and identically distributed (i.i.d.) samples from N distributions with means, $b_1, b_2, \ldots, b_N$, where $b_1 > b_2 \geq \ldots \geq b_N$ and the max-next distance of the means is $b_1 - b_2 = \Delta$. The complete observed action-salience sample at time t is an N-dimensional vector, denoted, $\mathbf{x}(t) = x_1(t), \ldots, x_n(t)$. The problem is to identify the time series with the largest mean; *i.e.*, the channel with mean b_1, from the samples alone, in a way that optimizes the speed-accuracy trade-off.

2.2 Model of evidence accumulation: MSPRT

For the evidence accumulation model we use a previously proposed model of decision-making in the basal ganglia that describes a plausible neural implementation of the MSPRT algorithm (Bogacz and Gurney, 2007; Lepora and Gurney, 2012). Briefly, the MSPRT uses a Bayesian framework to make decisions between multiple alternative hypotheses using a rule that says if the calculated posterior probability of the i^{th} hypothesis exceeds some threshold, Θ_i, then that hypothesis, H_i, is selected as the most likely one. To implement the goal of selecting the action channel with the highest mean action salience we construct hypotheses by imposing the assumptions that: i) all the channel data streams are Gaussian distributed with the same variance, and ii) all channels have the same mean μ_1, except for one channel with mean $\mu_2 > \mu_1$ (see Figure 1a).

Explicitly, we assume H_i is the hypothesis that observations from the i^{th} channel, x_i, are sampled from a Gaussian distribution with mean μ_2 and standard deviation, σ, and observations from all other channels, $x_{j\neq i}$, are sampled from a Gaussian with mean μ_1 and standard deviation, σ. This implies that H_i is the hypothesis that the i^{th} channel represents the action with greatest salience, and so selecting H_i implements the decision.

Starting from this construction, it can be shown that the log likelihood of the hypothesis, H_i, conditioned on the input history, $\mathbf{x}(1), \ldots, \mathbf{x}(T)$, is given by a combination of a simple signal integrator and a global inhibitory signal that depends on the activity of all channel accumulators, (Bogacz and Gurney, 2007):

$$\log p(H_i|\mathbf{x}(1), \ldots, \mathbf{x}(T)) = y_i(T) - \log \sum_{j=1}^{N} \exp y_j(T), \tag{1}$$

$$y_i(T) = \sum_{t=1}^{T} g x_i(t), \tag{2}$$

where $g = (\mu_2 - \mu_1)/\sigma^2$ is a scaling constant for the observations that depends on the separation and width of the hypothesised Gaussian distributions.

The above result is derived by first using Bayes' theorem and assuming flat priors over all hypotheses. Taking logs and substituting in the equations for the hypothesised Gaussian distributions results in equations (1) and (2) (demonstrated in Appendix A of this paper for completeness).

Although the original MSPRT decision rule employs a threshold on the posterior, we can equivalently apply a log-transformed threshold to the log of the posterior, $\theta_i = \log(\Theta_i)$,

$$y_i(T) - \log \sum_{j=1}^{N} \exp y_j(T) > \theta_i. \tag{3}$$

This rule implements the MSPRT exactly if the action salience distributions are Gaussian with equal widths, and provided $b_1 > b_2$ with $b_2 = b_3 = \ldots = b_N$. If the assumptions about the distributions are not true, then the model offers an approximation of the MSPRT that instead implements a race model between accumulators with global non-linear inhibition, which is a common schema for other proposed multi-alternative decision models with plausible neural implementations. How far the approximation deviates from true MSPRT decisions will depend on how far the underlying distributions that generate the evidence samples deviate from the equal-width Gaussian assumption that is implicit to the above derivation.

2.3 Mapping to Basal Ganglia anatomy

Now we turn to the biological plausibility of the model, in particular how the action selection and threshold-learning components of the algorithm may map on to activity within the cortico-basal ganglia-thalamic circuit. Representing the evidence in the decision model as a log posterior means that the quantity on the left of the inequality in Eq. (3) is always negative (since the posterior is a probability and taking logarithm of quantities less than 1 results in a negative value). Since a negative neural firing rate is not possible, Bogacz and Gurney (2007) proposed a model for the activity of an output channel in the basal ganglia which instead represented the negative of the log posterior for the corresponding hypothesis,

$$O_i(T) = -y_i(T) + \log \sum_{j=1}^{K} \exp y_j(T) < \theta_i(T), \tag{4}$$

where $O_i(T)$ denotes the activity of the i^{th} output channel.

Note that the threshold is now approached from above, and action selection is initiated when activity drops below a particular threshold. This is consonant with what we know about the connectivity of the basal ganglia where distinct parallel circuits in the basal ganglia govern cortical activity through a process of focused disinhibition. That is, action selection is initiated by a reduction in activity in the winning channel, thereby releasing its target from its inhibitory effect (Chevalier et al., 1985; Chevalier and Deniau, 1990).

2.3.1 Go and No-Go pathway activity as positive and negative threshold shifts

We introduce one further change to the functional form of the decision rule to increase the biological plausibility of the model that highlights the possible functional mapping of the decision rule onto the regions of the basal ganglia. We introduce a decision threshold shift term $\Delta\theta_i = \Delta\theta_i^{\text{N}} - \Delta\theta_i^{\text{G}}$,

$$O_i(T) = -y_i(T) + \log \sum_{j=1}^{K} \exp y_j(T) + \Delta\theta_i < \theta^*, \tag{5}$$

where $\Delta\theta_i^{\text{G}}$ and $\Delta\theta_i^{\text{N}}$ terms represent a positive or negative shift to the decision threshold that is due to the activity of striatal medium spiny neurons (MSN) populations in the so-called Go and No-Go pathways, respectively. In the inequality given in Eq. 5, the threshold shift is implemented as an additive gain on the normalized accumulator for that channel. The quantity on the left of the inequality represents the fully transformed action salience signal, which is then compared with a fixed threshold in cortex. The motivation for this form of the rule is so it corresponds with experimental observations, made during monkey perceptual-decision study, of ramping activity from LIP neurons during a random dot motion task, in which the decision process terminated when the firing rate of the associated neurons reached a common firing rate across task conditions implying a common threshold on the decision variable across task conditions (Gold and Shadlen, 2007).

The $\Delta\theta_i^{\text{G}}$ term represents a positive shift in the decision threshold, reducing the amount of accumulated evidence required for action initiation (remembering that the threshold is approached from above) and so a positive threshold shift increases the tendency to action, hence the Go signal; whereas a negative shift in the threshold (communicated by activity of MSN in the No-Go pathway and represented by the $\Delta\theta_i^{\text{N}}$ term) decreases the tendency towards action since this increases the gap between the accumulated evidence and the effective decision threshold.

A crucial point is that this tendency to action is manipulated by positive and negative shifts of the decision threshold, and that this affects the decision response time and accuracy.

Finally, as previously suggested in Lepora and Gurney (2012), an explicit mapping of the terms in the decision rule to regions of the basal ganglia is given in Figure 1). Briefly, the idea is that the basal ganglia performs two functions: firstly, it contributes an inhibitory ‘competition’ signal to the individual channels in the cortico-basal ganglia-thalamo-cortical loops that depends on the global activity, thereby preventing hasty decisions in the face of conflicting evidence. In the proposed mapping this takes place in the STN region, which has been cited as an important region for inhibiting motor decisions during phases of uncertainty (Frank, 2007; Zaghloul et al., 2012) and is associated with triggering non-motor decisions via a disinhibition mechanism (Zavala et al., 2017). This competition signal is then relayed from the STN, via diffuse excitatory projections, to the GPi/Snr complex.

The second function is to regulate the decision time by manipulating the decision threshold via signalling in the opposing Go/No-Go pathways which is a consequence of the activity of the D1 and D2 populations of MSNs in the striatum. The

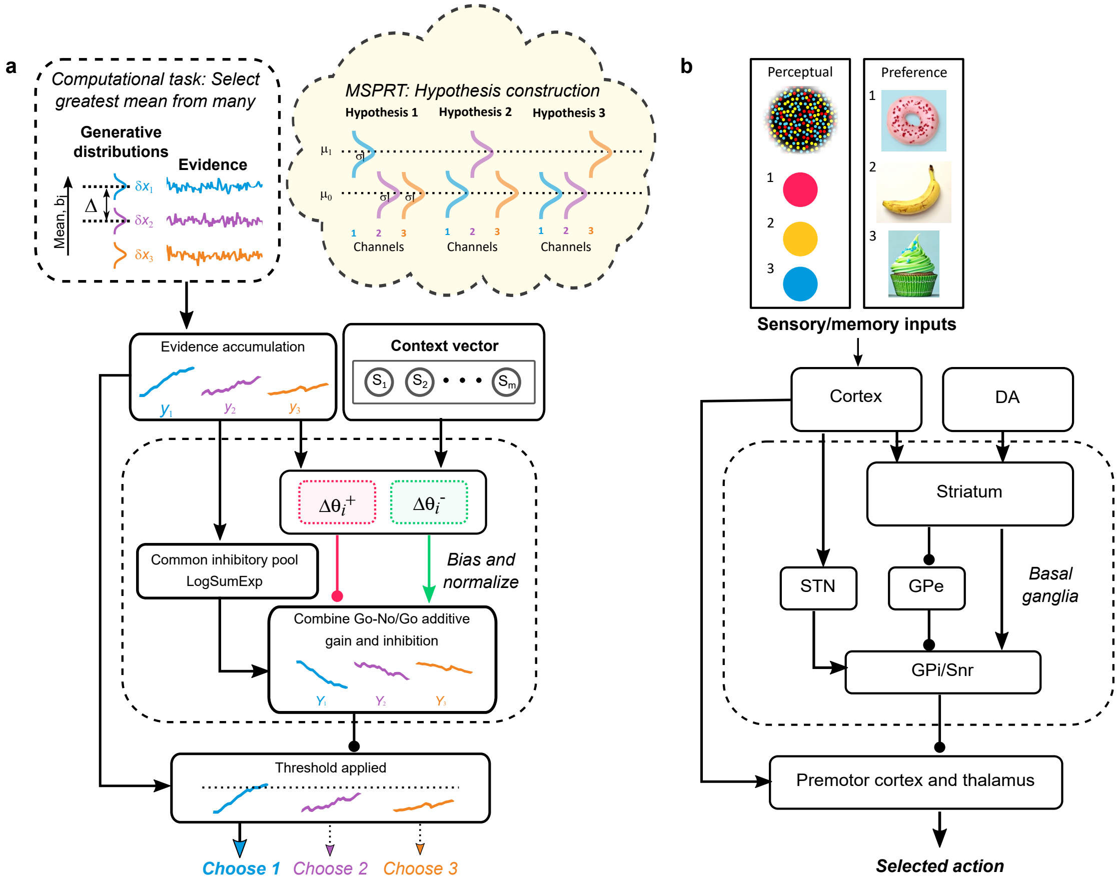

Figure 1: Proposed map between the architecture of the decision model and the anatomy of the basal ganglia. Panels show: a) System-level view of the decision model and, b) Basal ganglia anatomy. The computational task to be solved is 'the problem of the greatest one'; *i.e.* to identify the channel with greatest activity. The evidence sampled from the generative distributions correspond to cortical inputs to striatum, which are assumed to represent action salience values generated in response to contextual information provided by sensing of the external world, internal state and memory. In this schematic, the transformed evidence in favour of choice 1 crosses the threshold and so initiates the action associated with option 1 (bottom of panel (a)).

decision thresholds can therefore be context-dependent. Cortical signalling is assumed to convey contextual information which determines striatal dopamine levels which then govern the activity of the D1 and D2 MSN populations and hence the effective threshold on the accumulated evidence for action initiation (Lo and Wang, 2006; Lepora and Gurney, 2012).

2.4 Reinforcement learning of thresholds

As discussed in the introduction, previous models have combined sequential sampling decision models with reinforcement learning by learning the association between stimulus and reward, a value-based learning rule. The present model differs in that we assume this association has already been learned (the evidence distributions that generate the action-salience values), and instead we are learning the association between a context cue and the cost of time. The goal of the agent is to find the threshold that identifies the action channel with the highest expected salience value in such a way that balances the accuracy of the decision with the cost of the time taken to make the decision.

The optimal balance is defined as decisions that minimize the expected cost of the decision, which in this case is given by a linear combination of the error rate and the decision time. Explicitly, the problem is defined as follows below.

Optimality problem definition Let $W(j,k)$ denote the cost of deciding H_k when H_j is true, and assume that $W(j,k) > 0$ for $j \neq k$ and $W(k,k) = 0$. The optimality problem is to select the optimal decision thresholds, θ^* that minimize the expected cost, U_c,

$$\theta^* = \arg\min_x U_c(\theta), \tag{6}$$

$$U_c(\theta) = \mathbb{E}\left[cN(\theta) + W\left(H, \delta\left(\theta\right)\right)\right], \tag{7}$$

where c is cost per observation, N is stopping time, δ is decision and H is the true hypothesis. We assume equal cost for any error, so $W(j,k) = C_j$ for all k.

Since we assume equal error costs for all error types, there is a unique map between the cost ratio C_j/c and the optimal decision threshold θ_j^* on hypothesis H_j (Lepora, 2016; Griffith et al., 2021). Furthermore, for simplicity in presenting the results in this paper, we assume equal error costs across all options so that $W(j,k) = C$ for all j and k. This means there is a single optimal decision threshold for all evidence streams, although in general this need not be the case.

2.4.1 REINFORCE algorithm for learning decision thresholds

In applications of reinforcement learning, agents interact with their environment to learn about their actions and the surrounding environment through observations of state and reward signals. It is important to define how we will interpret 'action' in this RL framework. Usually, 'action' in a decision-making context is taken to mean 'option selection', that is the physical manifestation of the output of the decision process. However, within the RL perspective as we frame it here, the 'actions' are the decision threshold values and it is the decision task, in combination with the evidence accumulation model, that constitutes our agent's 'environment'. So, on each trial, the decision model takes the chosen decision thresholds as inputs and returns a stochastic reward thus driving the agent's threshold learning. For a decision with N alternatives there are N scalar thresholds that can take any real-valued number, i.e., the action space for the threshold-finding problem is N-dimensional and continuous.

Policy parameterization for sampling decision thresholds A common approach for problems with multi-dimensional, continuous action spaces is to parameterize the policy using a multi-variate normal distribution (Williams, 1992; van Hasselt and Wiering, 2007). The threshold samples drawn from this distribution could be thought of as representing a snap shot of a noisy, continuous physical variable such as the firing rate of neurons in signalling pathways in the basal ganglia (Bogacz and Gurney, 2007). Decision thresholds are therefore sampled at the beginning of each trial from a normal distribution with state-dependent parameters,

$$\theta \sim \mathcal{N}\left(\mu\left(s, \mathbf{w}\right), \sigma\left(s, \mathbf{w}\right)\right), \tag{8}$$

or, using standard RL notation, where the policy $\pi(a|s, \mathbf{w})$ denotes the action distribution conditioned on the state, s, and parameter vector $\mathbf{w}$,

$$\pi(a|s, \mathbf{w}) \doteq \frac{1}{\sigma(s, \mathbf{w})\sqrt{2\pi}} \exp\left(-\frac{(a - \mu(s, \mathbf{w}))^2}{2\sigma(s, \mathbf{w})^2}\right), \tag{9}$$

and the policy parameter vector is split into two parts, $\mathbf{w} = [\mathbf{w}_\mu, \mathbf{w}_\sigma]^\mathsf{T}$,

$$\mu(s, \mathbf{w}) \doteq \mathbf{w}_\mu^\mathsf{T}\mathbf{x}(s) \quad \text{and} \quad \sigma(s, \mathbf{w}) \doteq \exp\left(\mathbf{w}_\sigma^\mathsf{T}\mathbf{x}\left(s\right)\right), \tag{10}$$

where $\mathbf{x}(s)$ is a feature vector that represents the state, s.

REINFORCE The REINFORCE update occurs after each trial and is determined by the reward feedback, R, for that trial. The update rules for the weights in a Gaussian-parameterized policy follow from the policy gradient theorem (Williams, 1992) and are given by the following 3-factor learning rule:

$$\mathbf{w}_\mu \leftarrow \mathbf{w}_\mu + \alpha\left(R - \hat{v}\left(s\right)\right)\frac{1}{\sigma(s, \mathbf{w}_\sigma)^2}\left(a - \mu\left(s, \mathbf{w}_\mu\right)\right)\mathbf{x}(s), \tag{11}$$

$$\mathbf{w}_\sigma \leftarrow \mathbf{w}_\sigma + \alpha\left(R - \hat{v}\left(s\right)\right)\left(\frac{\left(a - \mu\left(s, \mathbf{w}_\mu\right)\right)^2}{\sigma\left(s, \mathbf{w}_\sigma\right)^2} - 1\right)\mathbf{x}(s), \tag{12}$$

where $\hat{v}$ is a state-value approximator, or so-called critic, with parameters $\mathbf{w}_v$:

$$\hat{v} = \mathbf{w}_v^\mathsf{T}\mathbf{x}(s), \tag{13}$$

which update according to stochastic gradient ascent,

$$\mathbf{w}_v \leftarrow \mathbf{w}_v + \alpha_v\left(R - \hat{v}\left(s, \mathbf{w}_v\right)\right)\mathbf{x}(s), \tag{14}$$

which was found to be effective for the critic as it implements an exponentially-weighted average of the reward received in state, s.

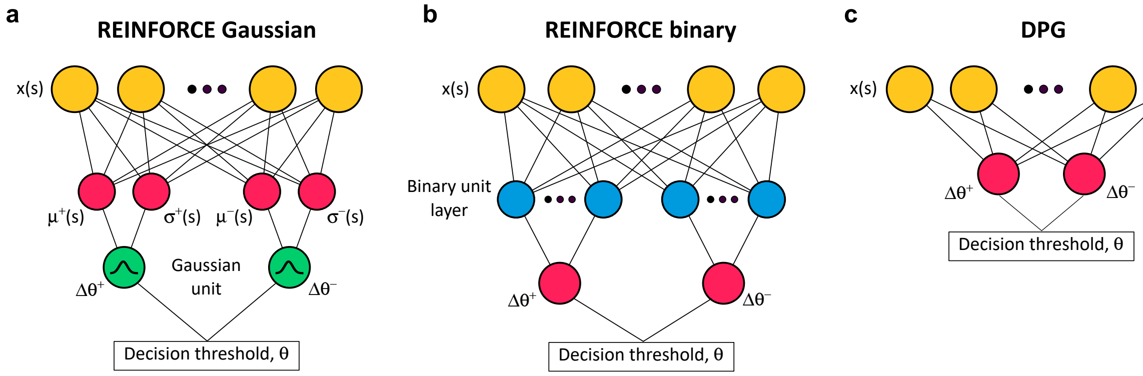

Figure 2: Network representations of: a) Gaussian parameterization, b) binary unit parameterization, and c) deterministic policy parameterization.

2.4.2 Alternative REINFORCE parameterization using binary units

The Gaussian parameterization is well suited to this type of action space, but its biophysical plausability is not so obvious. For comparison, we use a different parameterization where the decision threshold uses a linear combination of the stochastic output of binary units, or 'artificial neurons', to represent the threshold (Lepora, 2016; Williams, 1992), and so could be considered more biologically plausible. The decision threshold is given by,

$$\theta = \sum_{j=1}^{N_b} B_j y_j, \tag{15}$$

where B_j are fixed coefficients that weight the output of the j^{th} unit, and y_j is the stochastic binary output of the j^{th} unit so $y_j = \{0, 1\}$. The fixed coefficients are scaled exponentially, so the range of possible decision threshold values is from zero to $\theta_{\max}$:

$$B_j = \frac{(1/2)^j}{1-(1/2)^{N_b}} \theta_{\max}, \tag{16}$$

where N_b is the number of binary units contributing to the threshold. The conditional probability of the binary units 'firing' or being 'on' ($y_j = 1$) is distributed according to a logistic function, which is itself a function of the state and parameter vectors,

$$y_j \sim p(y_j|s_j) = f(s_j)y_j + (1 - f(s_j))(1 - y_j), \quad f(s_j) = \frac{1}{1 + e^{-s_j}}, \tag{17}$$

where s_j is a linear combination of the j^{th} parameter vector and the state vector $\mathbf{x}(s)$,

$$s_j = \sum_{i=1}^{N_b} w_{i,j} x_i. \tag{18}$$

The REINFORCE update rule for the weights is then given by

$$\Delta \mathbf{w}_j = \alpha_b [y_j(t) - f(s_j)][R(t) - \hat{v}]\mathbf{x}(s) \tag{19}$$

The form of the rule is similar to the Gaussian-update rule, and again contains features of a Hebbian 3-factor learning rule. In the spirit of the original REINFORCE paper (Williams, 1992), both parameterizations are represented as connectionist networks in Figure 2.

2.4.3 Deterministic Policy Gradient algorithm for learning thresholds

To provide a comparison for the REINFORCE method we also investigated a simple deterministic policy for the thresholds, which is learned using the deterministic policy gradient (DPG) method (Silver et al., 2014). In this approach, rather than sampling the threshold stochastically from an underlying learned distribution, we learn a policy denoted, $\mu_u = u^{\mathsf{T}}\mathbf{x}(s)$, which is deterministic and is chosen so it is given by the dot product of the state vector, $\mathbf{x}(s)$, and the learned parameter vector, u. The learning algorithm is *on-policy*, which means that the learned policy μ_u is the policy used to select the threshold during learning. This is in constrast to an *off-policy* method, which uses a separate policy to sample actions and rewards in order to update the deterministic policy, μ_u. Often, an off-policy method is needed to adequately explore the action space, but the reward in this task is stochastic enough to adequately drive exploration of the action space.

The threshold parameters are learned using the on-policy deterministic actor-critic algorithm, described by (Silver et al., 2014), using trial-by-trial updates, since it is assumed that the state is independent of the action taken. In general, the on-policy deterministic actor-critic algorithm is as follows:

$$\delta_t = R_t - Q^w(s_t, \theta_t), \tag{20}$$
$$w_{t+1} = w_t + \alpha_w \delta_t \nabla_w Q^w(s_t, \theta_t), \tag{21}$$
$$u_{t+1} = u_t + \alpha_u \nabla_u \mu_u(s_t) \nabla_\theta \left. Q^w(s_t, \theta_t)\right|_{\theta=\mu_u(s_t)}, \tag{22}$$

where δ denotes the reward prediction error, $Q^w(s_t, \theta_t)$ is a compatible differentiable action-value function (where selection of θ is the action), which is parameterized by w, and α_w and α_u are learning rates for the w and u parameter vectors respectively.

For our choice of deterministic policy, $\mu_u = u^\mathsf{T}\mathbf{x}(s)$, a compatible approximate state-action value function is $Q^w(s, \theta) = \theta\mathbf{w}^\mathsf{T}\mathbf{x}(s) + \mathbf{v}^\mathsf{T}\mathbf{x}(s)$, where the state-action value parameter vector is split into two parts, $w = [\mathbf{w}, \mathbf{v}]^\mathsf{T}$. Given these assumptions, it follows that the specific update rules for this task, according to the on-policy deterministic actor-critic algorithm, are given by:

$$\delta_t = R_t - Q^w(s_t, \theta_t), \tag{23}$$
$$u_{t+1} = u_t + \alpha_u(\mathbf{w}^\mathsf{T}\mathbf{x}(s_t))\mathbf{x}(s_t), \tag{24}$$
$$w_{t+1} = w_t + \alpha_w \delta_t \theta_t \mathbf{x}(s_t), \tag{25}$$
$$v_{t+1} = v_t + \alpha_v \delta_t \mathbf{x}(s_t). \tag{26}$$

The DPG weights can also be represented as a simple connectionist network similar to the REINFORCE representations (Fig. 2)

2.4.4 Optimising learning rates for model comparison

To assess the performance of the different learning algorithms in finding the optimal decision threshold, we used *regret* as our performance metric. Regret, in this context, can be thought of as the total reward that is ‘left behind’ due to a sub-optimal decision threshold. Regret was calculated as the difference between the maximum total expected reward over all trials (as if the optimal decision threshold were used for all trials) and the actual total reward obtained on average over 5000 trials from a naive initial state on the standard 3-choice perceptual decision-making task.

To ensure a fair comparison of the algorithms, we optimised the critic and policy learning rates, which are the relevant hyperparameters for the models. Optimization was done by random search (10^5 samples) of the parameter space defined by $\alpha_c \in [10^{-3}, 10^{-2}]$ and $\alpha_p \in [10^{-4}, 10^{-3}]$. The pairs of hyperparameters that minimized the regret, starting from a naive initial state, were selected.

3 Results

3.1 Summary of decision model outputs on multiple-choice decision task

We first characterise the output from the decision model on a standard 3-choice decision task for a range of decision thresholds. There is no learning at this stage, so we are aiming to characterise the performance of the decision making part of the model for different threshold values on a standard 3-choice decision task. The decision model receives action salience observations as a 3-vector in a time series that represents the signals from 3 different signalling pathways that represent action salience for 3 different actions. The goal is to identify the signalling pathway with the highest mean action salience.

Action salience observations are sampled from a normal distribution, $\mathbf{x}(t) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where we set $\boldsymbol{\mu} = [10.3, 10, 10]$ and the covariance matrix was set to the identity matrix $\boldsymbol{\Sigma} = \boldsymbol{I}_3$. Observations were therefore in essence from 3 independent Gaussian distributions with the same standard deviation ($\sigma = 1$) and the gap between the largest mean and the next largest mean, $\Delta = 0.3$. This gives a challenging maximum-finding task in which it would be difficult to improve on a chance-model from just one observation.

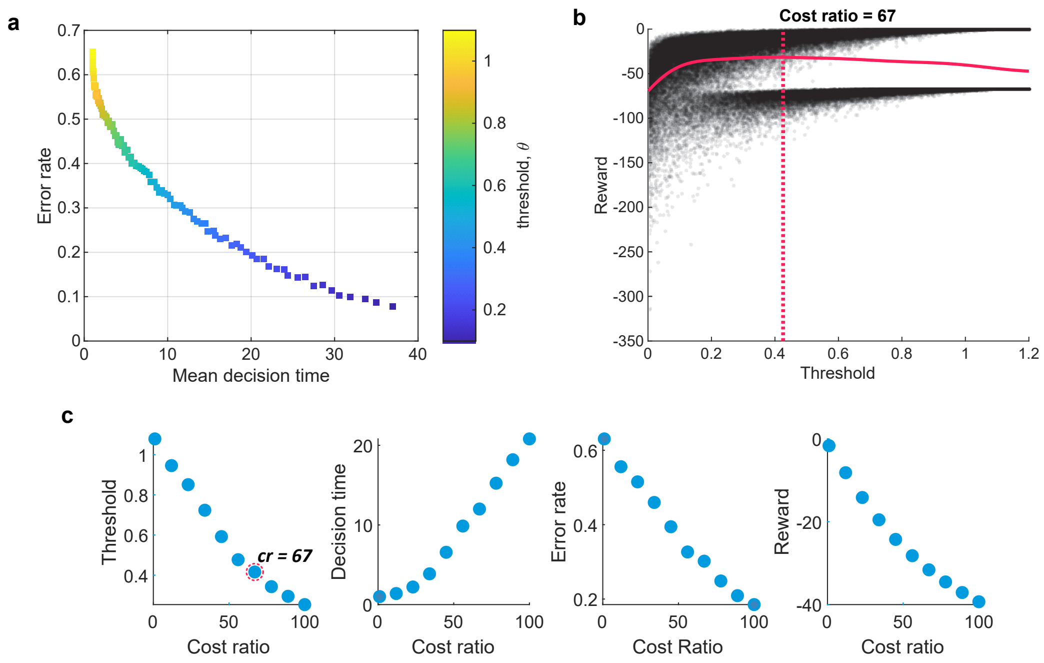

Figure 3: Typical decision model outputs on a 3 choice-task: a) Speed-accuracy curve, generated from 100 linearly-spaced decision-threshold values, for the 3-choice decision task described in the main text. The error rate is plotted against the mean decision time (averaged over 10^4 trials) and shows that as threshold value gets smaller, the decision time increases and the error rate decreases. Balancing these outcomes is achieved through decision threshold selection and the optimal threshold value depends on the objective function for the task. b) An estimate of the expected reward as a function of the decision threshold. The reward function is also parameterized by the cost ratio, which is assumed to be specific to the task. In this illustrative example, to visualise the reward as a function of the threshold only we have fixed the ratio of the cost of an error to the cost of an observation to $C/c = 67$. The vertical dashed line shows the maximum of the expected reward function and hence the predicted optimal decision-threshold for this case. c) (first panel) Optimal thresholds plotted against the cost ratio. The highlighted data point corresponds to the argmax of the reward function $(\mathrm{cr} = 67)$ shown in panel (b). (second panel) The expected decision time for optimal performance as a function of the cost ratio, (third panel) The expected error rate for optimal performance plotted against the cost ratio. (last panel) Expected reward from optimal decision-making plotted again cost ratio.

As can be seen from Figure 3a, when the threshold value results in a decision after a single sample, the accuracy of the decision model is no better than chance, demonstrating the importance of evidence integration over time for accurate decision making. As the threshold value decreases (remembering that in our decision model, the decision threshold is approached from above due to the release of inhibition mechanism in the basal ganglia) this allows the collection of more evidence samples, increasing the decision time, and reducing probability of making an error (Figure 3a).

To demonstrate the relationship between the expected reward on the task and the threshold on the evidence we plotted the expected reward as a function of the decision threshold (Figure 3b, data shown is for cost ratio of 67, where error cost is set to 1). The expected reward was calculated by sampling rewards from the decision model using 10^5 decision thresholds that were randomly sampled from a uniform distribution $[10^{-3}, 1.2]$ and then regressing the output reward over the decision thresholds with a Guassian Process model. The reward samples are plotted as a scatter and show a bimodal distribution due to the difference in reward resulting from error trials and correct trials. Variance in the reward increases for smaller thresholds (longer decision times) because, as is typically the case, in stochastic accumulator models such as the current model or the DDM, the response time distributions become wider as the expected response time increases (Ratcliff, 1978; Bogacz et al., 2006). By finding the maximum of the expected reward we can find the decision threshold value that is optimized for maximizing the expected reward on the task. The goal of a threshold learning agent would therefore be to efficiently find this threshold in order to maximize performance on the task.

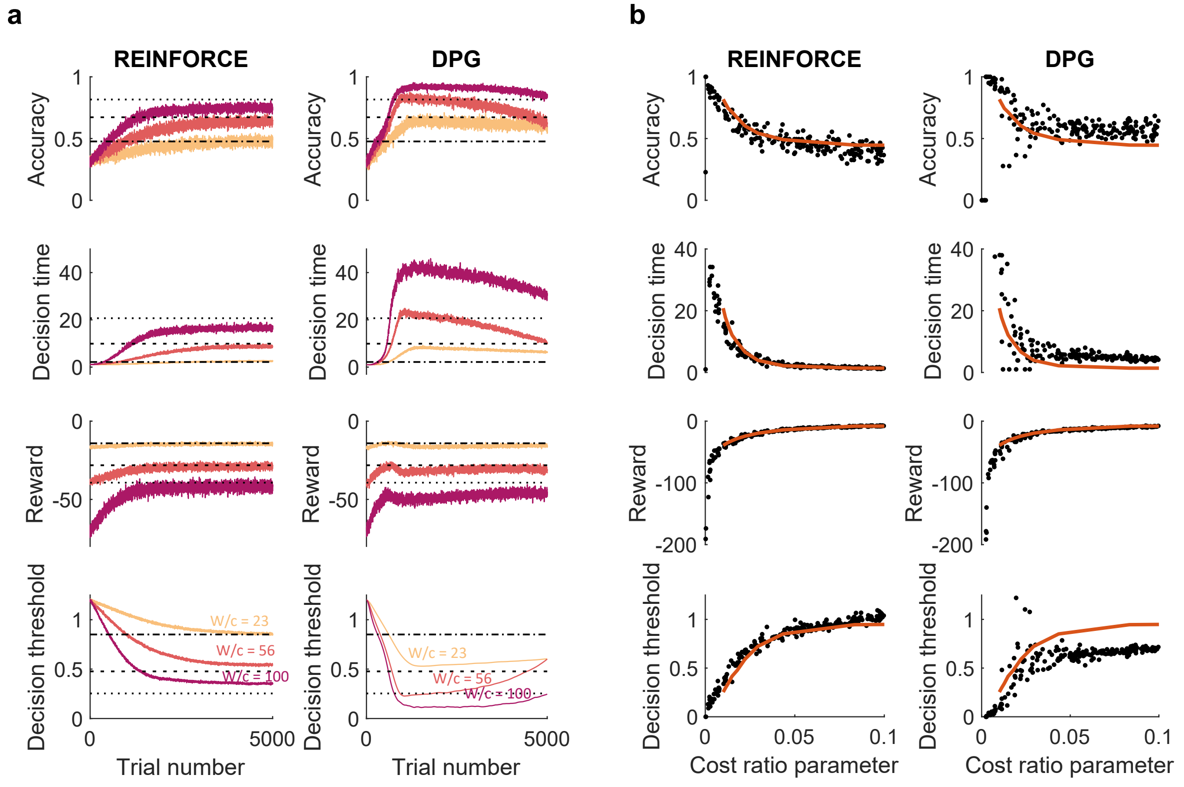

Figure 4: Threshold learning results: a) Learning curves (averaged over 100 episodes) for best performing REINFORCE model and DPG model. Different curves correspond to different cost ratios. Dashed horizontal lines indicate the expected values arising from optimal decision-making. b) REINFORCE and DPG learning of decision threshold over 200 learning episodes with the cost ratio parameter, c/W, sampled uniformly from $[0, 0.1]$. Results indicate the converged values after 5000 trials (averaged over last 100 trials). Red curve is optimal performance found by exhaustive search described in methods.

We characterized the effect of changing the relative importance of making an accurate vs fast decision in terms of maximizing reward by varying the cost ratio in the task's reward function. As the ratio of the cost of the error relative to the cost of evidence sampling increases, then the optimal threshold decreases to collect more samples. This decrease in optimal threshold has the effect of rebalancing the speed-accuracy trade-off by increasing the optimal expected decision time and reducing the optimal error rate on the task (Figure 3c).

3.2 REINFORCE learns the optimal decision thresholds for multi-alternative decisions

Next we used the same decision task to test the threshold learning component of the model to examine whether the optimal performance can be learned from a naive initial state.

As expected, when challenged with the three-choice task, REINFORCE learned decision thresholds that converged to values that approximated the optimal decision threshold (Figure 4; REINFORCE column shows only the Gaussian parameterization as this was the best performing version of the algorithm). This was confirmed by comparison to the best threshold values obtained through an exhaustive search of the threshold space over a range of cost ratios (Figure 3). The results shown in Figure 4 demonstrate that REINFORCE learns a good approximation to the optimal speed-accuracy trade-off on this multi-alternative decision task.

The DPG algorithm was less successful at finding stable decision thresholds even though the resulting reward, when averaged over many learning episodes, converged to a value that approximated the optimal reward (Figure 4a). The DPG algorithm under-performed relative to REINFORCE at higher cost ratios (Figure 4a,b). The decision threshold values governed by the DPG algorithm responded more swiftly but were less stable, due to the deterministic threshold

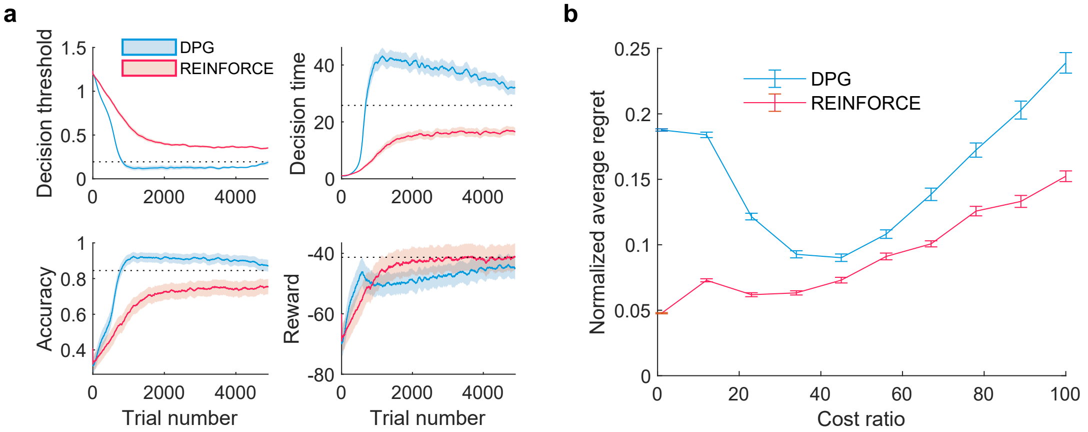

Figure 5: Performance comparison of REINFORCE and DPG learning models: a) Learning curves (averaged over 100 episodes) for cost ratio = 100. The horizontal dashed line represents optimal decision-making. The shaded regions are standard error of the mean (SEM). b) Normalized regret (averaged over 100 episodes) plotted against the cost ratio. Error bars are SEM. Results compare the overall performance of the learning models for a variety of cost ratio conditions. The REINFORCE learning model outperforms the DPG learning model over all cost ratios tested. The DPG model tends to respond more quickly but the REINFORCE model learns more precise thresholds, in the long run REINFORCE learning leads to lower regret.

policy of the DPG, than those found by REINFORCE. This was because rewards are highly stochastic and so the true gradient may not be followed for many trials before the error can be corrected: the deterministic threshold policy seems to exacerbate the problem.

It transpires that, for both learning algorithms, higher cost ratios led to poorer approximations of the optimal value. This poorer behaviour is because high cost ratios require large thresholds for optimal decisions. At large decision thresholds there tend to be fewer errors but greater variability in the decision times, which results in a flatter reward landscape in the threshold space and hence a reduced sensitivity of the expected reward to the threshold value. There is less of a gradient for the algorithm to follow and so the learning slows to a ‘good enough’ value. This is less of a problem at low cost ratios because the reward function is still quite sensitive to the threshold value around the optima and so there is adequate gradient for convergence to a closer approximation.

Performance comparison of threshold learning models To compare the performance of the REINFORCE and DPG learning rules, we used the best model versions (those with optimized learning rates) that minimized regret over a learning period of 5000 trials from an initial naive state. Regret was normalized to the maximum reward possible for that particular cost ratio. REINFORCE outperformed DPG across all cost ratios tested (Figure 5).

Looking at the time evolution of the model variables for the best performing model versions, DPG adapts the threshold value more quickly and so initially outperforms the REINFORCE algorithm. However, it then tends to overshoot the optimal threshold value leading to accurate but slow decisions. REINFORCE is slower to respond and converges to an underestimate of the threshold. However, due to the non-linear dependence of reward on the threshold, REINFORCE approximates optimal decision-making whereas DPG is does not achieve that level of performance on the task.

3.3 REINFORCE learning of thresholds is robust to diverse decision task conditions

Relearning – changing reward functions: Our next consideration was to compare the learning models’ ability to cope with changing reward functions. We challenged the models with a task that involved continuous presentation of perceptual stimuli, as in previous tasks, except that the cost-ratio parameter was changed at equal time intervals (25%, 50%, 75%) of the learning episode. In consequence, the model had to learn a new optimal decision threshold after having already learned a previous speed-accuracy trade-off. The challenge for the algorithm was to recognise the changed reward function and adapt the decision threshold by following the new gradient to learn the new optimal speed-accuracy trade-off.

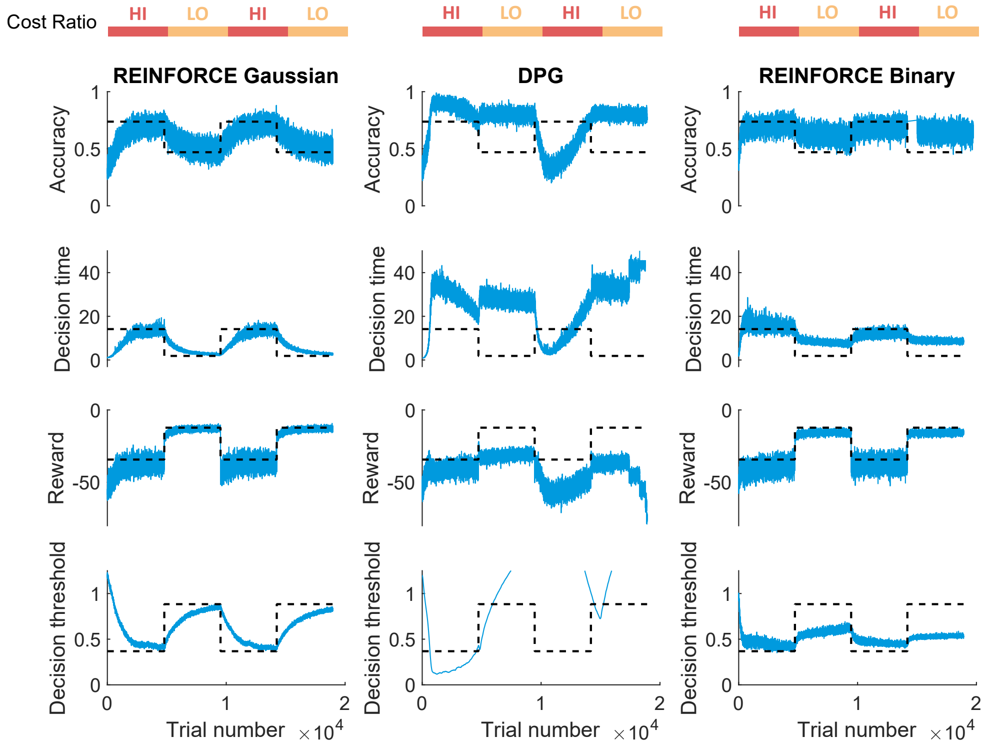

Figure 6: Learning and relearning curves for changing cost ratio conditions. REINFORCE learns and relearns optimal decision thresholds for non-static objectives. High and low task conditions correspond to cost ratios of 80 and 20 respectively. Learning curves are averaged over 100 episodes. Dashed lines represent optimal decision-making.

The Gaussian-parameterized REINFORCE algorithm performed well at this version of the task to result in model variables that converged to near optimal values (Figure 6, left column). The binary-parameterized REINFORCE method also resulted in a good return over the learning episode with approximately optimal decision-making as seen from the reward learning curve. However, the threshold value itself did not adapt as well as the Gaussian parameterized version and was less flexible to relearning due to the exponential weighting of the binary unit outputs (Figure 6, right column). (Although, as discussed, the reward is not so sensitive to the decision threshold value in the close-to-optimal region, and so this has less effect on the reward obtained.) The DPG was not well suited to this type of task as the changing reward function exposed the tendency of the algorithm to instability as it failed to converge to stable values (Figure 6, central column).

Context recognition – associating SAT with context cues: One way to solve the relearning problem is for the model to recognise that optimal speed-accuracy trade-offs may be contextual. Our model can associate reward functions with a particular context and so adapt to novel speed-accuracy trade-offs without forgetting previously learned thresholds-context associations. We demonstrated the model's ability to do this by using a task design similar to the relearning task, where the cost ratio changed at regular intervals, but those changes were also signalled by a context cue, which was represented in the model by the context feature-vector, $\mathbf{x(t)}$.

On first presentation of the cues, all models improved performance by learning to associate decision thresholds with both contexts. When the previously learned context cue reappeared, models immediately switched to the previously learned decision threshold, with no need for relearning, therefore outperforming the models that do not recognise context cues (Figure 7). These results demonstrate how context recognition and learning to associate with a learned threshold in models allows instantaneous optimal decision-making without relearning. DPG model variables failed to converge to stable values although context recognition mitigated some of the issues that REINFORCE binary parameterization and DPG encountered in the previous relearning task (Figure 7).

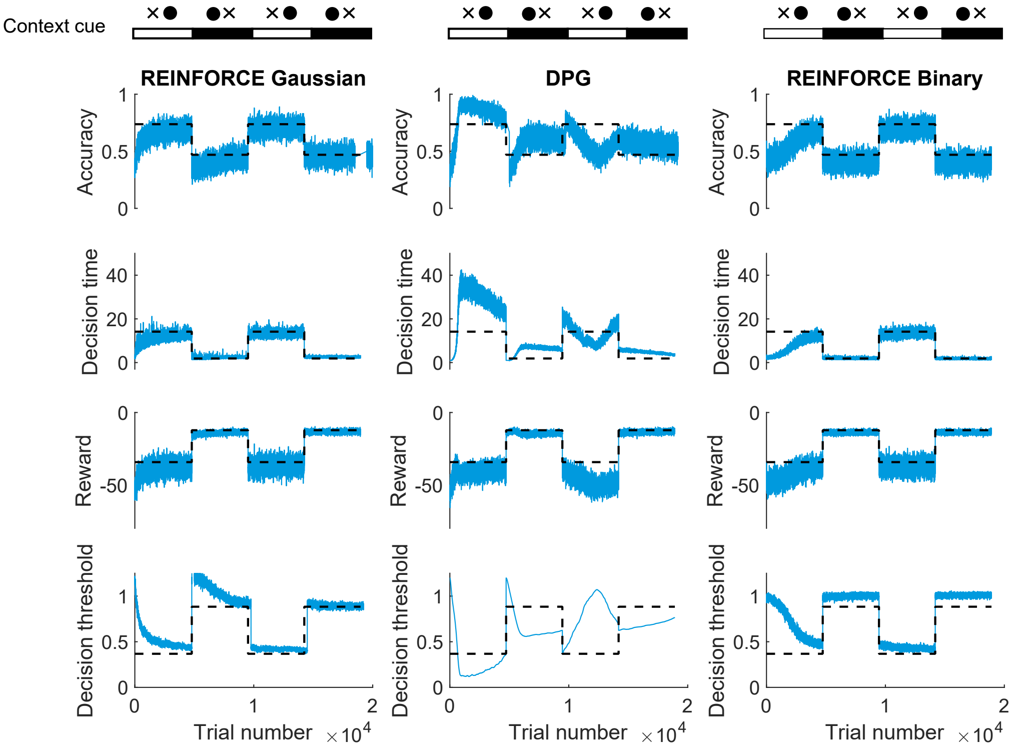

Figure 7: Learning and context switching. The cost ratio conditions are as previous simulation (cost ratios in order are [80, 20, 80, 20]) but changes are signalled by a context cue represented by the symbols and colour changes in the bar at the top of the figure.

Case of repeated trials – no explicit time cost: Stochastic accumulation to bound models, such as the SPRT and DDM, are often associated with reward functions that are simple linear combinations of the error rate and the decision time. Such reward functions lack biological realism, mainly because they are only concerned with maximizing expected reward on individual trials, leading to counter-intuitive decisions that require large deliberation times in order to obtain the necessary predicted accuracy. To illustrate the ability of REINFORCE to improve performance on a more realistic task we challenged the algorithm over repeated decisions within a finite time frame and removed the explicit time cost from the reward function. The total time per trial was 2000ms and there was a waiting time of 100ms between each decision taken in that time frame. In this version of the decision task, the reward function was simply the total reward obtained at the conclusion of the trial, which may have included many individual decisions. The cost of time was therefore implicit: the model needed to learn that deliberating on any individual decision for too long came with an opportunity cost because it would miss out on future decisions and therefore the possibility of greater reward, even though individual decisions may be less accurate. This makes the problem equivalent to maximizing the reward rate, a decision objective often considered in the literature (Gold and Shadlen, 2002).

As expected, REINFORCE learned to make faster but less accurate decisions when the total expected reward of an individual decision was increased but the task difficulty remained the same. This replicated an effect observed in human participant decision tasks, namely the magnitude effect. The magnitude effect is the tendency for people to make faster decisions on tasks where the expected reward is greater even if the task difficulty does not change (Fontanesi et al., 2019b). The model converged to decision thresholds that replicated this effect so that for tasks where the total reward received per trial was greater, the decision time was reduced even though difficulty of discriminating the three choices remained the same (Figure 8b). Threshold learning by reinforcement therefore naturally accounts for the magnitude effect by encoding the opportunity cost of the decision time in the adapted threshold.

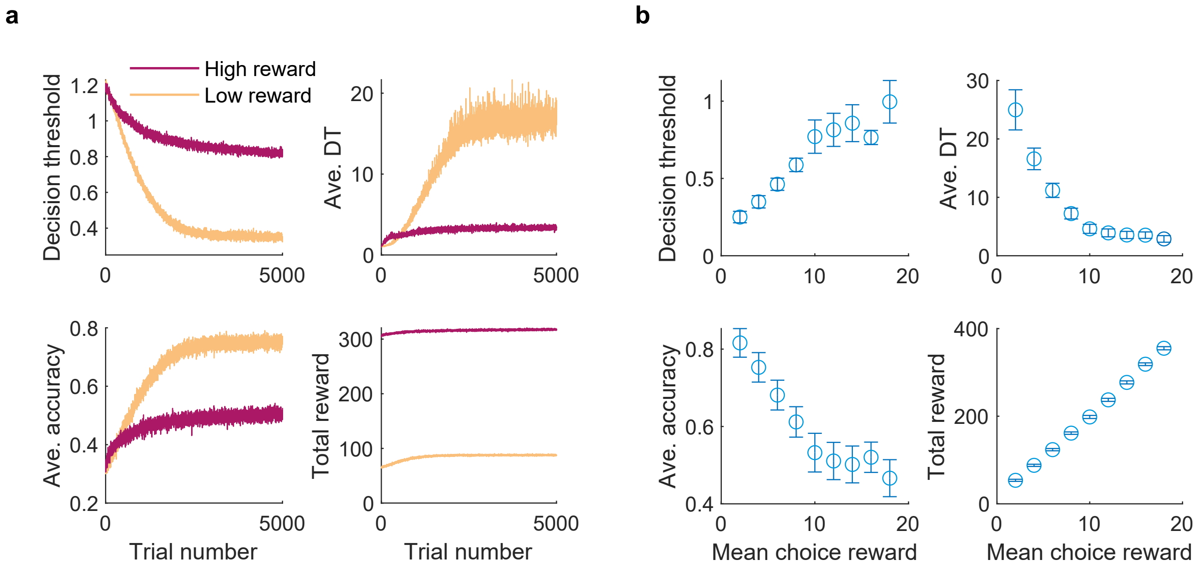

Figure 8: Results from repeated trials simulation demonstrate the magnitude effect. a) Learning curves (averaged over 100 episodes) for high reward ($R_1 = 18, R_{2,3} = 14$) and low reward ($R_1 = 6, R_{2,3} = 2$) conditions. b) Learning and decision model results for different expected value conditions. Results are converged values after 5000 trials (averaged over last 100 trials) and averaged over 10 learning episodes. Error bars are SEM.

4 Discussion

In this paper we have contributed to ongoing research efforts to bring together the common RL and evidence-accumulation approaches to decision making by proposing an integrated RL-evidence accumulation model that is designed to be consistent with knowledge of the neural circuitry of the basal ganglia. We have suggested a biologically-plausible implementation, in which the Go and No-go pathways of the basal ganglia modulate decision thresholds through context-sensitive gain control. This is in contrast to the view that the basal ganglia explicitly encode choice values via the Go and No-go pathways. A key conceptual proposal of the model, then, is that choice values are learnt separately from the value of time. Moreover, a key assumption of the model is that the principal role of the basal ganglia in decision-making is to transform the choice evidence and, in a context-sensitive manner, allocate the time for the decision-making process.

We have evaluated the model by demonstrating REINFORCE's ability to learn the time cost in various scenarios. These scenarios included those with and without context cues, under non-static reward functions and over repeated trials. Notably, under repeated trials, the model reproduced the magnitude effect: a phenomenon where shorter decision times are observed for decisions with higher expected rewards, even if decisions are equivalently difficult. This contrasts with the RL-DDM model, which uses a phenomenological model for the decision time using the learned choice values as a parameter (Fontanesi et al., 2019a). Under our model, the observed phenomenon is a natural consequence of the model's adaptive decision threshold and the task reward conditions.

Biophysical plausibility Learning in the basal ganglia is thought to occur via a 3-factor Hebbian learning rule. This rule posits that the change in synaptic strength is proportional to both pre-and post-synaptic activation, as well as the local concentration of dopamine. These three components have a correspondence to the factors in the REINFORCE update, bridging the gap between the biological-implementation level and the algorithmic level. The three correspondences are as follows: i) the context vector, $\mathbf{x}(s)$, corresponds to presynaptic activation from cortical signalling to the basal ganglia; ii) the perturbation to the mean, $a - \mu\left(s, \mathbf{w}_\mu\right)$, the difference between the sampled decision threshold (denoted as the action, a) and the current value of the mean of the Gaussian parameterised by weights $\mathbf{w}_\mu$, corresponds to postsynaptic activation of the striatal MSNs, and iii) the prediction error, $R - \hat{v}(s)$, corresponds to the supposed teaching signal from phasic dopamine signals in striatum.

A similar correspondence does not necessarily follow for the delta term in the update rule for the parameter vector for σ. Therefore, although this form of the rule results in effective exploration and convergence of the policy weights, this is a potential criticism of the model. An alternative method for controlling the standard deviation of the policy parameterization, or the effective exploration of the policy, that would hold biophysical plausibility (and functionality) would be to have the standard deviation of the sampled threshold be proportional to the current estimated mean, $\sigma = f\left(\mu\left(s, \mathbf{w}\right)\right)$. This would replicate the Fano factor rule that describes the variability of neural firing rates. The flattening of the reward landscape with increasing decision times, could give this term a benefit in potentially preventing the algorithm from getting stuck on flat regions of the reward landscape, and so help drive the learning away from previously learned thresholds that are no longer valid for a given context.

Basal ganglia and dopamine Previous functional accounts of the Go/No-Go signalling pathways in basal ganglia have suggested that the Go and No-Go pathways convey and encode evidence explicitly, either for or against actions coded by specific channels. In contrast, a key assumption we make for the function of the Go and No-Go pathways in our model is that, rather than explicitly encoding option-reward associations, they convey positive and negative shifts to the effective decision thresholds that are associated with particular contexts. This account of the function of the Go/No-Go signalling pathways in the basal ganglia aligns with recent proposals that levels of striatal dopamine govern the effective threshold for decision making in the basal ganglia (Berke, 2018; Chakroun et al., 2023). High striatal dopamine levels are associated with increased activity in the D1 (Go) pathway and decreased activity in the D2 (No-go) pathway, which in our model corresponds to a net negative change to the decision threshold and therefore a greater tendency towards action rather than inaction.

Further supporting evidence for this perspective has come from a recent study that used sequential sampling models to estimate the effective decision threshold used by human participants in a simple decision task (Chakroun et al., 2023). In this study, Chakroun et al. found that participants who underwent a pharmacological intervention whose effect was to boost activity of the Go pathway and a group where signalling in the No-Go pathway was suppressed had the same effect of reducing the effective decision threshold. This manifested in faster decision times but also less accurate decision making since this meant less accumulated evidence was necessary to trigger a decision. This behaviour is replicated by the model in that it learns to associate rewarding task context with lower decision thresholds and therefore more frequent and faster decisions.

Relationship to other decision-making models The work in this paper is closely related to work by Bogacz and Gurney (2007) and Lepora and Gurney (2012). In those studies, the authors used the same Bayesian framework as our model to implement the MSPRT. The evidence accumulation models in those studies and our paper are mathematically equivalent, but we implement the proposal from Lepora and Gurney (2012) that the Go and No-Go pathways apply gain to the evidence thereby implementing positive and negative shifts to the decision threshold. Our model therefore has redundancy in that the learned parameters representing the positive and negative shifts to the threshold, done by the dual-pathway, are equivalent to a single value.

The function of the dual pathway in basal ganglia has been debated in the literature (Calabresi et al., 2014; Jaskir and Frank, 2023). Previous models have used the architecture to underlie assumptions regarding positive and negative reward prediction errors and their role in encoding the expected choice values (Collins and Frank, 2014; Möller and Bogacz, 2019). An assumption in our model is that there is no computational advantage conferred by a dual pathway model and that the reason such an architecture would evolve is to do with the biophysics of swiftly adapting the threshold over very short timescales (for instance, urgency signals) using a balanced push-pull control for gain modulation of the output nuclei (Abbott and Chance, 2005; Johnson et al., 2012). Other authors have also suggested that dopamine-modulated competitive dynamic of Go and No-go pathways provide the basis for flexible and reactive control of behaviour (Dunovan and Verstynen, 2016).

Our decision model shares features (such as multiple accumulators and global inhibition terms) with other models in the literature. For example, for their investigation into optimal policies for multi-alternatives, Tajima et al. (2019) proposed a model that implemented global inhibition by projecting that accumulated evidence onto a non-linear manifold (found by solving the Bellman equation for the task). They used a simple polynomial to approximate the full manifold. The resulting transformation of the evidence resembles that due to the log-sum-exp inhibition term in our model. Another example comes from Kriener et al. (2020) who proposed a modified winner-takes-all neural network where weakly active neurons were prevented from contributing fully through global inhibition in a similar manner to the feed-forward inhibition in our basal ganglia decision model. Kriener et al.'s proposed network satisfied Hick's law (decision time scales as $\log(N)$, as does MSPRT (McMillen and Holmes, 2006)) but also demonstrated that the network was not as efficient in finding the maximum as the full MSPRT model.

Our model also bears resemblance to the advantage racing diffusion model (Heathcote, 2022; Miletić et al., 2021), which eschews a global inhibition term but instead employs multiple accumulators that represent pair-wise choice *advantages*. These are quantities calculated by weighted sum of the pair-wise difference in estimated Q-values and the sum total of the Q-values over all choices. One issue with this approach, however, is that when extended to multiple choice decisions, the need for pairwise comparisons between all choices leads to growth in accumulators required for each decision, since in general an n-choice task requires ${}^{n}P_2$ accumulators. For example, 6 accumulators are required for 3 choices and 12 accumulators are required for 4 choices, and so on.

Our threshold learning model is related to several previous studies that also use reinforcement or reward tracking to adapt decision thresholds on accumulation-to-bound decision models. For instance, Myung and Busemeyer (1989) used feedback, simple error correction and hill-climbing methods to adapt the decision threshold on a free-response task similar to the task in this paper. Their threshold adaptation algorithm explained the observed performance improvements in human participants on a binary-choice, free-response task. In another example, Erev (1998) proposed a phenomenological update rule, that incorporated reinforcement, generalization and recency-bias components, to adjust the decision threshold in a single-observation, binary-choice task. In contrast to both of these examples, our model parameterizes the decision threshold so it is multi-dimensional enabling it to flexibly be applied to decisions with multiple alternatives as well as under different contexts.

Another interesting and neurally plausible approach to decision threshold adaptation was proposed by Simen et al. (2006) who used a system of stochastic differential equations to rapidly adjust the decision threshold parameter in order to optimize the speed-accuracy trade-off on a free-response binary choice task. The threshold is entirely determined by the reward rate estimate through an assumed linear relationship between the two variables. The rapid adaptation of the threshold therefore comes from the accumulation of reward feedback signals by the reward-rate estimator. One disadvantage of this approach is that due to the leaky accumulation of reward signals by the reward-rate estimator and the direct relationship to the threshold, the threshold parameter is not stored and decays over time. Our model differs in that the threshold parameter is learned directly and stable over time, meaning it can be applied to individual decisions with large time gaps between them without the decision threshold value decaying.

Further work and conclusion In the future it would be interesting to investigate the effect on the efficiency of the basal ganglia decision model under different learned action salience distributions, i.e., non-Gaussian, or where the separation, Δ, between distribution max-next means is not static. One way to mitigate the loss in efficiency from a non static or unknown separation, Δ, would be to have an adaptive $g = \Delta/\sigma$ term in the decision model, acting as gain on individual evidence channel observations, $x_i(t)$.

Previous efforts to integrate RL and accumulator models have predominantly centered around learning choice values, which then dictate the rate of accumulating evidence for that choice, e.g., drift rate in RL-DDM (Fontanesi et al., 2019a,b) and advantage parameters in (Miletić et al., 2021). In contrast, we have assumed that the action salience distributions that generate the decision evidence have already been learned and instead have focused on learning the cost of time, represented by the decision threshold, through reinforcement. However, both types of learning are completely complementary and could be combined so that choice values are learned concurrently as the basal ganglia learns to allocate decision time by setting the decision threshold, given the current context. Both types of learning may take place over different timescales, potentially explaining some discrepancies in the decision-making literature regarding learning rates on different decision tasks. For instance, response-time adaptation on binary-choice tasks has been observed to converge rapidly (over tens of trials) (Bogacz et al., 2010), possibly reflecting faster learning of the action salience distributions or tactical decision policies, as well as more gradually (over hundreds of trials) (Myung and Busemeyer, 1989), perhaps indicative of slower time-cost learning. The accumulation-to-threshold paradigm in combination with RL therefore offers a powerful, context-dependent framework for multi-alternative decision-making because it offers a mechanism for learning action-value distributions as well as how to efficiently allocate the time given to generating and integrating samples from those action-value distributions as evidence.

Acknowledgements NL and TG were supported in part by a Leverhulme Trust Research Leadership Award on ‘A Biomimetic Forebrain for Robot Touch’ (RL2016-039). SB was supported by an EPSRC PhD scholarship. Typeset with NeurIPS Template available at https://neurips.cc/Conferences/2025/CallForPapers, licensed under CC BY 4.0

Author contributions TG wrote the article and performed the analysis; NL and SB participated in critical review and discussion; NL led the research.

A Derivation of decision rule from MSPRT

The derivation here follows the example from (Bogacz and Gurney, 2007). Starting from Bayes' theorem and assuming equal priors for all hypotheses gives the posterior of the i^{th} hypothesis, H_i, in the same form as the original MSPRT paper (Baum and Veeravalli, 1994),

$$p(H_i|\mathbf{x}(1),\ldots,\mathbf{x}(T)) = \frac{p(\mathbf{x}(1),\ldots,\mathbf{x}(T)|H_i)}{\sum_{j=1}^{N} p(\mathbf{x}(1),\ldots,\mathbf{x}(T)|H_j)}. \tag{27}$$

Taking the natural log of the posterior gives,

$$\log p(H_i|\mathbf{x}(1),\ldots,\mathbf{x}(T)) = \log p(\mathbf{x}(1),\ldots,\mathbf{x}(T)|H_i) - \log\sum_{j=1}^{N}\exp(\log p(\mathbf{x}(1),\ldots,\mathbf{x}(T))|H_j). \qquad (28)$$

The decision rule is obtained by first substituting into the log-likelihood term in equation (28), the Gaussian density functions, $f_0 = \mathcal{N}(\mu_0,\sigma)$ and $f_1 = \mathcal{N}(\mu_1,\sigma)$ to represent the the hypotheses discussed in the methods section (Figure 1a),

$$\log p(\mathbf{x}(1),\ldots,\mathbf{x}(T)|H_i) = \log\prod_{t=1}^{T}\left(f_1\left(x_i\left(t\right)\right)\prod_{\substack{j=1\\ j\neq i}}^{N} f_0\left(x_j\left(t\right)\right)\right), \qquad (29)$$

$$= \sum_{t=1}^{T}\log f_1(x_i(t)) + \sum_{\substack{j=1\\ j\neq i}}^{N}\sum_{t=1}^{T} f_0(x_j(t)), \qquad (30)$$

$$= \sum_{t=1}^{T}\left[\log\frac{1}{\sqrt{2\pi}\sigma} - \frac{(x_i(t)-\mu_1)^2}{2\sigma^2}\right] + \sum_{\substack{j=1\\ j\neq i}}^{N}\sum_{t=1}^{T}\left[\log\frac{1}{\sqrt{2\pi}\sigma} - \frac{(x_j(t)-\mu_0)^2}{2\sigma^2}\right]. \qquad (31)$$

For legibility, we initially expand the first term on the R.H.S. of equation (31),

$$\sum_{t=1}^{T}\left[\log\frac{1}{\sqrt{2\pi}\sigma} - \frac{(x_i(t)-\mu_1)^2}{2\sigma^2}\right] = T\log\frac{1}{\sqrt{2\pi}\sigma} - \frac{T\mu_1^2}{2\sigma^2} - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\left[x_i(t)^2 - 2x_i(t)\mu_1\right]. \qquad (32)$$

Then, we rewrite the second term on the R.H.S of equation (31), so that the sum over the j index includes all evidence channels (subtracting the i^{th} channel term so both expressions are equivalent) and expand again,

$$\sum_{\substack{j=1\\ j\neq i}}^{N}\sum_{t=1}^{T}\left[\log\frac{1}{\sqrt{2\pi}\sigma} - \frac{(x_j(t)-\mu_0)^2}{2\sigma^2}\right] = \sum_{j=1}^{N}\sum_{t=1}^{T}\left[\log\frac{1}{\sqrt{2\pi}\sigma} - \frac{(x_j(t)-\mu_0)^2}{2\sigma^2}\right] - \sum_{t=1}^{T}\left[\log\frac{1}{\sqrt{2\pi}\sigma} - \frac{(x_i(t)-\mu_0)^2}{2\sigma^2}\right], \qquad (33)$$

$$= NT\log\frac{1}{\sqrt{2\pi}\sigma} - \sum_{j=1}^{N}\sum_{t=1}^{T}\left[\frac{(x_j(t)-\mu_0)^2}{2\sigma^2}\right] - T\log\frac{1}{\sqrt{2\pi}\sigma} + \frac{T\mu_0^2}{2\sigma^2} + \frac{1}{2\sigma^2}\sum_{t=1}^{T}\left[x_i(t)^2 - 2x_i(t)\mu_0\right]. \qquad (34)$$

We combine the R.H.S expressions of (31) and (34) to arrive at the full equation for the log-likelihood,

$$\log p(\mathbf{x}(1),\ldots,\mathbf{x}(T)|H_i) = NT\log\frac{1}{\sqrt{2\pi}\sigma} - \frac{T\mu_1^2}{2\sigma^2} + \frac{T\mu_0^2}{2\sigma^2} - \sum_{j=1}^{N}\sum_{t=1}^{T}\left[\frac{(x_j(t)-\mu_0)^2}{2\sigma^2}\right]$$

$$- \frac{1}{2\sigma^2}\sum_{t=1}^{T}\left[x_i(t)^2 - 2x_i(t)\mu_1\right] + \frac{1}{2\sigma^2}\sum_{t=1}^{T}\left[x_i(t)^2 - 2x_i(t)\mu_0)\right]. \qquad (35)$$

Rearranging the last two terms on the R.H.S. of equation (35), and setting the first four terms on the R.H.S of equation (35) to C, for legibility, gives,

$$\log p(\mathbf{x}(1),\ldots,\mathbf{x}(T)|H_i) = C + g\sum_{t=1}^{T}x_i(t), \qquad (36)$$

where $g = (\mu_1 - \mu_0)/\sigma^2$. Substituting back into equation (28) gives,

$$\log p(H_i|\mathbf{x}(1),\ldots,\mathbf{x}(T)) = C + g\sum_{t=1}^{T}x_i(t) - \log\left(\exp(C)\sum_{j=1}^{N}\exp\left(g\sum_{t=1}^{T}x_j(t)\right)\right). \qquad (37)$$

The C terms cancel, and by defining the accumulator variables, $y_i(T) = g\sum_{t=1}^{T}x_i(t)$ we arrive at equation (1),

$$\log p(H_i|\mathbf{x}(1),\ldots,\mathbf{x}(T)) = y_i(T) - \log\sum_{j=1}^{N}\exp y_j(T).$$

References

Larry F Abbott and Frances S Chance. Drivers and modulators from push-pull and balanced synaptic input. *Progress in brain research*, 149:147–155, 2005.

Carl W. Baum and Venugopal V. Veeravalli. A Sequential Procedure for Multihypothesis Testing. *IEEE Transactions on Information Theory*, 40(6):1994–2007, 1994.

Joshua D. Berke. What does dopamine mean? *Nature Neuroscience*, 21(6):787–793, 2018.

Rafal Bogacz and Kevin Gurney. The Basal Ganglia and Cortex Implement Optimal Decision Making Between Alternative Actions. *Neural Computation*, 19:442–477, 2007.

Rafal Bogacz, Eric Brown, Jeff Moehlis, Philip Holmes, and Jonathan D. Cohen. The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, 113(4):700–765, 2006.

Rafal Bogacz, Peter T Hu, Philip J Holmes, and Jonathan D Cohen. Do humans produce the speed–accuracy trade-off that maximizes reward rate? *Quarterly journal of experimental psychology*, 63(5):863–891, 2010.

Paolo Calabresi, Barbara Picconi, Alessandro Tozzi, Veronica Ghiglieri, and Massimiliano Di Filippo. Direct and indirect pathways of basal ganglia: a critical reappraisal. *Nature neuroscience*, 17(8):1022–1030, 2014.

Karima Chakroun, Antonius Wiehler, Ben Wagner, David Mathar, Florian Ganzer, Thilo Van Eimeren, Tobias Sommer, and Jan Peters. Dopamine regulates decision thresholds in human reinforcement learning in males. *Nature Communications*, pages 1–14, 2023.

G Chevalier, S Vacher, JM Deniau, and M Desban. Disinhibition as a basic process in the expression of striatal functions. i. the striato-nigral influence on tecto-spinal/tecto-diencephalic neurons. *Brain research*, 334(2):215–226, 1985.

Gilles Chevalier and Jean Michel Deniau. Disinhibition as a basic process in the expression of striatal functions. *Trends in neurosciences*, 13(7):277–280, 1990.

Anne K. Churchland, Roozbeh Kiani, and Michael N. Shadlen. Decision-making with multiple alternatives. *Nature Neuroscience*, 11(6):693–702, 2008.

Paul Cisek. Cortical mechanisms of action selection: The affordance competition hypothesis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1485):1585–1599, 2007.

Paul Cisek and John F. Kalaska. Neural Mechanisms for Interacting with a World Full of Action Choices. *Annual Review of Neuroscience*, 33(1):269–298, 2010.

Anne G E Collins. Advances in modeling learning and decision-making in neuroscience. *Neuropsychopharmacology*, (April):1–15, 2021.

Anne G E Collins and Michael J. Frank. Opponent Actor Learning: Modeling Interactive Effects of Striatal Dopamine on Reinforcement Learning and Choice Incentive. *Psychological Review*, 121(3):337–366, 2014.

Long Ding and Joshua I Gold. Caudate Encodes Multiple Computations for Perceptual Decisions. 30(47):15747–15759, 2010.

Long Ding and Joshua I Gold. The Basal Ganglia' s Contributions to Perceptual Decision Making. *Neuron*, 79(4): 640–649, 2013.

Jan Drugowitsch, Ruben Moreno-Bote, Anne K. Churchland, Michael N. Shadlen, and Alexandre Pouget. The Cost of Accumulating Evidence in Perceptual Decision Making. *Journal of Neuroscience*, 32(11):3612–3628, 2012.

Kyle Dunovan and Timothy Verstynen. Believer-Skeptic meets actor-critic: Rethinking the role of basal ganglia pathways during decision-making and reinforcement learning. *Frontiers in Neuroscience*, 10(MAR):1–15, 2016.

Ido Erev. Signal Detection by Human Observers: A Cutoff Reinforcement Learning Model of Categorization Decisions under Uncertainty. *Psychological Review*, 105(2):280–298, 1998. ISSN 0033295X.

Laura Fontanesi, Sebastian Gluth, Mikhail S. Spektor, and Jörg Rieskamp. A reinforcement learning diffusion decision model for value-based decisions. *Psychonomic Bulletin & Review*, 26(4):1099–1121, 2019a.

Laura Fontanesi, Stefano Palminteri, and Maël Lebreton. Decomposing the effects of context valence and feedback information on speed and accuracy during reinforcement learning: a meta-analytical approach using diffusion decision modeling. *Cognitive, Affective and Behavioral Neuroscience*, 19(3):490–502, 2019b.

Birte U Forstmann, Alfred Anwander, Andreas Schäfer, Jane Neumann, Scott Brown, and Eric-jan Wagenmakers. Cortico-striatal connections predict control over speed and accuracy in perceptual decision making. *PNAS*, 107(36): 15916–15920, 2010.

Michael J Frank. Hold your horses: A dynamic computational role for the subthalamic nucleus in decision making. *Neural Networks*, 19(8):1120–1136, oct 2006.

Michael J. Frank. Hold Your Horses: Impulsivity, Deep Brain Stimulation, and Medication in Parkinsonism. *Science*, 1309(November):1309–1312, 2007.

Michael J Frank and Eric D Claus. Anatomy of a decision: Striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychological Review*, 113(2):300–326, 2006.

Michael J. Frank, Ahmed A. Moustafa, Heather M. Haughey, Tim Curran, and Kent E. Hutchison. Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings of the National Academy of Sciences of the United States of America*, 104(41):16311–16316, 2007.

Michael J. Frank, C. Gagne, E. Nyhus, S. Masters, T. V. Wiecki, J. F. Cavanagh, and D. Badre. fMRI and EEG Predictors of Dynamic Decision Parameters during Human Reinforcement Learning. *Journal of Neuroscience*, 35(2): 485–494, 2015.

Joshua I Gold and Michael N. Shadlen. Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions, and reward. *Neuron*, 36(2):299–308, 2002.

Joshua I Gold and Michael N. Shadlen. The Neural Basis of Decision Making. *Annual Review of Neuroscience*, 30(1): 535–574, 2007.

Thom Griffith, Sophie-anne Baker, and Nathan F Lepora. The statistics of optimal decision making: Exploring the relationship between signal detection theory and sequential analysis. *Journal of Mathematical Psychology*, 103: 102544, 2021.

Kevin Gurney, T. J. Prescott, and P. Redgrave. A computational model of action selection in the basal ganglia. II. Analysis and simulation of behaviour. *Biological Cybernetics*, 84(6):411–423, 2001.

Timothy D. Hanks and Christopher Summerfield. Perceptual Decision Making in Rodents, Monkeys, and Humans. *Neuron*, 93(1):15–31, 2017.

Andrew Heathcote. Winner Takes All! What Are Race Models, and Why and How Should Psychologists Use Them? *Current Directions in Psychological Science*, 2022.

Richard P. Heitz. The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience*, 8(8 JUN):1–19, 2014.

Alana Jaskir and Michael J Frank. On the normative advantages of dopamine and striatal opponency for learning and choice. *eLife*, 12:e85107, mar 2023. ISSN 2050-084X.

Michael D Johnson, Allison S Hyngstrom, Marin Manuel, and CJ Heckman. Push–pull control of motor output. *Journal of Neuroscience*, 32(13):4592–4599, 2012.

Arash Khodadadi, Pegah Fakhari, and Jerome R Busemeyer. Learning to allocate limited time to decisions with different expected outcomes. *Cognitive Psychology*, 95:17–49, 2017.

Ian Krajbich and Antonio Rangel. Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences*, 108(33), 2011.

Birgit Kriener, Rishidev Chaudhuri, and Ila R Fiete. Robust parallel decision-making in neural circuits with nonlinear inhibition. *PNAS*, 117(41), 2020.

Nathan F Lepora. Threshold Learning for Optimal Decision Making. *Nips*, pages 3756–3764, 2016.

Nathan F Lepora and Kevin Gurney. The Basal Ganglia Optimize Decision Making over General Perceptual Hypotheses. *Neural Computation*, 24:2924–2945, 2012.

Chung-chuan Lo and Xiao-jing Wang. Cortico – basal ganglia circuit mechanism for a decision threshold in reaction time tasks. *Nature neuroscience*, 9(7):956–963, 2006.

Tyler McMillen and Philip Holmes. The dynamics of choice among multiple alternatives. *Journal of Mathematical Psychology*, 50(1):30–57, 2006.

Steven Miletić, Russell J. Boag, and Birte U. Forstmann. Mutual benefits: Combining reinforcement learning with sequential sampling models. *Neuropsychologia*, 136:107261, 2020.

Steven Miletić, Russell J. Boag, Anne C. Trutti, Niek Stevenson, Birte U. Forstmann, and Andrew Heathcote. A new model of decision processing in instrumental learning tasks. *eLife*, 10:1–55, 2021.

Moritz Moeller, Sanjay Manohar, and Rafal Bogacz. Uncertainty–guided learning with scaled prediction errors in the basal ganglia. *PLOS Computational Biology*, 18(5):e1009816, 2022.

Moritz Möller and Rafal Bogacz. Learning the payoffs and costs of actions. *PLoS Computational Biology*, 15(2):1–32, 2019.

In Jae Myung and Jerome R. Busemeyer. Criterion Learning in a Deferred Decision-Making Task. *The American Journal of Psychology*, 102(1):1, 1989. ISSN 00029556.

Mads Lund Pedersen, Michael J Frank, and Guido Biele. The drift diffusion model as the choice rule in reinforcement learning. *Psychonomic Bulletin and Review*, (24):1234–1251, 2017.

Roger Ratcliff. A theory of memory retrieval. *Psychological Review*, 85(2):59–108, 1978.

Roger Ratcliff, Philip L. Smith, Scott D. Brown, and Gail McKoon. Diffusion Decision Model: Current Issues and History. *Trends in Cognitive Sciences*, 20(4):260–281, 2016.

P. Redgrave, T. J. Prescott, and Kevin Gurney. The basal ganglia: A vertebrate solution to the selection problem? *Neuroscience*, 89(4):1009–1023, 1999.

David K Sewell, Hayley K Jach, Russell J Boag, and Christina A Van Heer. Combining error-driven models of associative learning with evidence accumulation models of decision-making. *Psychonomic Bulletin & Review*, 26: 868–893, 2019.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. *31st International Conference on Machine Learning, ICML 2014*, 1:605–619, 2014.

Patrick Simen, Jonathan D. Cohen, and Philip Holmes. Rapid decision threshold modulation by reward rate in a neural network. *Neural Networks*, 19(8):1013–1026, 2006. ISSN 08936080.

Satohiro Tajima, Jan Drugowitsch, Nisheet Patel, and Alexandre Pouget. Optimal policy for multi-alternative decisions. *Nature Neuroscience*, 22:1503–1511, 2019.

David Thura and Paul Cisek. The Basal Ganglia Do Not Select Reach Targets but Control the Urgency of Commitment. *Neuron*, 95(5):1160–1170.e5, 2017.

Hado van Hasselt and Marco A. Wiering. Reinforcement learning in continuous action spaces. In *2007 IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning*, pages 272–279, 2007.

Don van Ravenzwaaij, Scott D. Brown, A. A.J. Marley, and Andrew Heathcote. Accumulating Advantages: A New Conceptualization of Rapid Multiple Choice. *Psychological Review*, 127(2):186–215, 2020.

Venugopal V. Veeravalli and Carl W. Baum. Asymptotic Efficiency of A Sequential Multihypothesis Test. *IEEE Transactions on Information Theory*, 41(6):1994–1997, 1995.

A Wald. Sequential Tests of Statistical Hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945.

Ronald J. Williams. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Machine Learning*, 8(3):229–256, 1992.

Kareem A Zaghloul, Christoph T Weidemann, Bradley C Lega, Jurg L Jaggi, Gordon H Baltuch, and Michael J Kahana. Neuronal activity in the human subthalamic nucleus encodes decision conflict during action selection. *Journal of Neuroscience*, 32(7):2453–2460, 2012.

Baltazar A Zavala, Anthony I Jang, and Kareem A Zaghloul. Human subthalamic nucleus activity during non-motor decision making. *eLife*, 6:e31007, dec 2017. ISSN 2050-084X.