OPEN ACCESS Freely available online

PLOS ONE

Reward Optimization in the Primate Brain: A Probabilistic Model of Decision Making under Uncertainty

Yanping Huang, Rajesh P. N. Rao*

Department of Computer Science and Engineering, University of Washington, Seattle, Washington, United States of America

Abstract

A key problem in neuroscience is understanding how the brain makes decisions under uncertainty. Important insights have been gained using tasks such as the random dots motion discrimination task in which the subject makes decisions based on noisy stimuli. A descriptive model known as the drift diffusion model has previously been used to explain psychometric and reaction time data from such tasks but to fully explain the data, one is forced to make ad-hoc assumptions such as a time-dependent collapsing decision boundary. We show that such assumptions are unnecessary when decision making is viewed within the framework of partially observable Markov decision processes (POMDPs). We propose an alternative model for decision making based on POMDPs. We show that the motion discrimination task reduces to the problems of (1) computing beliefs (posterior distributions) over the unknown direction and motion strength from noisy observations in a Bayesian manner, and (2) selecting actions based on these beliefs to maximize the expected sum of future rewards. The resulting optimal policy (belief-to-action mapping) is shown to be equivalent to a collapsing decision threshold that governs the switch from evidence accumulation to a discrimination decision. We show that the model accounts for both accuracy and reaction time as a function of stimulus strength as well as different speed-accuracy conditions in the random dots task.

Citation: Huang Y, Rao RPN (2013) Reward Optimization in the Primate Brain: A Probabilistic Model of Decision Making under Uncertainty. PLoS ONE 8(1): e53344. doi:10.1371/journal.pone.0053344

Editor: Floris P. de Lange, Radboud University Nijmegen, the Netherlands

Received October 12, 2012; **Accepted** November 27, 2012; **Published** January 22, 2013

Funding: This project was supported by the National Science Foundation (NSF) Center for Sensorimotor Neural Engineering (EEC-1028725), NSF grant 0930908, Army Research Office (ARO) award W911NF-11-1-0307, and Office of Naval Research (ONR) grant N000140910097. YH is a Howard Hughes Medical Institute International Student Research fellow. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: rao@cs.washington.edu

Introduction

Animals are constantly confronted with the problem of making decisions given noisy sensory measurements and incomplete knowledge of their environment. Making decisions under such circumstances is difficult because it requires (1) inferring hidden states in the environment that are generating the noisy sensory observations, and (2) determining if one decision (or action) is better than another based on uncertain and delayed reinforcement. Experimental and theoretical studies [1–6] have suggested that the brain may implement an approximate form of Bayesian inference for solving the hidden state problem. However, these studies typically do not address the question of how probabilistic representations of hidden state are employed in action selection based on reinforcement. Daw, Dayan and their colleagues [7,8] explored the suitability of decision theoretic and reinforcement learning models in understanding several well-known neurobiological experiments. Bogacz and colleagues proposed a model that combines a traditional decision making model with reinforcement learning [9] (see also [10]). Rao [11] proposed a neural model for decision making based on the framework of partially observable Markov decision processes (POMDPs) [12]; the model focused on network implementation and learning but assumed a deadline to explain the collapsing decision threshold. Drugowitsch et al. [13] sought to explain the collapsing decision threshold by combining a traditional drift diffusion model with reward rate maximization. Other recent studies have used the general framework of POMDPs to explain experimental data in decision making tasks such as those involving a stop-signal [14,15] and different types of prior knowledge [16].

In this paper, we derive from first principles a POMDP model for the well-known random dots motion discrimination task [17]. We show that the task reduces to the problems of (1) computing beta-distributed beliefs over the unknown direction and motion strength from noisy observations, and (2) selecting actions based on these beliefs in order to maximize the expected sum of future rewards. Without making ad-hoc assumptions such as a hypothetical deadline, a collapsing decision threshold emerges naturally via expected reward maximization. We present results comparing the model's predictions to experimental data and show that the model can explain both reaction time and accuracy as a function of stimulus strength as well as different speed-accuracy conditions.

Methods

POMDP framework

We model the random dots motion discrimination task as a POMDP. The POMDP framework assumes that at any particular time step, the environment is in a particular *hidden* state, μ, that is not directly accessible to the animal. This hidden state however can be inferred by making a sequence of sensory measurements. At each time step t, the animal receives a sensory measurement

(observation), o_t, from the environment, which is determined by an *emission* probability distribution $P(o_t|\mu)$. Since the hidden state μ is unknown, the animal must maintain a *belief* (posterior probability distribution) over the set of possible states given the sensory observations seen so far: $b_t(\mu|o_{1:t})$, where $o_{1:t}$ represents the sequence of observations that the animal has accumulated so far. At each time step, an action (decision) $a_t \in \mathcal{A}$ made by the animal can affect the environment by changing the current state to another according to a *transition* probability distribution $P(\mu'|\mu,a_t)$ where μ is the current state, and μ' is a new state. The animal then gets a reward $R(\mu,a_t)$ from the environment, depending on the current state and the action taken. During training, the animal learns a policy, $\pi(b) \in \mathcal{A}$, which indicates which action a to perform for each belief state b. We make two main assumptions in the POMDP model. First, the animal uses Bayes rule to update its belief about the hidden state after each new observation o_{t+1}: $P(\mu|o_{1:t+1}) = \frac{P(\mu|o_{1:t}) \times P(o_{t+1}|\mu)}{P(o_{t+1}|o_{1:t})}$. Second, the animal is trained to follow an *optimal policy* $\pi^*(b)$ that maximizes the animal's expected total future reward in the task. Figure 1 illustrates the decision making process using the POMDP framework. Note that in the decision making tasks that we model in this paper, the hidden state μ is fixed by experimenters within a trial and thus there is no transition distribution to include in the belief update equation. In general, the hidden state in a POMDP model follows a Markov chain, making the observations $o_{1:t}$ temporally correlated.

Random dots task as a POMDP

We now describe how the general framework of POMDPs can be applied to the random dots motion discrimination task as shown in Figure 1. In each trial, experimenter chooses a fixed direction $d \in \{-1,+1\}$ corresponding to leftward and rightward motion respectively, and a stimulus strength (motion coherence) $c \in [0,1]$, where 0 corresponds to completely random motion and 1 corresponds to 100% coherent motion (i.e., all dots moving in the same direction). Intermediate values of c represent a corresponding fraction of dots moving in the coherent direction (e.g., 0.5 represents 50% coherent motion). The animal is shown a movie of randomly moving dots, a fraction c of which are moving in the same direction d.

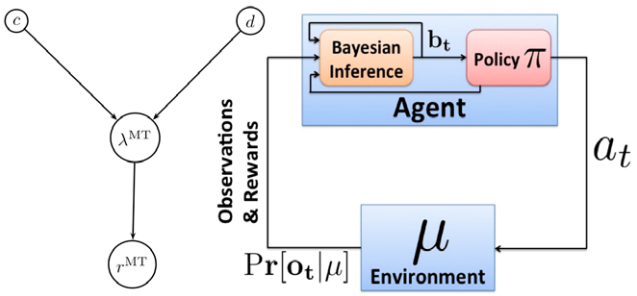

Figure 1. POMP Framework for Decision Making. *Left:* The graphical model representing the probabilistic relationship between random variables c, d, λ and r. In the POMDP model, the hidden state μ corresponds to coherence c and direction d jointly. The observation o_t corresponds to MT response $r^{MT}(t)$. The relations between these variables are summarized in table 1. *Right:* In order to solve a POMDP problem, the animal maintains a belief b_t, which is a posterior probability distribution over hidden states $\mu =$ of the world given observations $o_{1:t}$. At a current belief state b_t, an action is selected according to the learned policy π, which maps belief states to actions.
doi:10.1371/journal.pone.0053344.g001

In a given trial, neither the direction d nor the coherence c is known to the animal. We therefore regard (c,d) as the joint hidden environment state μ in the POMDP model. Neurophysiological evidence suggests that information regarding random dot motion is received from neurons in cortical area MT [18–21]. Therefore, following previous models (e.g., [22–24]), we define the observation model $P(o_t|\mu)$ in the POMDP as a function of the responses of MT neurons. Let the firing rate of MT neurons preferring rightward and leftward direction be λ_R^{MT} and λ_L^{MT} respectively. We can define:

$$\lambda_R^{\mathrm{MT}}(c,d) = \rho_{\mathrm{pref}} \frac{d+1}{2} c + \rho_{\mathrm{null}} \frac{1-d}{2} c + \lambda_0^{\mathrm{MT}}$$

$$\lambda_L^{\mathrm{MT}}(c,d) = \rho_{\mathrm{pref}} \frac{1-d}{2} c + \rho_{\mathrm{null}} \frac{1+d}{2} c + \lambda_0^{\mathrm{MT}} \quad (1)$$

where $\lambda_0^{\mathrm{MT}} = 20$ spikes/second is the average spike rate for 0% coherent motion stimulus, and $\rho_{\mathrm{pref}} = 40$ and $\rho_{\mathrm{null}} = -20$ are the "drive" in the preferred and null directions respectively. These constants (ρ_{pref}, ρ_{null} and λ_0^{MT}) are based on fits to experimental data as reported in [23,25]. Let τ_t be the elapsed time between time steps t and $t+1$. Then, the number of spikes emitted by MT neurons r^{MT} within τ_t follows a Poisson distribution:

$$\Pr[r^{\mathrm{MT}}] = \frac{e^{-\lambda^{\mathrm{MT}} \tau_t} (\lambda^{\mathrm{MT}} \tau_t)^{r^{\mathrm{MT}}}}{r^{\mathrm{MT}}!}. \quad (2)$$

We define the observation o_t at time t as the spike count from MT neurons preferring rightward motion, given the total spike count from rightward and leftward-preferring neurons, i.e., the observation is a conditional random variable $o_t = r_R^{\mathrm{MT}} | n_t$ where $n_t = r_R^{\mathrm{MT}} + r_L^{\mathrm{MT}}$. Then o_t follows a stationary Binomial distribution $\mathrm{Bino}(n,\mu)$. Note that the duration of each POMDP time step need not be fixed, and we can therefore adjust τ_t such that $n_t = n$ for some fixed n, i.e., the animal updates the posterior distribution over hidden state each time it receives n spikes from the MT population. τ_t is exponentially distributed, and the standard deviation of τ_t will approach zero as n increases. When $n=1$, o_t becomes an indicator random variable representing whether a spike was emitted by a rightward motion preferring neuron or not.

It can be shown [26] that o_t follows a Binomial distribution $\mathrm{Bino}(n,\mu)$ with

$$\mu = \frac{\lambda_R^{\mathrm{MT}}}{\lambda_R^{\mathrm{MT}} + \lambda_L^{\mathrm{MT}}} = \frac{\rho_{\mathrm{pref}} \frac{d+1}{2} c + \rho_{\mathrm{null}} \frac{1-d}{2} c + \lambda_0^{\mathrm{MT}}}{(\rho_{\mathrm{pref}} + \rho_{\mathrm{null}}) c + 2\lambda_0^{\mathrm{MT}}} \quad (3)$$

$\mu \in [0,1]$ represents the probability that the MT neurons favoring rightward movement will spike given that there is a spike in the MT population. Since μ is a joint function of c and d, we could equivalently regard it as the hidden state of our POMDP model: $\mu > 0.5$ indicates rightward direction ($d=+1$) while $\mu < 0.5$ indicates the opposite direction ($d=-1$). The coherence $c=0$ corresponds to $\mu = 0.5$ while $c=1$ corresponds to the two extreme values $\mu = 0$ or 1 for direction d being left or right respectively. Note that both direction d and coherence c are unknown to the animal in the experiments, but they are held constant within a trial.

Bayesian inference of hidden state

Given the framework above, the task of deciding the direction of motion of the coherently moving dots is equivalent to the task of deciding whether $d=1$ or not, and deciding when to make such a decision. The POMDP model makes decisions based on the "belief" state $b_t(\mu)=P(\mu|o_{1:t})$, which is the posterior probability distribution over $\mu=\frac{cd+1}{2}$ given a sequence of observations $o_{1:t}$:

$$\begin{aligned} b_t(\mu) &= \frac{\Pr[o_t|\mu]\Pr[\mu|o_{1:t-1}]}{\Pr[o_t|o_{1:t-1}]} \\ &= \frac{\mu^{m_R(t)}(1-\mu)^{m_L(t)}\Pr[\mu]}{Pr[o_{1:t}]} \end{aligned} \tag{4}$$

where $m(t)=\sum_t n_t = n*t$, $m_R(t)=\sum_{\tau=1}^{t} o_\tau$, and $m_L(t)=m(t)-m_R(t)$. To facilitate the analysis, we represent the prior probability $\Pr[\mu]$ as a beta distribution with parameters α_0 and β_0. Note that the beta distribution is quite flexible: for example, a uniform prior can be obtained using $\alpha_0=\beta_0=1$. Without loss of generality, we will fix $\alpha_0=\beta_0=1$ throughout this paper. The posterior distribution can now be written as:

$$\begin{aligned} b_t(\mu) &\propto \mu^{m_R+\alpha_0-1}(1-\mu)^{m_L+\beta_0-1} \\ &= \mathrm{B}eta[\mu|\alpha=m_R+\alpha_0,\beta=m_L+\beta_0] \end{aligned} \tag{5}$$

The belief state b_t at time step t thus follows a beta distribution with two parameters α and β as defined above. Consequently, the posterior probability distribution over μ depends only on the number of spikes m_R and m_L for rightward and leftward motion respectively. These in turn determine $\hat{\mu}$ and t, where

$$\hat{\mu}=\frac{m_R+\alpha_0}{m_R+m_L+\alpha_0+\beta_0} \tag{6}$$

is the point estimator of μ, and $t=\frac{m_R+m_L}{n}$. The animal only needs to keep track of $\hat{\mu}$ and t in order to encode the belief state $b_t=\mathrm{B}eta[\mu|\alpha=\hat{\mu}(nt+\alpha_0+\beta_0),\beta=(1-\hat{\mu})(nt+\alpha_0+\beta_0)]$. After marginalizing over coherence c, we have the posterior probability over direction d:

$$\Pr[d=1|o_{1:t}]=\int_{\mu=0.5}^{1}\mathrm{B}eta(\mu|\alpha,\beta)d\mu=1-I_{0.5}(\alpha,\beta) \tag{7}$$

$$\Pr[d=-1|o_{1:t}]=\int_{\mu=0}^{0.5}\mathrm{B}eta(\mu|\alpha,\beta)d\mu=I_{0.5}(\alpha,\beta). \tag{8}$$

where $I_x(\alpha,\beta)=\int_{\mu=0}^{x}\mathrm{B}eta(\mu|\alpha,\beta)d\mu$ is the regularized incomplete beta function.

Actions, rewards, and value function

The animal updates its belief after receiving the current observation o_t, and chooses one of the three actions (decisions) $a\in\{A_R,A_L,A_S\}$, denoting rightward eye movement, leftward eye movement, and sampling (i.e., waiting for one more observation) respectively. The model assumes the animal receives rewards $R(\mu,a)$ as follows (rewards are modeled using real numbers). When the animal makes a correct choice, *i.e.*, a rightward eye movement A_R when $d=1$ ($\mu>1/2$) or a leftward eye movement A_L when $d=-1$ ($\mu<1/2$), the animal receives a positive reward $R_P>0$. The animal receives a negative reward (i.e., penalty) or nothing when an incorrect action is chosen $R_N\leq 0$. We further assume that the animal is motivated by hunger or thirst to make a decision as quickly as possible. This is modeled using a unit penalty $R_S=-1$ for each observation the animal makes, representing the cost the animal needs to pay when choosing the sampling action A_S.

Recall that a belief state b_t is determined by the parameters α,β. The goal of the animal is to find an optimal "policy" π^* that maximizes the "value" function $v^\pi(b_t)$, defined as the expected sum of future rewards given the current belief state:

$$v^\pi(b_t)=\mathrm{E}[\sum_{k=1}^{\infty}R(b_{t+k},\pi(b_{t+k}))|b_t=\mathrm{B}eta(\mu|\alpha,\beta)] \tag{9}$$

where the expectation is taken with respect to all future belief states $(b_{t+1},\ldots,b_{t+k},\ldots)$. The reward term $R(b_t,a)$ above is the expected reward for the given belief state and action:

$$R(b_t,A_S)=nR_S \tag{10}$$

$$\begin{aligned} R(b_t,A_R) &= \sum_d\int_{c=0}^{1}R(c,d,A_R)\mathrm{Beta}(\mu|\alpha,\beta)dc \\ &= R_P\times[1-I_{0.5}(\alpha,\beta)]+R_N\times I_{0.5}(\alpha,\beta) \\ &= (R_P-R_N)\times[1-I_{0.5}(\alpha,\beta)]+R_N \\ R(b_t,A_L) &= (R_P-R_N)\times I_{0.5}(\alpha,\beta)+R_N \end{aligned}$$

The above equations can be interpreted as follows. When A_S is selected, the animal receives n more samples at a cost of nR_S. When A_R is selected, the expected reward $R(b_t,A_R)$ depends on the probability density function of the hidden parameter μ given belief state b_t. With probability $I_{0.5}(\alpha,\beta)$, the true parameter μ is less than 0.5, making A_R an incorrect decision with penalty R_N, and with probability $1-I_{0.5}(\alpha,\beta)$, action A_R is correct, earning the reward R_P.

Finding the optimal policy

A policy $\pi(b_t)$ defines a mapping from a belief state to one of the available actions a. A method for learning a POMDP policy by trial and error using the method of temporal difference (TD) learning was suggested in [11]. Here, we derive a policy from first principles and compare the result with behavioral data.

One standard way [12] to solve a POMDP is to first convert it into a Markov Decision Process (MDP) over belief state, and then apply standard dynamical programming techniques such as value iteration [27] to compute the value function in equation 9. For the corresponding *belief MDP*, we need to define the transition probabilities $T(b_t|b_{t-1},a_{t-1})$. When $a_{t-1}=A_S$, the belief state can be updated using the previous belief state and current observation based on Bayes' rule:

$$\begin{aligned} T(b_t|b_{t-1},A_S) &= \Pr[\alpha',\beta'|\alpha,\beta,A_S] \\ &= \Pr[o_t|\alpha,\beta]\delta_{\alpha'=\alpha+o_t}\delta_{\beta'=\beta+n-o_t} \end{aligned} \tag{11}$$

for all $o_t\in\{0,\ldots,n\}$. In the above equation, $\delta(.)$ is the Kronecker delta, and $\Pr[o_t|\alpha,\beta]$ is the expected value of the likelihood function $\Pr[o_t|\mu]=\mu$ over the posterior distribution b_t:

$$\Pr[o_t|\alpha,\beta]=\binom{n}{o_t}\frac{\alpha^{o_t}\beta^{n-o_t}}{(\alpha+\beta)^n}, \quad (12)$$

which is a stationary distribution independent of time t. When the selected action is A_R or A_L, the animal stops sampling and makes an eye movement. To account for such cases, we include an additional state Γ, representing a terminal state, with zero reward $R(\Gamma,a)=0$ and absorbing behavior, $T(\Gamma|\Gamma,a)=1$ for all actions a. Formally, the transition probabilities with respect to the absorbing (termination) state are defined as $\Pr[\Gamma|b_t,a\in\{A_R,A_L\}]=1$ for all b_t, indicating the end of a trial.

Given the time-independent belief state transition $\Pr\left[b_t'|b_t,a\right]$, the optimal value $v^\star$ and policy $\pi^\star=\arg\max_\pi v^\pi$ can be obtained by solving Bellman's equation:

$$\pi^\star(b_t)=\underset{a}{argmax}[R(b_t,a)+\sum_{b_t'}\Pr\left[b_t'|b_t,a\right]v^*(b_t')]$$

$$v^\star(b_t)=\max_a[R(b_t,a)+\sum_{b_t'}\Pr\left[b_t'|b_t,a\right]v^*(b_t')] \quad (13)$$

Before we proceed to results from the model, we note that the one-step belief transition probability matrix $T(b_t|b_{t-1},A_S)$ with $n=n_0$ can be shown be mathematically equivalent to the n_0-steps transition matrix $T^{n_0}(b_t|b_{t-1},A_S)$ with $n=1$. The solution to Bellman's equation 13 is independent of n. Therefore, unless otherwise mentioned, the results are based on the most general scenario where the animal needs to select an action whenever a new spike is received, *i.e.*, $n=1$.

We summarize the model variables as well as their statistical relationships in table 1.

Results

Optimal value function and policy

Figure 2 (a) shows the optimal value function computed by applying value iteration [27] to the POMDP defined in the Methods and Analysis section, with parameters $R_P=50$, $R_N=0$, and $R_S=-0.1$. The x-axis of Figure 2 (a) represents the total number of observations $m=m_R+m_L$ encountered thus far, which is equal to the elapsed time t in the trial. The y-axis represents the ratio $\hat{\mu}=\frac{m_R+\alpha_0}{m+\alpha_0+\beta_0}$, which is the estimator of the hidden parameter μ. In general, the model predicts a high value when $\hat{\mu}$ is close to 1 or 0, or equivalently, when the estimated coherence is close to 1. This is because at these two extremes, selecting the appropriate action has a high probability of receiving a large positive reward R_P. On the other hand, for $\hat{\mu}$ near 0.5 (estimated c near 0), choosing A_L or A_R in these states has a high chance of resulting in an incorrect decision and a large negative reward R_N (see [11] for a similar result using a different model and under the assumption of a deadline). Thus, belief states with $m_R\sim m_L$ have a much lower value compared to belief states with $m_R\gg m_L$ or $m_R\ll m_L$.

Figure 2 (b) shows the corresponding optimal policy $\pi^\star$ as a joint function of $\hat{\mu}$ and t. The optimal policy $\pi^\star$ partitions the belief space into three regions: Π^R, Π^L, and Π^S, representing the set of belief states preferring actions A_R, A_L and A_S respectively. Let Π^a_m be the set of belief states preferring action a after m observations, for $a\in\{A_R,A_L,A_S\}$ and $m=m_R+m_L$. Early in a trial, when m is small, the model selects the sampling action A_S regardless of the value of $\hat{\mu}$. This is because for small m, the variance of the point estimator $\hat{\mu}(m)$ is high. For example, even when $\hat{\mu}=1$ when $m=2$, the probability that the true $\mu<0.5$ is still high. The sampling action A_S is required to reduce this variance by accruing more evidence. As m becomes larger, the variance of $\hat{\mu}$ decreases, and the deviation between $\hat{\mu}$ and the true value of μ diminishes by the law of large numbers. Consequently, the animal will pick action A_R even when $\hat{\mu}$ is only slightly above 0.5. This gradual decrease in the threshold over time for choosing the overt actions A_R or A_L has been called a "collapsing bound" in the decision making literature [28–30].

The optimal policy $\pi^\star$ is entirely determined by three reward parameters $\{R_P,R_N,R_S\}$. At a given belief state, $\pi^\star$ picks one of the three available actions that leads to the largest expected future reward. Thus, the choice is determined by the relative, not the absolute, value of the expected future reward for the different actions. From equation 10, we have

$$R(\alpha,\beta,A_L)-R(\alpha,\beta,A_R)\propto R_N-R_P. \quad (14)$$

If we regard the sampling penalty R_S as specifying the unit of reward, the optimal policy $\pi^\star$ is determined by the ratio $\frac{R_N-R_P}{R_S}$ alone. Figure 2 (c) shows the relationship between $\frac{R_N-R_P}{R_S}$ and the optimal policy $\pi^\star$ by showing the rightward decision boundaries $\phi^R(t)$ for different values of $\frac{R_N-R_P}{R_S}$. As $\frac{R_N-R_P}{R_S}$ increases (e.g., by making the sampling cost R_s smaller), the boundary $\phi^R(t)$ gradually moves towards the upper right corner, giving the animal more time to make decisions which results in more accurate decisions. To better understand this relationship, we fit the decision boundary to a hyperbolic function:

$$\phi^R(t)-0.5\propto\frac{t}{t+\tau_{1/2}} \quad (15)$$

We find that $\tau_{1/2}$ exhibits nearly logarithmic growth with $\frac{R_N-R_P}{R_S}$. Interestingly, a collapsing bound is obtained even with extremely small R_S because the goal is reward maximization across trials: it is better to terminate a trial and accrue reward in future trials than to continue sampling noisy (possibly 0% coherent) stimuli.

Model predictions: psychometric function and reaction time

We compare predictions of the model based on the learned policy π^* with experimental data from the reaction time version (rather than the fixed duration version) of the motion discrimination task [31]. As illustrated in Figure 3, the model assumes that motion information regarding the random dots on the screen is processed by MT neurons. These neurons provide the observations o_t (and $n-o_t$) to right- and left-direction coding LIP neurons, which maintain the belief state $b_t=\{\alpha=\sum_t o_t,\beta=\sum_t(n-o_t)\}$. Actions are selected based on the optimal policy π^*. If $b_t\in\Pi^R_t$ or $b_t\in\Pi^L_t$, the animal makes a rightward or leftward decision respectively and terminates the trial. When $b_t\in\Pi^S_t$, the animal chooses the sampling action and gets a new observation o_{t+1}.

Table 1. Summary of model variables and paramters.

POMDP Variables	Descriptions
μ	The hidden variable of POMDP, $\mu=\frac{c\times d+1}{2}\in[0,1]$. In the random dots task, μ is a constant over time
c	The coherence (motion strength) of the random dots task. $c\in[0,1]$. c is fixed during a task.
d	The underlying direction of the random dots task. $d\in\{\pm 1\}$. d is fixed during a task.
$\lambda_{R,L}^{\mathrm{MT}}$	The average spike rate of MT neurons preferring rightward or leftward direction, respectively, as a function of both coherence c and d described in equations 1.
$r_{R,L}^{MT}$	The number of spikes emitted by MT neurons preferring rightward or leftward direction, respectively during one POMDP step. r^{MT} follows a Poisson distribution with mean λ^{MT}
n_t	Total number of spikes emitted by MT neurons during one POMDP step. $n_t=r_R^{MT}+r_L^{MT}$
o_t	The noisy observation at time step t, which is a conditional random variable $o_t=r_R^{MT}\|n_t$ following a Binomial distribution $Bino(n_t,\mu)$. Note that $o_1,\ldots,o_t$ are conditional dependent of each other given the hidden variable μ
b_t	The belief (posterior distribution) $b_t=P(\mu\|o_{1:t})$. With a beta-distributed initial belief $b_0=Beta(\alpha_0,\beta_0)$, b_t is also beta distributed due to the binomial distributed emission probability $P(o_t\|\mu)$. Without loss of generality, $\alpha_0=\beta_0=1$ throughout the paper.
a_t	Action chosen by the animal at time t. $a_t\in\{A_S,A_R,A_L\}$.
Model Parameters	
R_S	A negative reward associated with the cost of an observation.
R_P	A positive reward associated with a correct eye movement.
R_N	A negative reward associated with an incorrect eye movement.
RT_{step}	The duration of a single observation, the real elapsed time per POMDP step. Only used to translate the number of POMDP time steps to real elapsed time when comparing with experimental data.
RT_0	Non-decision residual time. Both RT_{step} and RT_0 are obtained from a linear regression to compare model predictions (in unit of POMDP steps) with animals' response time (in unit of seconds), independent of the POMDP model.

doi:10.1371/journal.pone.0053344.t001

The performance on the task using the optimal policy π^* can be measured in terms of both the accuracy of direction discrimination (the so-called psychometric function), and the reaction time required to reach a decision (the chronometric function). In this section, we derive the expected accuracy and reaction time as a function of stimulus coherence c, and compare them to the psychometric and chronometric functions of a monkey performing the same task [31].

The sequence of random variables $\{\hat{\mu}_1,\hat{\mu}_2,\ldots,\hat{\mu}_t\}$ forms a (non-stationary) Markov chain with transition probabilities determined by equation 11. Let $\Psi(\hat{\mu}_t,t|\mu)$ be the joint probability that the animal keeps selecting A_S until time step t:

$$\Psi(\hat{\mu}_t,t|\mu=Pr\left[\hat{\mu}_1\in\Pi_1^S,\hat{\mu}_2\in\Pi_2^S,\ldots,\hat{\mu}_t\in\Pi_t^S\right]. \tag{16}$$

At $t=0$, the animal will select A_S regardless of $\hat{\mu}$ under π^*, making $\psi(\hat{\mu},0|\mu)=\Pr[\hat{\mu}_0]$. At $t\geq 1$, $\Psi(\hat{\mu}_t,t|\mu)$ can be expressed recursively as:

$$\Psi(\hat{\mu}_t,t|\mu)=\sum_{\hat{\mu}_{t-1}\in\Pi_{t-1}^S}\Pr[\hat{\mu}_t|\hat{\mu}_{t-1}]\Psi(\hat{\mu}_{t-1},t-1|\mu) \tag{17}$$

Let $\Pr[t,R|\mu]$ and $\Pr[t,L|\mu]$ be the joint probability mass functions that the animal makes a right or left choice at time t, respectively. These correspond to the probability that the point estimator $\hat{\mu}(t)$ crosses the boundary of Π^R or Π^L for the first time at time t:

$$\begin{aligned}\Pr[t,R|\mu]&=\Pr\left[\hat{\mu}_t\in\Pi_t^R,\hat{\mu}_{t-1}\in\Pi_{t-1}^S,\ldots,\hat{\mu}_1\in\Pi_1^S|\mu\right]\\&=\sum_{\hat{\mu}_t\in\Pi_t^R}\sum_{\hat{\mu}_{t-1}\in\Pi_{t-1}^S}Pr[\hat{\mu}_t|\hat{\mu}_{t-1}]\Psi(\hat{\mu}_{t-1},t|\mu)\end{aligned} \tag{18}$$

$$\Pr[t,L|\mu]=\sum_{\hat{\mu}_t\in\Pi_t^L}\sum_{\hat{\mu}_{t-1}\in\Pi_{t-1}^S}\Pr[\hat{\mu}_t|\hat{\mu}_{t-1}]\Psi(\hat{\mu}_{t-1},t|\mu) \tag{19}$$

The probabilities of making rightward or leftward eye movement are the marginal probabilities summing over all possible crossing times: $\Pr[R|\mu]=\sum_{t=1}^{\infty}\Pr[t,R|\mu]$ and $\Pr[L|\mu]=\sum_{t=1}^{\infty}\Pr[t,L|\mu]$. When the underlying motion direction is rightward, $\Pr[R|\mu]$ represents the accuracy of motion discrimination and $\Pr[L|\mu]$ represents the error rate. The mean reaction times for correct and error choices are the expected crossing times over the conditional probability that the animal makes decision A_R and A_L respectively at time t:

$$RT_R(\mu)=\sum_{t=1}^{\infty}t\frac{\Pr[t,R|\mu]}{\Pr[R|\mu]} \tag{20}$$

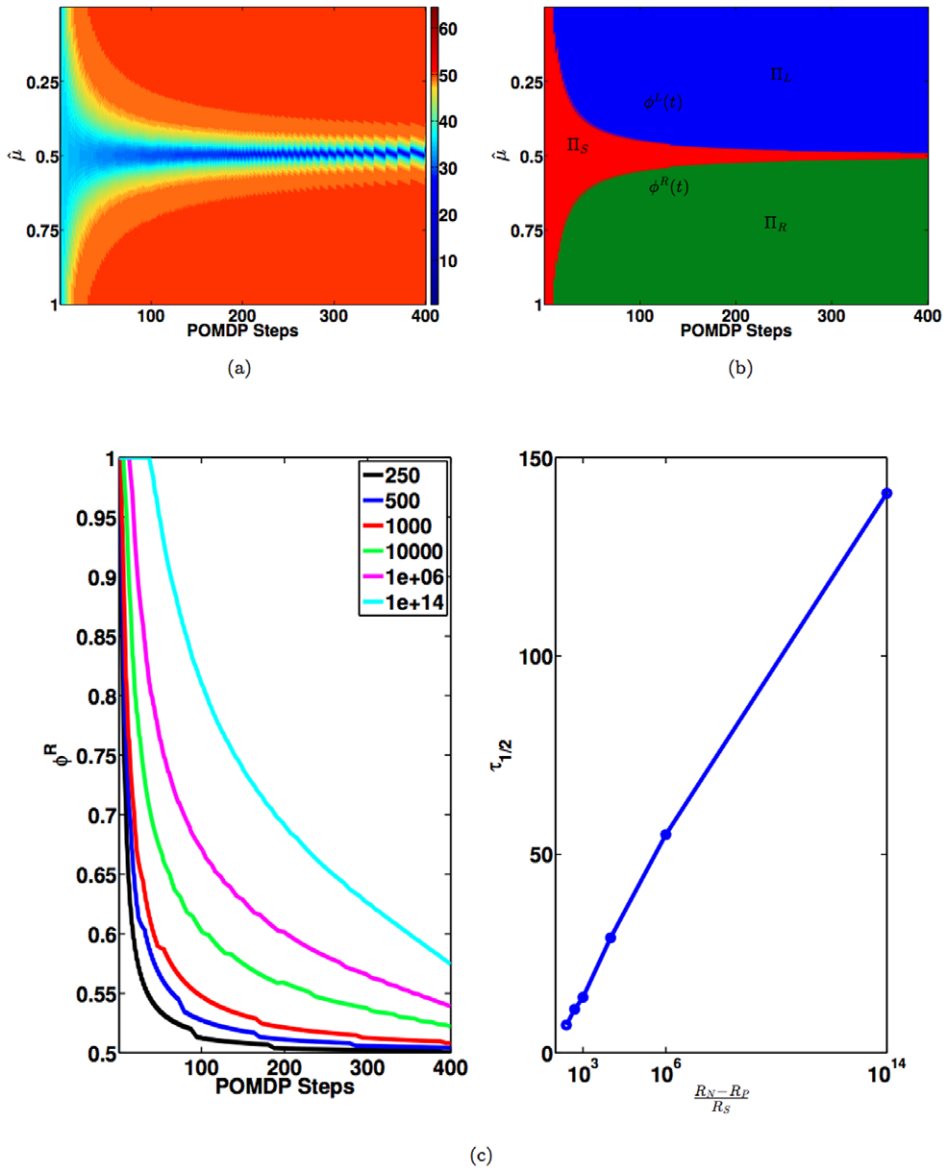

Figure 2. Optimal Value and Policy for the Random Dots Task. (a) Optimal value as a joint function of $\hat{\mu}=\frac{m_R+\alpha_0}{m+\alpha_0+\beta_0}$ and the number of POMDP steps t. (b) Optimal Policy as a function of $\hat{\mu}$ and the number of POMDP steps t. The boundaries $\phi^R(t)$ and $\phi^L(t)$ divide the belief space into three areas: Π_S (red), Π_R (green), and Π_L (blue), each of which represents belief states whose optimal actions are A_S,A_R and A_L respectively. Model parameters: $R_P=50$, $R_S=-0.1$, and $R_N=0$. (c) *Left:* The rightward decision boundary $\phi^R(t)$ for different values of $\frac{R_N-R_P}{R_S}$. *Right:* The half time $\tau_{1/2}$ of $\phi^R(t)$ for different values of $\frac{R_N-R_P}{R_S}$, where $\phi^R(\tau_{1/2})=\frac{\phi^R(0)-\phi^R(\infty)}{2}$.
doi:10.1371/journal.pone.0053344.g002

$$RT_L(\mu)=\sum_{t=1}^{\infty} t\frac{\Pr[t,L|\mu]}{\Pr[L|\mu]} \tag{21}$$

The left panel of Figure 4 shows performance accuracy as a function of motion strength c for the model (solid curve) and a monkey (black dots). The model parameters are the same as those in Figure 2, obtained using a binary search within $R_p\in\{0,2000\}$ with a minimum step size 10.

The right panel of Figure 4 shows for the same model parameters the predicted mean reacton time $RT_R(\mu)$ for correct choices as a function of coherence c (and fixed direction $d=1$) for the model (solid curve) and the monkey (black dots). Note that $RT_R(\mu)$ represents the expected number of POMDP time steps for making a rightward eye movement A_R. It follows from the Poisson spiking process that the duration of each POMDP time step follows a exponential distribution with its expectation proportional to $\lambda_R(\mu)+\lambda_L(\mu)$. In order to make a direct comparison to the monkey data $RT^*_R(\mu)$, which is in units of real time, a linear regression was used to to determine the duration RT_{step} of a single

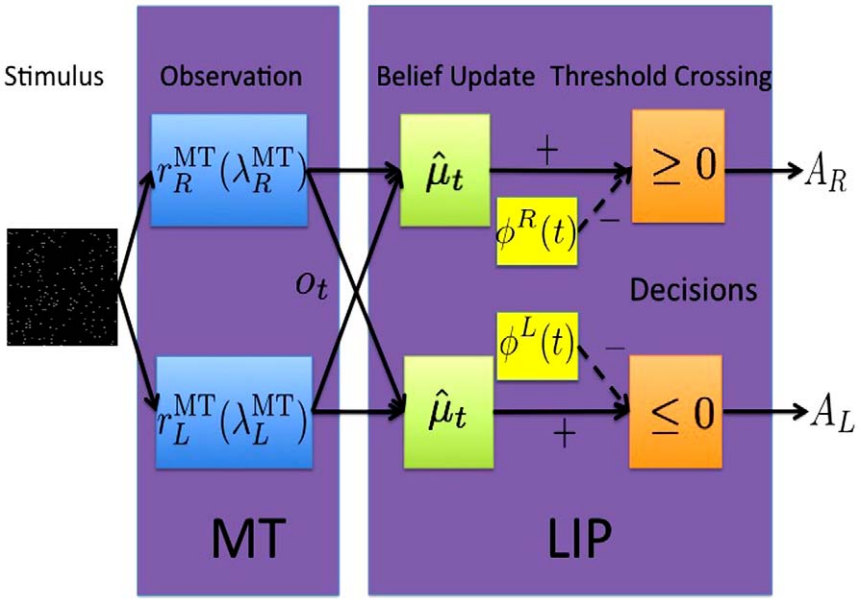

Figure 3. Relationship between Model and Neural Activity. The input to the model is a random dots motion sequence. Neurons in MT with tuning curves λ^{MT} emit r^{MT} spikes at time step t, which constitutes the observation o_t in the POMDP model. The animal maintains the belief state b_t by computing $\hat{\mu}_t$ (b_t can be parameterized by $\hat{\mu}_t$ and t - see text). The optimal policy is implemented by selecting rightward eye movement A_R when $\hat{\mu}_t \geq \phi^R(t)$, or equivalently, when $(\hat{\mu}_t - \phi^R(t)) \geq 0$ (and likewise for leftward eye movement A_L).
doi:10.1371/journal.pone.0053344.g003

observation and the onset of decision time RT_0:

$$RT_R^*(\mu) = RT_{step} * (\lambda_R(\mu) + \lambda_L(\mu)) * RT_R(\mu) + RT_0. \quad (22)$$

Note that the reaction time in a trial is the sum of decision time plus the non-decision delays whose properties are not well understood. The offset RT_0 represents the non-decision residual time. We applied the experimental mean reaction time reported in [31] with motion coherence $c = \{0.032, 0.064, 0.128, 0.256, 0.512\}$

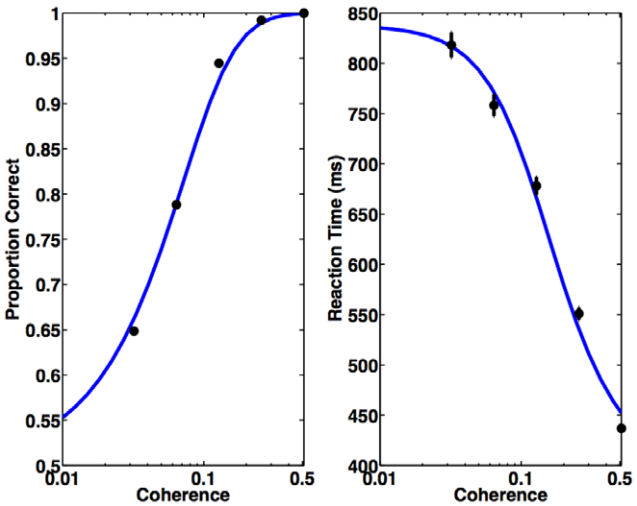

Figure 4. Comparison of Performance of the Model and Monkey. Black dots with error bars represent a monkey's decision accuracy and reaction time for correct trials. Blue solid curves are model predictions ($RT_R(\mu)$ and $RT_R(\mu)$ in the text) for parameter values $R_P = 50, R_S = -0.1$, and $R_N = 0$. Monkey data from [31].
doi:10.1371/journal.pone.0053344.g004

to compute the two coefficients RT_{step} and RT_0. The unit duration per POMDP step $RT_{step} = 9.20$ ms/step, and the offset $RT_0 = 358.5$ ms, which is comparable to the 300 ms non-decision time on average reported in the literature [23,32].

There is essentially one parameter in our model needed to fit the experimental accuracy data, namely, the reward ratio $\frac{R_N - R_P}{R_S}$. The other two parameters RT_{step} and RT_0 are independent of the POMDP model, and are used only to translate the POMDP time steps into real elapsed time. This reward ratio has direct physical interpretation and can be easily manipulated by the experimenters. For example, changing the amount of awards for the correct/incorrect choices, or giving subjects different speed instructions will effectively change $\frac{R_N - R_P}{R_S}$. In Figure 5 (a), we show performance accuracies $\Pr[R|\mu]$ and predicted mean reaction time $RT_R(\mu)$ with different values of $\frac{R_N - R_P}{R_S}$. With fixed R_N and R_P, decreasing R_S makes the observations more affordable and allows subjects to accumulate more evidence, in turn leads to a longer decision time and higher accuracy. Our model thus provides a quantitative framework for predicting the effects of reward parameters on the accuracy and speed of decision making. To test our theory, we compare the model predictions with the experimental data from a human subject, reported by Hanks et al [33], under different speed-accuracy regimes. In their experiments, human subjects were instructed to perform the random dots task under different speed-accuracy conditions. The red crosses in Figure 5 (b) represent the response time and accuracy of a human subject in the direction discrimination task with instructions to perform the task more carefully at a slower speed, while the black dots represent the task under normal speed conditions. The slower speed instruction encourages human subjects to accumulate more observations before making the final decision. In the model, this amounts to reducing the negative cost

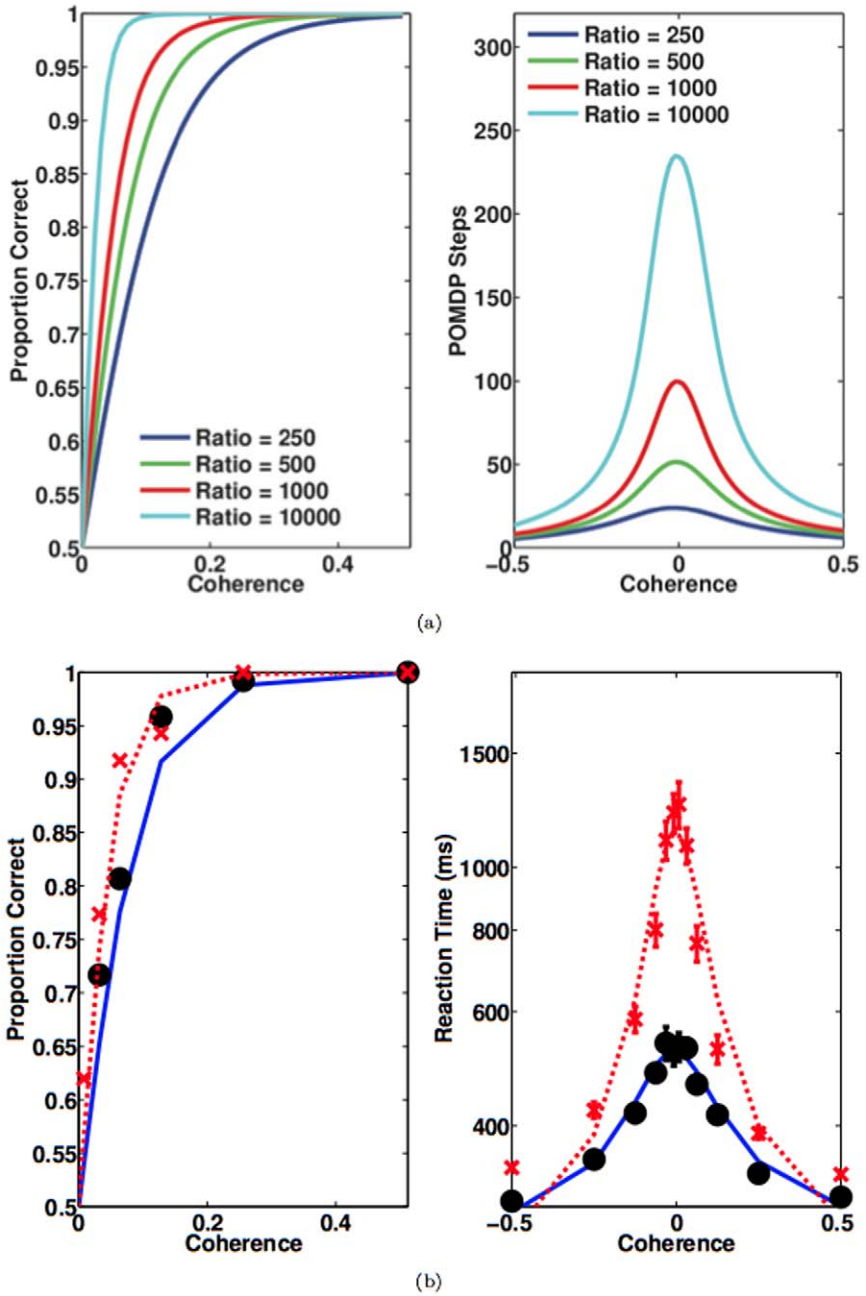

Figure 5. Effect of $\frac{R_N - R_P}{R_S}$ on speed-accuracy tradeoff. (a) Model predictions of psychometric and chronometric functions for different values of $\frac{R_N - R_P}{R_S}$. (b) Comparison of model predictions and experimental data for different speed-accuracy regimes. The black dots represent the response time and accuracy of a human subject in the direction discrimination task under normal speed conditions, while the red crosses represent data with a slower speed instruction. The model predictions are plotted as black solid curves (with $\frac{R_N - R_P}{R_S} = 450$) and red dashed lines ($\frac{R_N - R_P}{R_S} = 1250$), respectively. The per-step duration and non-decision residual time are fixed to be the same for both conditions: $RT_{step} = 7.7$ ms/step, and $RT_0 = 204$ ms. Human data are from human subject LH in [33].
doi:10.1371/journal.pone.0053344.g005

associated with each sample R_s. Indeed, this tradeoff between speed and accuracy was consistent with predicted effects of changing the reward ratio. We first fit the model parameters to experimental data under normal speed conditions, based on fitting $\frac{R_N - R_P}{R_S}$, $RT_{step} = 7.7$ ms/step, and $RT_0 = 204$ ms (Figure 5 (b), black solid curves). The red dashed lines shown in Figure 5 (b) are model fits to the data under slower speed instruction. There is just one degree of freedom in this fit, as all model parameters except the reward ratio were fixed to the values used to fit data in the normal speed regime.

Neural response during direction discrimination task

From Figure 2 (b), it is clear that for the random dots task, the animal does not need to store the whole two dimensional optimal policy but only the two one-dimensional decision boundaries ϕ^R and ϕ^L. This naturally suggests a neural mechanism for decision making similar to that in drift diffusion models: LIP neurons compute the belief state from MT responses and employ divisive normalization to maintain the point estimate $\hat{\mu}_t = \frac{m_R + \alpha_0}{m + \alpha_0 + \beta_0}$. We now explore the hypothesis that the response of LIP neurons represents the difference between $\hat{\mu}$ and the optimal decision threshold $\phi^R(t)$. In this model, a rightward eye movement is initiated only when the difference $\frac{m_R}{m_R + m_L} - \phi^R$ reaches a fixed bound (in this case, 0). Therefore, we modeled the firing rates in the lateral intraparietal area (LIP) λ^{LIP} as:

$$\lambda_R^{LIP}(t) = \lambda_0^{LIP} + \hat{B}\left(\frac{m_R + \alpha_0}{m + \alpha_0 + \beta_0} - \phi^R(t) + \frac{\beta_0}{\alpha_0 + \beta_0}\right) \qquad (23)$$

where λ_0^{LIP} is the spontaneous firing rate for LIP neurons. Since $\phi^R(0) = 1$, a constant $\frac{\beta_0}{\alpha_0 + \beta_0}$ is added to make $\lambda_R^{LIP}(0) = \lambda_0^{LIP}$. $\hat{B}$ represents the termination bound; $\hat{B} = 49.6$ spikes s^{-1} from [30]. The firing rate λ_L^{LIP} is defined similarly.

The above model makes two testable predictions about neural responses in LIP. The first is that the neural response to 0% coherent motion (the so called "urgency" signal [30,34]) encodes the decision boundary $\phi^R(t)$ (or $\phi^L(t)$ for leftward-preferring LIP neurons). In Figure 6a, we plot the model response to 0% coherent motion, along with a fit to a hyperbolic function $u(t) \propto \frac{t}{t + \tau_{1/2}}$, the same function that Churchland et al [30] used to parametrize the experimentally observed "urgency signal." The parameter $\tau_{1/2}$ is the time taken to reach 50% of the maximum. The estimate of $\tau_{1/2}$ for the model from Figure 6 (a) is 123 ms, which is consistent with the $\tau_{1/2} = 133.2$ ms estimated from neural data [30].

The second prediction concerns the buildup rate (in units of spikes s^{-2} coh^{-1}) of the LIP firing rates. The buildup rate of LIP at each motion strength is calculated from the slope of a line fit to model LIP firing rate during the first 120 ms of decision time. As shown in Figure 6 (b), buildup rates scaled approximately linearly as a function of motion coherence. The effect of a unit change in coherence on the buildup rate can be estimated from the slope of the fitted line to be 227.7 spike s^{-2} coh^{-1}, similar to what has been reported in the literature [30] (222.5 spike s^{-2} coh^{-1}).

Discussion

The random dots motion discrimination task has provided a wealth of information regarding decision making in the primate brain. Much of this data has previously been modeled using the drift diffusion model [35,36], but to fully account for the experimental data, one has to sometimes use ad-hoc assumptions. This paper introduces an alternative model for explaining the monkey's behavior based on the framework of partially observable Markov decision processes (POMDPs).

We believe that the POMDP model provides a more versatile framework for decision making compared to the drift diffusion model, which can be viewed as a special case of sequential statistical hypothesis testing (SSHT) [37]. Sequential statistical hypothesis testing assumes that the stimuli (observations) are independent and identically distributed whereas the POMDP model allows observations be temporally correlated. The observations in the POMDP are conditionally independent given the hidden state μ, which evolves according to a Markov chain. Thus, the POMDP framework for decision making [11,14,16,38,39] can be regarded as a strictly more general model than the SSHT models. We intend to explore the applicability of our POMDP

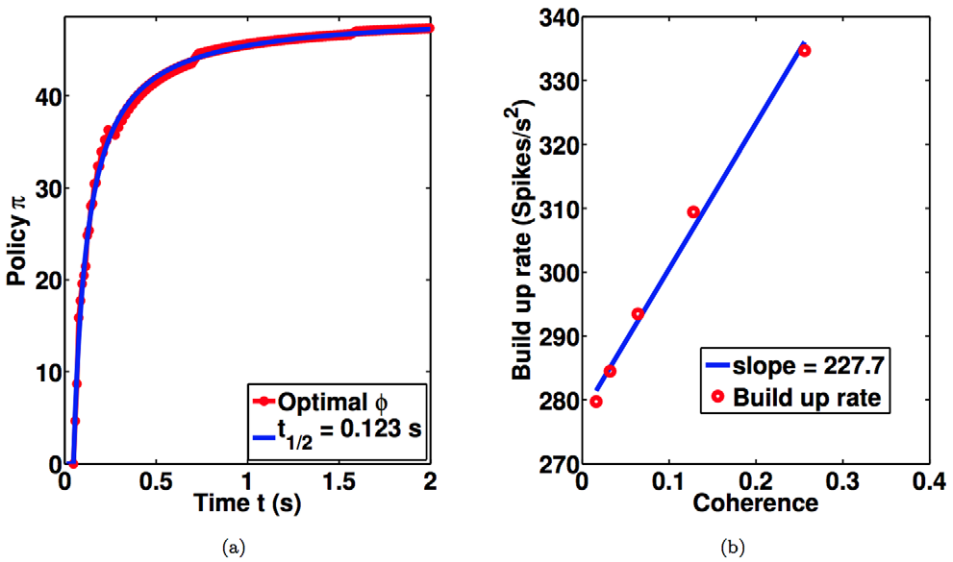

Figure 6. Comparison of Model and Neural Responses. (a) Model response to 0% coherence motion is shown in red. Blue curve depicts a fit using a hyperbolic function $u(t) = u_\infty \frac{t}{t + \tau_{1/2}}$ where $\tau_{1/2} = 123$ ms, which is comparable to the value of 133.2 ms estimated from neural data [30]. (b) The first 120 ms of decision time was used to compute the buildup rate from the model response following the procedure in [30]. The red points show model buildup rates estimated for each coherence value. The effect of a unit change in the coherence on buildup rate can be estimated from the slope of the blue fitted line: this value, 227.7 spike s^{-2} coh^{-1}, is similar to the corresponding value 222.5 spike s^{-2} coh^{-1} estimated from neural data [30].
doi:10.1371/journal.pone.0053344.g006

model to time-dependent stimuli, such as temporally dynamic attention [40] and temporally blurred stimulus representations [41] in future studies.

Another advantage of a POMDP model is that the model parameters have direct physical interpretations and can be easily manipulated by the experimenter. Our analysis shows that the optimal policy is fully determined by the reward parameters $\{R_P,R_N,R_S\}$. Thus, the model psychometric and chronometric functions, which are derived from the optimal policy, are also fully determined by these model parameters. Experimenters can control these reward parameters by changing the amount of awards for the correct/incorrect choices, or by giving subjects different speed instructions. This allows our model to make testable predictions, as demonstrated by the effects of the change in the reward ratios on the speed-accuracy trade-off. It should be noted that these reward parameters can be subjective and may vary from individual to individual. For example, R_P can be directly related to the external food or juice reward provided by the experimenter while R_S may be linked to internal factors such as degree of hunger or thirst, drive, and motivation. The precise relationship between these reward parameters and the external reward/risk controlled by the experimenter remains unknown. Our model thus provides a quantitative framework for studying this relationship between internal reward mechanisms and external physical reward.

The proposed model demonstrates how the monkey's choices in the random dots task can be interpreted as being optimal under the hypothesis of reward maximization. The reward maximization hypothesis has previously been used to explain behavioral data from conditioning experiments [8] and dopaminergic responses under the framework of temporal difference (TD) learning [42]. Our model extends these results to the more general problem of decision making under uncertainty. The model predicts psychometric and chronometric functions that are quantitatively close to those observed in monkeys and humans solving the random dots task.

We showed through analytical derivations and numerical simulation that the optimal threshold for selecting overt actions is a declining function of time. Such a collapsing decision bound has previously been obtained for decision making under a deadline [11,29]. It has also been proposed as an ad-hoc mechanism in drift diffusion models [28,30,43] for explaining finite response time at zero percent coherence. Our results demonstrate that a collapsing bound emerges naturally as a consequence of reward maximization. Additionally, the POMDP model readily generalizes to the case of decision making with arbitrary numbers of states and actions, as well as time-varying state.

Instead of traditional dynamic programming techniques, the optimal policy π^* and value v^* can be learned via Monte Carlo approximation-based methods such as temporal difference (TD) learning [27]. There is much evidence suggesting that the firing rate of midbrain dopaminergic neurons might represent the reward prediction error in TD learning. Thus, the learning of value and policy in the current model could potentially be implemented in a manner similar to previous TD learning models of the basal ganglia [8,9,11,42].

Acknowledgments

The authors would like to thank Timothy Hanks, Roozbeh Kiani, Luke Zettlemoyer, Abram Friesen, Adrienne Fairhall and Mike Shadlen for helpful comments.

Author Contributions

Conceived and designed the experiments: YH RR. Performed the experiments: YH. Analyzed the data: YH. Contributed reagents/materials/analysis tools: YH. Wrote the paper: YH RR.

References

1. Knill D, Richards W (1996) Perception as Bayesian inference. Cambridge: Cambridge University Press.
2. Zemel RS, Dayan P, Pouget A (1998) Probabilistic interpretation of population codes. Neural Computation 10.
3. Rao RPN, Olshausen BA, Lewicki MS (2002) Probabilistic Models of the Brain: Perception and Neural Function. Cambridge, MA: MIT Press.
4. Rao RPN (2004) Bayesian computation in recurrent neural circuits. Neural Computation 16: 1–38.
5. Ma WJ, Beck JM, Latham PE, Pouget A (2006) Bayesian inference with probabilistic population codes. Nature Neuroscience 9: 1432–1438.
6. Doya K, Ishii S, Pouget A, Rao RPN (2007) Bayesian Brain: Probabilistic Approaches to Neural Coding. Cambridge, MA: MIT Press.
7. Daw ND, Courville AC, Touretzky D (2006) Representation and timing in theories of the dopamine system. Neural Computation 18: 1637–1677.
8. Dayan P, Daw ND (2008) Decision theory, reinforcement learning, and the brain. Cognitive, Affective and Behavioral Neuroscience 8: 429–453.
9. Bogacz R, Larsen T (2011) Integration of reinforcement learning and optimal decision making theories of the basal ganglia. Neural Computation 23: 817–851.
10. Law CT, Gold JI (2009) Reinforcement learning can account for associative and perceptual learning on a visual-decision task. Nat Neurosci 12: 655–663.
11. Rao RPN (2010) Decision making under uncertainty: A neural model based on POMDPs. Frontiers in Computational Neuroscience 4.
12. Kaelbling LP, Littman ML, Cassandra AR (1998) Planning and acting in partially observable stochastic domains. Artificial Intelligence 101: 99–134.
13. Drugowitsch J, Moreno-Bote R, Churchland AK, Shadlen MN, Pouget A (2012) The cost of accu-mulating evidence in perceptual decision making. J Neurosci 32: 3612–3628.
14. Shenoy P, Rao RPN, Yu AJ (2010) A rational decision-making framework for inhibitory control. Advances in Neural Information Processing Systems (NIPS) 23. Available: http://www.cogsci.ucsd.edu/~ajyu/Papers/nips10.pdf. Accessed 2012 Dec 24.
15. Shenoy P, Yu AJ (2012) Rational impatience in perceptual decision-making: a bayesian account of discrepancy between two-alternative forced choice and go/nogo behavior. Advances in Neural Information Processing Systems (NIPS) 25. Cambridge, MA: MIT Press.
16. Huang Y, Friesen AL, Hanks TD, Shadlen MN, Rao RPN (2012) How prior probability influences decision making: A unifying probabilistic model. Advances in Neural Information Processing Systems (NIPS) 25. Cambridge, MA: MIT Press.
17. Shadlen MN, Newsome WT (2001) Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. Journal of Neurophysiology 86.
18. Newsome WT, Britten KH, Movshon JA (1989) Neuronal correlates of a perceptual decision. Nature 341: 52–54.
19. Salzman CD, Britten KH, Newsome WT (1990) Cortical microstimulation inuences perceptual judgements of motion direction. Nature 346: 174–177.
20. Britten KH, Shadlen MN, Newsome WT, Movshon JA (1992) The analysis of visual motion: a comparison of neuronal and psychophysical performance. J Neurosci 12: 4745–4765.
21. Shadlen MN, Newsome WT (1996) Motion perception: seeing and deciding. Proc Natl Acad Sci 93: 628–633.
22. Wang XJ (2002) Probabilistic decision making by slow reverberation in cortical circuits. Neuron 36: 955–968.
23. Mazurek ME, Roitman JD, Ditterich J, Shadlen MN (2003) A role for neural integrators in per-ceptual decision-making. Cerebral Cortex 13: 1257–1269.
24. Beck JM, Ma W, Kiani R, Hanks TD, Churchland AK, et al. (2008) Probabilistic population codes for Bayesian decision making. Neuron 60: 1142–1145.
25. Britten KH, Shadlen MN, Newsome WT, Movshon JA (1993) Responses of neurons in macaque MT to stochastic motion signals. Vis Neurosci 10(6): 1157–1169.
26. Casella G, Berger R (2001) Statistical Inference, 2nd edition. Pacific Grove, CA: Duxbury Press.
27. Sutton RS, Barto AG (1998) Reinforcement Learning: An Introduction. Cambridge, MA: MIT Press.
28. Latham PE, Roudi Y, Ahmadi M, Pouget A (2007) Deciding when to decide. SocNeurosciAbstracts 740.
29. Frazier P, Yu A (2008) Sequential hypothesis testing under stochastic deadlines. Advances in Neural Information Processing Systems 20: 465–472.
30. Churchland AK, Kiani R, Shadlen MN (2008) Decision-making with multiple alternatives. Nat Neurosci 11: 693–702.
31. Roitman JD, Shadlen MN (2002) Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. Journal of Neuroscience 22(21): 9475–9489.

32. Luce RD (1986) Response times: their role in inferring elementary mental organization. Oxford: Oxford University Press.
33. Hanks TD, Mazurek ME, Kiani R, Hopp E, Shadlen MN (2011) Elapsed decision time affects the weighting of prior probability in a perceptual decision task. Journal of Neuroscience 31: 6339–6352.
34. Cisek P, Puskas G, El-Murr S (2009) Decisions in changing conditions: The urgency-gating model. Journal of Neuroscience 29: 11560–11571.
35. Palmer J, Huk AC, Shadlen MN (2005) The effects of stimulus strength on the speed and accuracy of a perceptual decision. Journal of Vision 5: 376–404.
36. Bogacz R, Brown E, Moehlis J, Hu P, Holmes P, et al. (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.
37. Lai TL (1988) Nearly optimal sequential tests of composite hypotheses. The Annals of Statistics 16(2): 856–886.
38. Frazier PL, Yu AJ (2007) Sequential hypothesis testing under stochastic deadlines. In Advances in Neural Information procession Systems 20. Cambridge, MA: MIT Press.
39. Yu A, Cohen J (2008) Sequential effects: Superstition or rational behavior. In Advances in Neural Information Processing Systems 21: 1873–1880.
40. Ghose GM, Maunsell JHR (2002) Attentional modulation in visual cortex depends on task timing. Nature 419(6907): 616–620.
41. Ludwig CJH (2009) Temporal integration of sensory evidence for saccade target selection. Vision Research 49: 2764–2773.
42. Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. Science 275: 1593–1599.
43. Ditterich J (2006) Stochastic models and decisions about motion direction: Behavior and physiology. Neural Networks 19: 981–1012.