

Policy Adjustment in a Dynamic Economic Game

Jian Li^{1,9}, Samuel M. McClure^{1,9*}, Brooks King-Casas¹, P. Read Montague^{1,2*}

1 Human Neuroimaging Laboratory, Center for Theoretical Neuroscience, Department of Neuroscience, Baylor College of Medicine, Houston, Texas, United States of America, **2** Menninger Department of Psychiatry and Behavioral Science, Baylor College of Medicine, Houston, Texas, United States of America

Making sequential decisions to harvest rewards is a notoriously difficult problem. One difficulty is that the real world is not stationary and the reward expected from a contemplated action may depend in complex ways on the history of an animal's choices. Previous functional neuroimaging work combined with principled models has detected brain responses that correlate with computations thought to guide simple learning and action choice. Those works generally employed instrumental conditioning tasks with fixed action-reward contingencies. For real-world learning problems, the history of reward-harvesting choices can change the likelihood of rewards collected by the same choices in the near-term future. We used functional MRI to probe brain and behavioral responses in a continuous decision-making task where reward contingency is a function of both a subject's immediate choice and his choice history. In these more complex tasks, we demonstrated that a simple actor-critic model can account for both the subjects' behavioral and brain responses, and identified a reward prediction error signal in ventral striatal structures active during these non-stationary decision tasks. However, a sudden introduction of new reward structures engages more complex control circuitry in the prefrontal cortex (inferior frontal gyrus and anterior insula) and is not captured by a simple actor-critic model. Taken together, these results extend our knowledge of reward-learning signals into more complex, history-dependent choice tasks. They also highlight the important interplay between striatum and prefrontal cortex as decision-makers respond to the strategic demands imposed by non-stationary reward environments more reminiscent of real-world tasks.

Citation: Li J, McClure SM, King-Casas B, Montague PR (2006) Policy Adjustment in a Dynamic Economic Game. PLoS ONE 1(1): e103. doi:10.1371/journal.pone.0000103

INTRODUCTION

Knowing how to behave adaptively reduces, in most circumstances, to knowing the consequences of available actions, or, how much reward each action will garner on average. Practically speaking, this is a daunting problem. This is particularly so since the reward associated with different actions depends on a wide variety of factors such as one's history of actions, the behavior of competitors, and even stochastic changes in the environment through time. For example, a bee's decision to harvest nectar from one flower has the inevitable consequence of decreasing the returns from that flower and increasing the returns from non-sampled flowers (the nectar levels can recover). In foraging theory, if the amount of available prey (i.e. reward) is greater than the appetite of the predators then the food supply will increase [1]. Sudden unexpected shocks can also have significant effects on the costs and benefits associated with different actions. In general, the reward available in the future depends in complex ways on a possibly overwhelming variety of environmental factors [2–3].

Despite this apparent complexity of action-reward relationships presented by the world, most work in neuroscience and psychology has focused on fixed action-reward dependencies and studied the change of action-reward contingencies in block design paradigms such as Wisconsin Card-sorting task (WCST), reversal learning paradigms and extinction paradigm where in certain block of trials the action-reward contingency is fixed [4–6]. This trend persists in recent neuroimaging studies in humans, in which operant learning paradigms have been studied extensively [7–16, but see 17]. These studies suggested that brain areas associated with the mesolimbic dopamine system (i.e. striatal structures, prefrontal cortex) play an important role in reward learning and action selection [18–21].

In this paper, we study changes in action selection reflective of changes in reward expectation in a series of tasks in which earned reward depends in complex ways on previous actions. The

rationale of this study was to fit each subject's behavior through time with a continuous error-based learning model (e.g. actor-critic model) [5,9–14] to predict subject's consequent action selection and correlated brain activity in a series of tasks where reward contingency is a function of both subject's immediate choice and choice history. We then asked how subject responded to unexpectedly introduced new reward structures since subjects are required to develop different strategies for these reward structures and usually the adjustment of strategies correspond to increasingly exploratory actions by the subjects; hence they would give us opportunities to observe the interplay between control signals in striatum and prefrontal cortex during the switches of periods when action-reward dependencies vary slowly through time and periods where subject's behavior becomes more variable and depends less on previous experience.

Academic Editor: Paul Zak, Claremont Graduate University, United States of America

Received: November 8, 2006; **Accepted:** November 18, 2006; **Published:** December 20, 2006

Copyright: © 2006 Li et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was funded by the Kane Family Foundation (P.R.M.), NINDS grant NS-045790 (P.R.M.) and NIDA grant DA-11723 (P.R.M.).

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

* **To whom correspondence should be addressed.** E-mail: ead@bcm.mc.edu

© The authors contribute all copyright.

✉ **Current address:** Department of Psychology, Princeton University, Princeton, New Jersey, United States of America

RESULTS

The experimental design and subjects' performance in these tasks has been described previously [9,22–23]. We reviewed subjects' behavioral tendencies, and note that all of the subjects performed in accord with these summaries on each task individually, and switched behavioral strategies rapidly when the tasks were switched in our current manipulation.

MS→RO task (Matching Shoulder→Rising Optimum)

In both the matching shoulders and rising optimum tasks, subjects have a strong tendency to perform near the crossing point in the reward functions (see Figure 1B legend for detail). This can be understood by considering how earned reward changes near the crossing point. Assuming that subjects choose A at the crossing point, the percent allocation to A (%A) will increase, resulting in decreased subsequent earned reward for selecting A (Figure 1B). Reward can be increased by switching to choice B, which also decreases the percent allocation to A, returning subjects to the crossing point. The converse sequence of events occurs if B is initially selected. As long as subjects tend to select in accord with which choice is expected to produce the greatest immediate reward (Herrnstein called this melioration, [3,24]), then they will perform at the crossing point in the reward functions (~33% allocation to A) in both the MS and RO tasks.

In the MS task, performing at the crossing point is the optimal solution [9]. However, in the RO task it is grossly sub-optimal. If subjects were to select button A on every choice in the RO task, they would experience a temporary decrease in earned reward that would subsequently reverse to produce the maximum average return (Figure 1B). This optimal policy (selecting A only) is an unstable equilibrium point in subjects' action selection policy due to the fact that at high allocation to A, choices to B produce greater immediate reward (Figure 1C). In the MS→RO task, subjects show evidence for both behavioral equilibria after the reward structure switch. They began performing near the crossing point, showing a temporary excursion to greater allocation to A (%A), and then reverted again to performing near the crossing point (Figure 2A).

FR→PR task (Flat return → Pseudorandom)

Regardless of the pattern of choices in the FR and PR tasks, the average earned reward will be the same (Figure 1D, dot line; Figure 2A). As in MS and RO, the FR reward structure possesses a crossing point in the reward functions that acts as a stable selection strategy (Figure 1D).

Subjects always performed the PR task immediately after the FR task. Furthermore, we configured the reward returns in PR so that the mean and variance in rewards that subjects experienced in PR task were equal to what the subjects earned on the FR task (Figure 1D). Unlike the FR task, reward in the PR paradigm was randomly determined and was not dependent on the subjects' choices (randomly drawn from a uniform distribution). Under these conditions (PR), subjects tend to perform randomly, evenly distributing their choices between A and B. The transition from performing near the crossing point in the FR task (~40% allocation to A) to equally distributing choice in the PR task (~50% allocation to A) occurs at variable delays across subjects (see below for discussion).

Reinforcement learning model of reward learning

Reward learning requires monitoring the expected reward for the available actions (A and B), and biasing choices in favor of the action with highest expected reward. We modeled this process

using the two reinforcement learning models described above (Figure 2B). The sigmoid model assumes that subjects tend to select the choice associated with the greatest model weight (i.e. more likely to selected A when $w_A > w_B$). Furthermore, the probability that subjects select the choice with the greater estimated weight is expected to scale with the difference in weights ($w_A - w_B$), while the ϵ -greedy method assumes a probability of $1 - \epsilon/2$ to the choice with bigger weight (w^*). To test these predictions, we calculated $w_A - w_B$ at the time of every choice and arranged choices in order of increasing weight difference for the sigmoid action selection method. We then compared the observed probability of selecting A (P_A) by subjects with the probability predicted by the logistic decision function (Green, Figure 2B). For all 4 reward structures, this analysis revealed a strong correlation between observed and estimated probabilities of selecting choice A (MS: $r = 0.97$, RO: $r = 0.99$, FR: $r = 0.97$, PR: $r = 0.97$). For the ϵ -greedy method, we assigned individual subject's probability to the choice associated with bigger weight to be $1 - \epsilon/2$ and probability for the other choice is thus $\epsilon/2$. We then arranged choices in order of increasing probability of choosing a specific choice (A or B) and then compared the observed probability of selecting A (P_A) by subjects with that predicted by the ϵ -greedy decision function (Pink, Figure 2B). This analysis revealed a similar fitting as the softmax action selection method (MS: $r = 0.97$, RO: $r = 0.99$, FR: $r = 0.95$, PR: $r = 0.99$) both in behavioral fitting and further neural correlates mapping.

Neural correlates of prediction error

Reinforcement learning model states that learning signals (prediction error) are used to update and monitor the value of choices. In our experiment, we used prediction errors estimated from the model and applied it as one of the regressors in a general linear model (GLM) to imaging data. We find that the prediction error signal estimated from two methods (softmax action selection and ϵ -greedy) correlates with activity in the ventral striatum in both of our tasks with different reward structures (Figure 3).

Correlation with reward prediction error

The reinforcement learning model fits to the behavioral data provide estimates of the reward prediction error experienced after every choice. These prediction errors were used to produce regressors that were further fit to subjects' functional imaging data.

In the FR→PR task, the BOLD signal in regions of the ventral putamen correlated significantly with estimated prediction error signals ($p < 0.005$, uncorrected) using both methods (softmax and ϵ -greedy). However, at a threshold of $p < 0.005$ (uncorrected), we find that no areas other than visual cortex are significantly correlated with estimated prediction error signals in the matching shoulders to rising optimum (MS→RO) task. We reasoned that the lack of correlation with estimated prediction error in this task may result from the fact that the large negative prediction error and prolonged recovery phase produced by the change in reward paradigm (Figure 2A, Figure 3, blue and red traces) may dominate the overall fitting. To test this, we excluded the period of time encompassing the first 25 choices following the onset and switch in reward structures from the analysis. With this correction, BOLD signals in ventral putamen correlated significantly with estimated prediction error signals from both methods ($p < 0.001$, uncorrected). This result suggests BOLD signals in ventral striatum (putamen) can be predicted by prediction errors (PE) when action-reward dependencies vary slowly through time where PE fluctuates around 0 (Table 1, Figure 3), but not in phases where subject's behavior becomes more variable and is less dependent on previous experience.

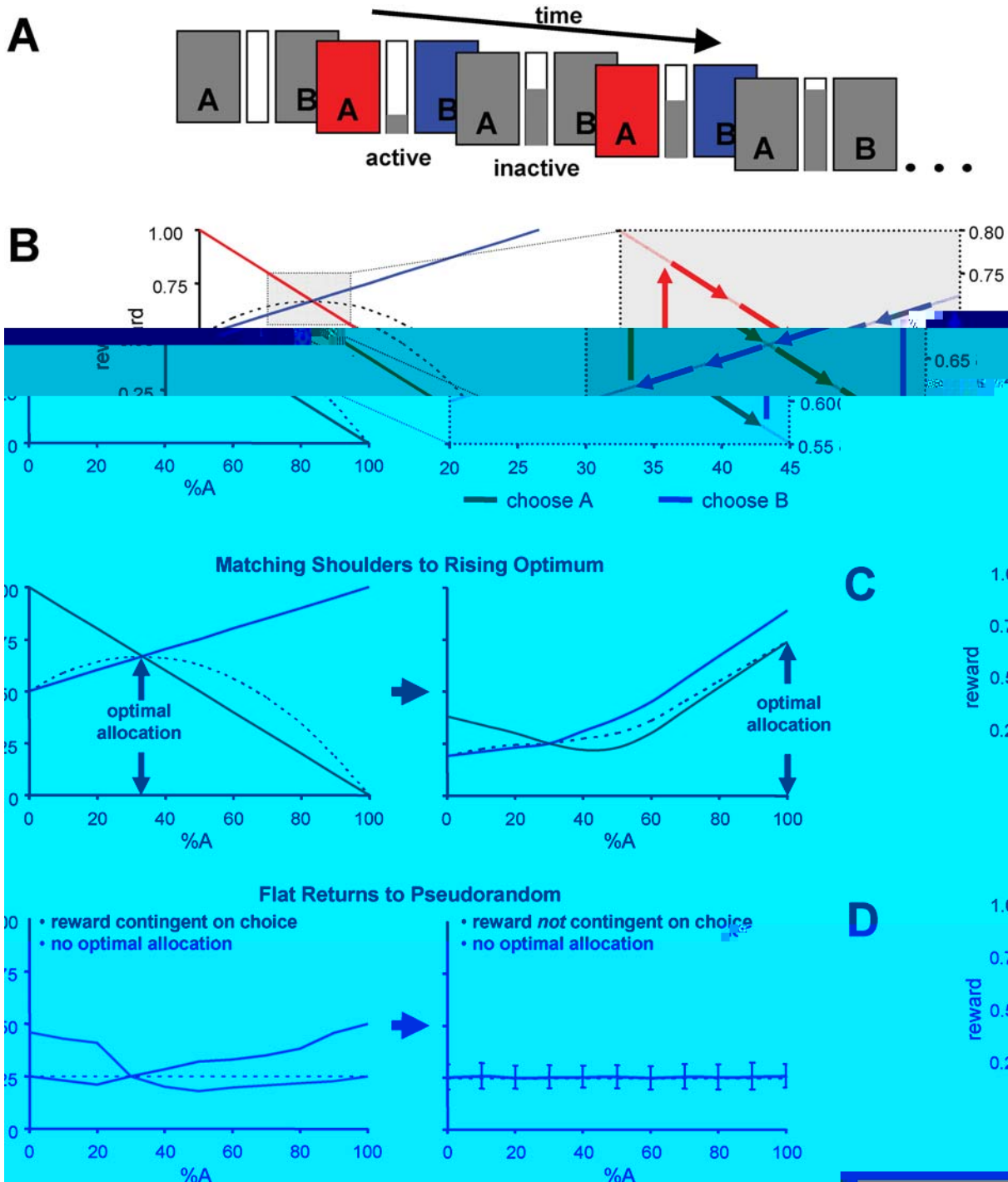


Figure 1. Experimental design. (A) Subjects engaged in a decision-making task in which they were instructed to choose from one of two actions (A or B) with the goal of obtaining and maintaining maximum earned reward indicated by a central location of a dashed line. (B) Example of a trial sequence: Central height (reward) depends on a variable: 1) chosen decision (A or B, Red and Blue arrows) and 2) the percentage of choice A (%A) made over the trial (x-axis). The initial %A is 50%. In cases in which a subject chooses choice A in more than 33% of the trial, a B choice is triggered and a B choice is earned than an A choice. In cases in which a subject chooses choice A in less than 33% of the trial, an A choice is triggered and a B choice is earned than an A choice. Thus, the higher the proportion of the rewarded function, 'B' choice becomes earned and more subjects are on the left. However, a subject must learn the proportion, a choice begins to yield greater reward than B choice and more subjects are on the right. In this example, the proportion is seen to the optimal allocation of A, a change in the allocation is indicated by the red arrow. (C) In the first task, a subject made 125 decisions in which they determined the matching height (MS) curve (left panel), followed by 125 decisions in which they determined the rising optimum (RO) curve (right panel). (D) In a second task, a subject began in the flat return (FR) curve (left panel) and switched to a pseudorandom (PR) task (right panel). In the FR task, all choices are equally rewarded (dashed line). In the PR condition, a random reward is determined independent of choice but a subject is given the same mean and variance of reward as earned in the FR task.

doi:10.1371/journal.pone.0000103.g001

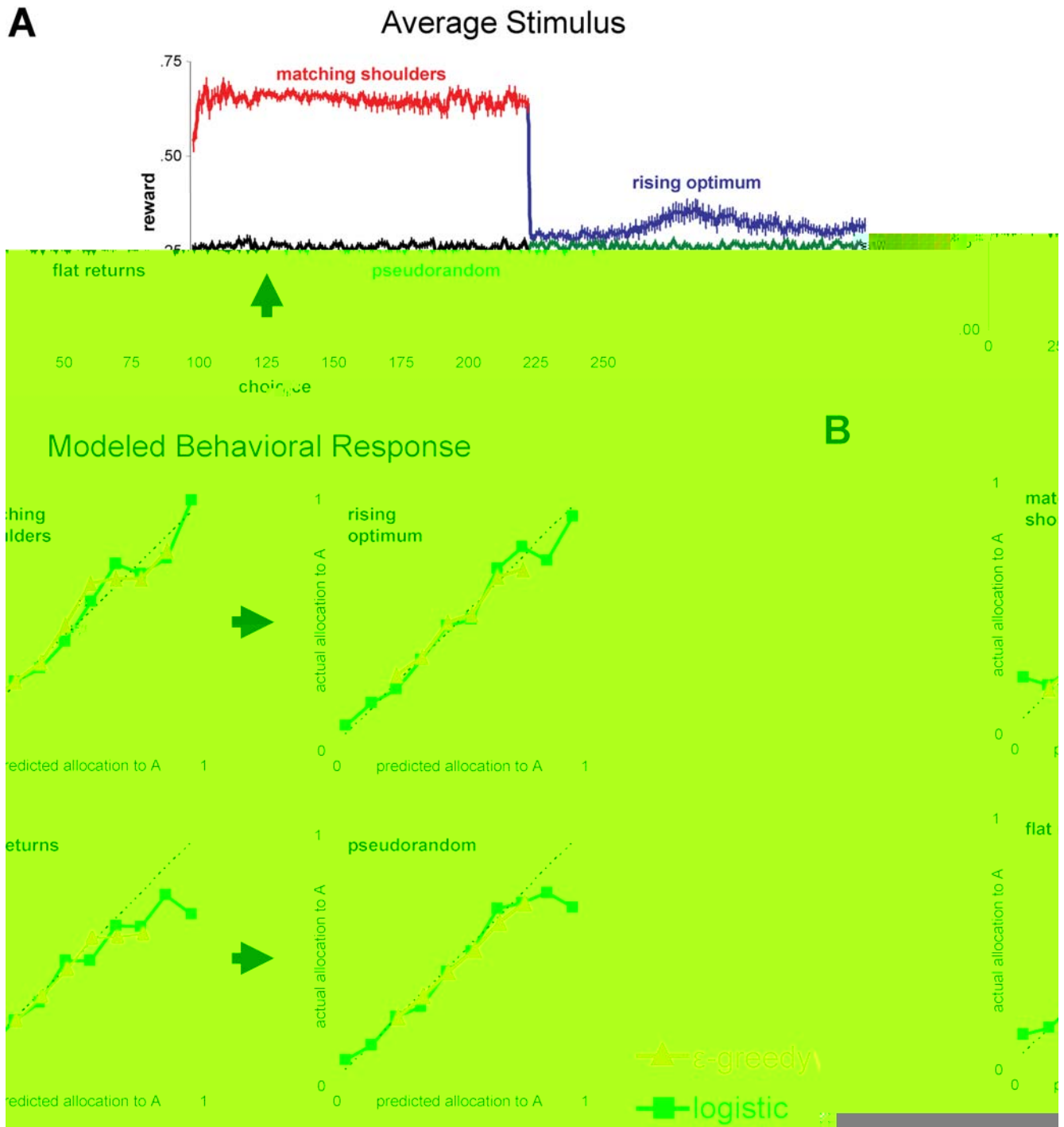


Figure 2. Experienced rewards and modeled behavioral response. (A) The choice from matching shoulders (MS) or rising optimum (RO) is a decision signaled by a large decrease in immediate reward (~60%). However, the choice from the flat returns (FR) or pseudorandom (PR) condition did not elicit a similar change in experienced reward. Read S.E. indicates behavioral bias at each choice. (B) Subjective decision is modeled using a reinforcement learning model in which different methods determine the probability of choice action (ϵ -greedy method and sigmoid method). For both methods, we have subjectively maintained independent estimates of the reward received for each choice, A and B, and used the relative balance of experienced reward using the Rescorla-Wagner learning algorithm. Choice probabilities are modeled by the sigmoid function (of method, green curve) or 2) have a fixed probability of $1-\epsilon/2$ for choice a priori based on predicted likelihood has been chosen A. Y-axis indicates the actual allocation to A for all choices in each bin. Lineage regression has been predicted and actual choice. (MS: ϵ -greedy = 0.97, RO: ϵ -greedy = 0.99, FR: ϵ -greedy = 0.95, PR: ϵ -greedy = 0.99 for ϵ -greedy method; MS: ϵ -greedy = 0.97, RO: ϵ -greedy = 0.99, FR: ϵ -greedy = 0.95, PR: ϵ -greedy = 0.99 for sigmoid method). ϵ -greedy method is a different probability of choice specific choice for each trial in the 250 trial task. This ϵ -greedy method does not cover all the methods of X-axis sigmoid method does.

doi:10.1371/journal.pone.0000103.g002

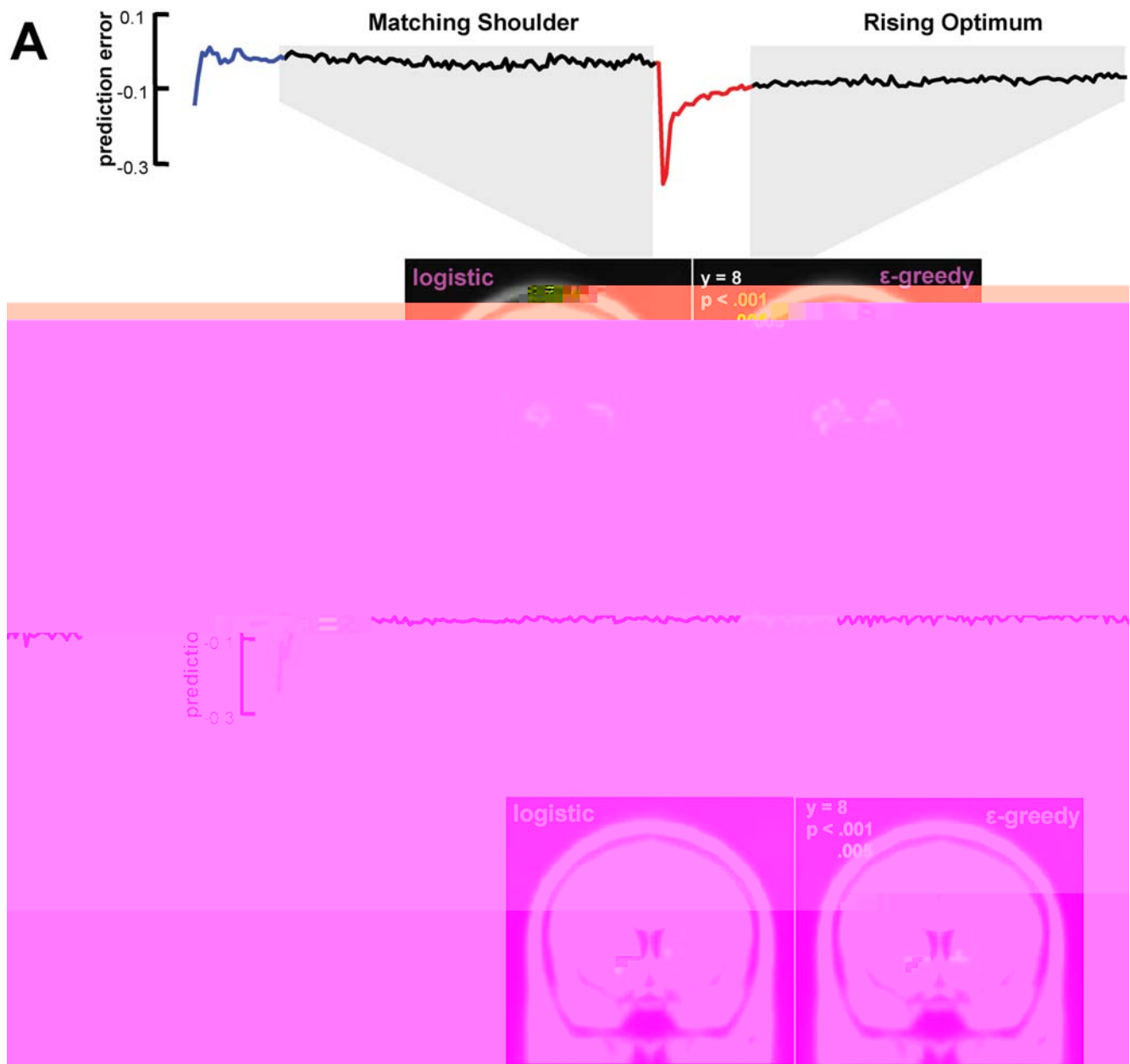


Figure 3. Activity in ventral striatum correlated to prediction errors with two different models. For each choice, a prediction error is generated by comparing the value experienced by the subject in the current choice to the modeled value for the choice. The change in prediction error, $\delta(t)$, is compared for two different tasks. Neural activity corresponding to prediction error has been generated from independent fitting information model of two different tasks (MS \rightarrow RO and FR \rightarrow PR) using two different methods (of maximum likelihood and ϵ -greedy) as identified through general linear model (GLM) analysis. Peak activity for left and right striatum are located at MNI $[-12, 8, 8]$ and $[24, 12, -4]$ for MS \rightarrow RO task and $[-16, 8, -4]$ and $[12, 8, 0]$ for FR \rightarrow PR task using of maximum likelihood. Activity in striatum is significantly greater in the prediction error in the left striatum of both tasks (left: <0.001 ; right: <0.005 , uncorrected).
doi:10.1371/journal.pone.0000103.g003

These results are further confirmed by fitting our learning model to each of the 4 sub-tasks (MS, RO, FR, and PR) independently. Prediction errors generated in this manner, omitting periods immediately after the introduction of new reward structures (25 trials, Figure S3, red and blue), correlate with BOLD signals in the same area of the ventral striatum in each of the four sub-tasks (Figure S3). This indicates that, when behavior is relatively stable, the ventral striatum is engaged to dynamically track ongoing reward estimation errors. Overall, these results

correspond well to a recent report that prediction error-like signals occur in the striatum in an operant learning paradigm [12].

Brain activity during periods of unexpected reward structure switches

We hypothesized that the unexpected salient events can possibly act to indicate possible changes in reward contingency or reward paradigm and they may trigger further exploratory behaviors by

Table 1

Areas of activation positively related to prediction error (MS→RO task, logistic fitting)

		<i>peak MNI</i>					
<i>region of activation</i>	<i>T</i>	<i>Z</i>	<i>x</i>	<i>y</i>	<i>z</i>	<i>voxels</i>	
lef amen	3.65	3.40	−12	8	8	11	
igh amen	4.28	3.90	24	12	−4	10	
occi ial	4.39	3.99	4	−88	20	71	
Pa ahi ocam al g	4.86	4.33	24	0	−12	5	
Infe io a ie al co e	3.76	3.48	−32	−44	44	5	

Areas of activation positively related to prediction error (FR→RO task, logistic fitting)

		<i>peak MNI</i>					
<i>region of activation</i>	<i>T</i>	<i>Z</i>	<i>x</i>	<i>y</i>	<i>z</i>	<i>voxels</i>	
lef amen	3.77	3.50	−16	8	−4	7	
igh amen	3.70	3.44	12	8	0	4	

Region i h 3 o g e e ignifican o el e e iden ified ing T- e , <.001 (nco ec ed).
doi:10.1371/jo nal. one.0000103. 002

subjects. There were three time points in the behavioral tasks that reliably signaled significant changes in reward paradigms which can also be confirmed from deviated prediction error signals (Figure 3). We consider brain areas that are activated in all of these instances as involved in abstract rules monitoring and detection and their activities can be triggered by salient events (dramatic immediate reward change in our case, [25–26]) and further help to promote more exploratory behaviors by subjects in order to determine more optimal strategies for current reward structure. We identified these areas using a conjunction analysis (intersection of areas significantly correlated at $p < 0.001$, uncorrected for multiple comparisons, in each instance, [27]). The three points include the beginning of both tasks (first choices in MS and first choices in FR), and the instant when the task paradigm switched from MS to RO. In the MS→RO task, subjects invariably played near the crossing point in the MS task (Figure S1), which resulted in a significant (60%) drop in earned reward when the task switched to RO reward structure. This change caused subjects to alter (even if briefly) behavioral strategy (Figure S2).

Three brain areas were identified by the conjunction analysis including the bilateral anterior insula and a region in the inferior

frontal gyrus (IFG) (See Table 2 for detailed description). The region identified in the insula has been implicated in responding to cognitive conflict and behavior inhibition [21,28]. The other region we identified lies on the IFG and extends into the frontal operculum (BA 44, Figure 4). This region has been identified under conditions requiring increased attention and changes in behavior [29].

Another time point in our experiment that might be associated with exploration is the task switch in the FR→PR paradigm. Subjects changed from performing near the crossing point (~40% A%) in the FR task to random selecting behavior in the PR task (~50% A%) (Figure S2). There are two differences between the task switch in FR→PR and MS→RO that underlie why we did not include it in the conjunction analysis. First, in FR→PR, subjects switch to a null behavioral strategy (random play). Thus it is unclear to what degree subjects are exploring new behavioral strategies as opposed to simply omitting strategic play. Secondly, the reward structure switch in FR→PR task does not produce as dramatic a signal of changed reward contingencies as in the MS→RO paradigm. In MS→RO task, the bar height decreases approximately 60% at the task switch. For FR→PR task, the mean

Table 2.

Areas of activation positively related to onset and switch of reward structures

		<i>peak MNI</i>					
<i>region of activation</i>	<i>T</i>	<i>Z</i>	<i>x</i>	<i>y</i>	<i>z</i>	<i>voxels</i>	
lef IFG	6.11	5.19	−48	4	28	21	
igh IFG	7.61	6.07	44	8	28	48	
lef in la	6.00	5.11	−32	24	4	28	
igh in la	4.63	4.16	36	20	8	18	
lef amen	5.83	5.00	−8	4	0	18	
igh amen	6.98	5.71	12	4	0	15	
midb ain	6.47	5.41	−4	−28	−4	44	
occi ial	7.95	6.25	−32	−88	16	639	

Region i h 15 o g e e ignifican o el e e iden ified ing T- e , <.0001 (nco ec ed).
doi:10.1371/jo nal. one.0000103. 002



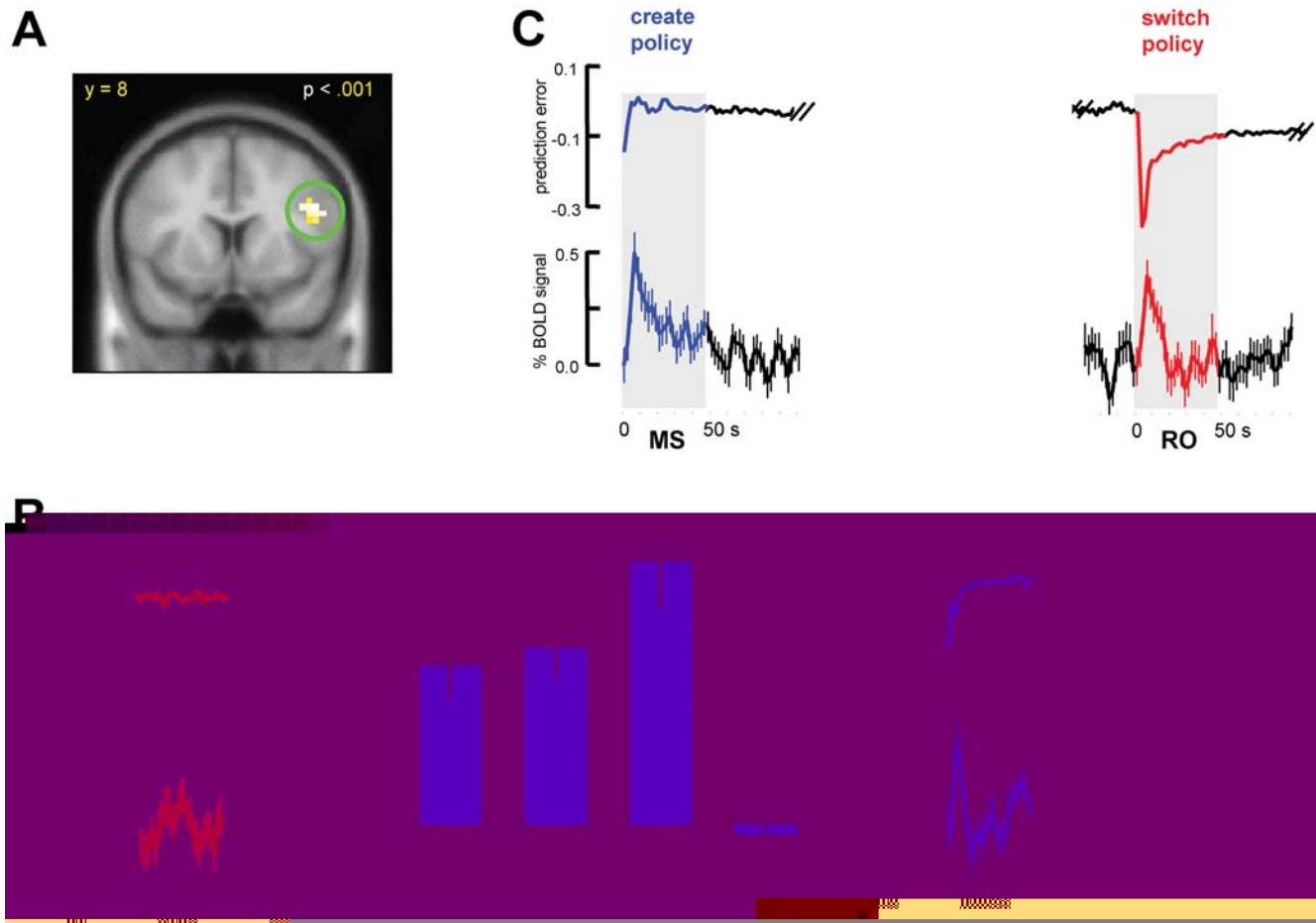


Figure 4. Inferior frontal gyrus (IFG) and anterior insula respond to reward structure switches. (A) The IFG activity is identified by a conjunction analysis in a brain volume on the condition of new reward choice. (B) Relative BOLD signal change of IFG in four different blocks are estimated by general linear model (GLM). Activity is shown for each of the four blocks (MS, RO, FR) and is significantly larger than zero (<0.001 , one-tailed test). (C) BOLD signal change of the confided brain region of interest (ROI) analysis in the four blocks. The period corresponds to initial learning the reward choice of the task (blue) and adapting to change in reward choice in the middle of the MS→RO task (red). All statistical results are available for the analysis in the text. doi:10.1371/journal.pone.0000103.g004

and variance in reward were unchanged after the reward structure switch. Subjects may therefore require more time to discover and respond to the change in reward contingencies and the time for each subject to discover and respond could vary across individuals. This is further confirmed by the fact that reward prediction error signal estimated from both of our models does not indicate any significant changes during the switch from FR to PR reward structure (Figure 3). Consistent with this, activity does increase in the area of PFC identified by the conjunction analysis following the switch in FR→PR. However, the amplitude is reduced and the duration of activity is prolonged relative to MS→RO task. This is the expected outcome were subjects to discover and respond to the task switch at more delayed and variable times.

DISCUSSION

Using a continuous decision-making task with four different dynamic reward structures underlying subjects' action-outcome contingencies, we found in this paper that differential involvement of brain areas in action selection and learning during different time periods of the task. Two model-based regression analyses showed BOLD activities in the ventral striatal structure correlate with prediction error signal. However, a sudden introduction of new

reward structures engages more complex control circuitry in the prefrontal cortex (inferior frontal gyrus and anterior insula) and is not captured by a simple actor-critic model. Separately, these two systems have been the subject of numerous investigations of decision-making. The first of these systems, the ventral striatum, is believed to be involved specifically in maintaining and updating the expected reward value of actions. This is accomplished through interactions with the mesolimbic dopamine system which activates in accord with ongoing reward prediction error signals [18]. In the striatum, dopamine is known to modulate synaptic plasticity [19] allowing for the activity of these neurons to encode action value [30]. Functional MRI studies have demonstrated that the striatum is clearly involved in biasing action selection in accord with current action values [13], and that activity in this structure changes in accord with ongoing prediction errors [10,31–32]. Our findings show that these results hold during periods of conditional action-selection learning in each of the four different reward structures when, presumably, the striatum is the primary determinant of behavior. Our results distinguished from previous studies by using different underlying reward functions originally derived from Herrnstein's matching law [3,22,24] that reward associated with each choice depends not only on the current

choice but also subject's previous choice history, while most of the previous studies adopted fixed action-reward contingencies paradigms [11–14,33]. Animals have to face a non-stationary world and the amount of reward expected from a contemplated action depends in complex ways on the history of an animal's choices. This can dramatically change the likelihood of rewards collected by the same choices in the near-term future, and our result indicates similar neural correlates are required in reward-learning tasks more reminiscent of real-world environment.

During periods where immediate reward return fluctuates dramatically (at the beginnings of both tasks, and in the middle of MS->RO task), activity in the striatum is not well captured by reinforcement learning models, suggesting that the striatum is not strongly involved in action selection during these periods. In these periods, brain activity is increased in two areas of the prefrontal cortex: in the inferior frontal gyrus as well as bilaterally in the anterior insula. The insula has primarily been implicated as responding to disgust, pain, and other aversive stimuli [34–36]. In terms of decision-making tasks such as ours, the insula has also

26 axial slices (3.4×3.4×4 mm width) parallel to the AC-PC line. Images were acquired with a repetition time (TR) of 2s, an echo time (TE) of 40ms, and flip angle of 90°.

Experimental Task

Subjects lay supine with their head in the scanner bore and observed the rear-projected computer screen via a 45° mirror mounted to the head coil. Choices were registered using two MRI-compatible button boxes. Selections to A were made by pressing any button with the left hand and selections to B by pressing any button with the right hand for half number of the subjects and in the reversed pattern for the other half. After each selection, the central reward bar obtained a new height dependent on earned reward. Following this, the buttons (A and B) on the screen were disabled and turned gray for 1.25 s. Subjects were instructed that they could not make further selection until the buttons on the screen turned back to normal color from gray.

Subjects engaged in two repeated play, two-alternative decision-making tasks in which they were instructed to choose from one of two actions (A or B) with the goal of obtaining and maintaining maximum earned reward (Figure 1A; [8,23]). The central bar height (reward) is controlled by two variables: 1) Current choice made (A or B), if the subject chooses A, then the reward received will be along the red line, otherwise, the reward will be on the blue line; 2) Subject' choice history: the percentage of choice "A" (%A) selected in the past 20 choices. The initial %A value is set to be 50%. As task proceeds, the %A is updated (a 20-choice moving window) as a result of each choice (A or B) subjects made. The tasks were modified for use in fMRI by pacing the rate at which choices are made to no faster than one every 1.25 s. The mean reaction time was slightly less than 2 s. Each task required subject to make 250 selections. After the first 125 selections, the reward structure was switched (Figure 1B, C). Subjects were not instructed that these switches would occur. In the first task, the reward structure was initially defined by the matching shoulders (MS) paradigm and was then switched to the rising optimum (RO) reward paradigm (MS→RO task; Figure. 1C). The other task began with the flat returns (FR) paradigm and was switched to pseudo-random (PR) returns at the switch (FR→PR task; Figure. 1D). In all reward paradigms except pseudo-random, earned reward depended on two variables: (1) the subject's choice (A or B; corresponding to red and blue reward curves in Figure 1C, D, respectively), and (2) the percent of the last 20 choices made to choice A (%A, allocation to A; x-axis on plots in Figure. 1B, C, D). Allocation to A was set to 50% at the beginning of both tasks.

Data Analysis

Imaging data was analyzed using SPM2 [50] and xjView (<http://people.hnl.bcm.tmc.edu/cuixu/xjView/>). Functional images were realigned, corrected for slice timing, coregistered with a canonical brain in MNI coordinates, resliced to 4x4x4mm and smoothed with an 8mm FWHM Gaussian kernel prior to analysis.

The prediction error signal, $\delta(t)$, determined by fitting the behavioral data, was used to produce a regressor through convolution with a canonical hemodynamic response kernel. To find brain voxels sensitive to changes in reward paradigm, we used a regressor with a single hemodynamic response function offset to the time of reward structure switch (beginning of each reward structure).

Regressors were fit independently to data from each voxel in the functional brain images using standard linear model methods. A random effects analysis was conducted by performing one-sample t-tests over best fitting beta amplitudes produced by linear model

fitting. Brain areas are considered significantly activated that are composed of at least 5 contiguous voxels significant at $p < 0.005$ with peak significance in the cluster of at least $p < 0.001$.

Behavior Fitting–Modeling

Subjects' decision-making was modeled with a reinforcement learning algorithm. We assume that subjects maintained independent estimates of the reward expected for each choice, A and B, and updated these values based on experienced rewards. In particular, we assume choice values (w_A and w_B) were updated according to a Rescorla-Wagner learning algorithm.

We used two methods to assign probabilities to each choice: 1) logistic method and 2) ϵ -greedy method. In logistic method, choices were assumed to be probabilistically related to choice values according to a sigmoid function with slope m :

$$P_A(t) = \frac{e^{mw_A(t)}}{e^{mw_A(t)} + e^{mw_B(t)}} \quad (1)$$

While in ϵ -greedy method, one of the two alternative choices will be assigned with probability $1-\epsilon/2$ if the weight associated with that choice is bigger than the other and the probability of choosing the other choice is thus $\epsilon/2$. When the weights associated with two choices are equal, then one choice will be randomly assigned with probability $1-\epsilon/2$, the other $\epsilon/2$.

$$\begin{aligned} P_A(t) &= H(w_A(t) - w_B(t)) * (1 - \epsilon/2) + \\ &H(w_B(t) - w_A(t)) * (\epsilon/2); \end{aligned} \quad (2)$$

Where $H(x)$ is a Heaviside step function and defined by

$$H(x) = \begin{cases} 0 & x < 0 \\ 1/2 & x = 0 \\ 1 & x > 0 \end{cases}$$

In both methods, for each choice (denote the choice by *), the reward experienced by the subject $r(t)$ was compared with the current modeled weight value $w_*(t)$ to produce a prediction error $\delta(t)$:

$$\delta(t) = r(t) - w_*(t) \quad (3)$$

The prediction error served as a learning signal that was used by both methods to improve modeled action weights by an amount governed by the learning rate λ :

$$w_*(t+1) = w_*(t) + \lambda \delta(t); \quad (4)$$

The quality of both model fittings was determined by how well F16d

parameter fitting, fitting was initiated from 20 randomly determined starting values and the best fit was taken across all final parameter values. The learning rate was restricted to values between 0 and 1; the sigmoid slope was restricted to positive values and the greedy parameter (ϵ) was restricted to values between 0 and 1.

SUPPORTING INFORMATION

Figure S1 Individual subject performance variability in both tasks. These four panels represent how individual subject averagely perform in 4 different reward structures (MS, RO, FR, PR). Immediate reward subject receive from each choice they make depends on two variables: 1) current decision (A or B, Red and Blue trace correspondingly) and 2) the percentage of choice A (%A) made over the past 20 trials (x-axis). Each subject's average behavior is represented by a triangle on each reward structure plot. Most subject perform around the optimal strategy (cross point of red and blue curve) in the MS task, while in RO task, subjects tend to split along the %A and many subject were restricted to the crossing point which is not the optimal strategy anymore. In FR task, subjects were still slightly attracted by the crossing point while in the PR task subjects were randomly distributed around the 50% %A point.

Found at: doi:10.1371/journal.pone.0000103.s001 (1.44 MB TIF)

Figure S2 Various behavioral responses subjects performed in both tasks. Subjects quickly adjusted to the optimal strategy at the beginning of both tasks (MS and FR). The switch from matching shoulders (MS) to rising optimum (RO) reward structures was signaled by a large decrease in immediate reward return (Fig. 2) and could possibly trigger the more exploratory behavior in the

RO task. However, the switch from the flat returns (FR) structure to the pseudorandom (PR) condition did not elicit a similar change in experienced reward and thus although the general behavioral patterns in FR and PR task are different ($\sim 40\%$ %A in FR task and $\sim 50\%$ %A in PR task), there is no evidence indicating a reliable exploratory phase in PR task. %A S.E. is indicated by vertical bars at each choice.

Found at: doi:10.1371/journal.pone.0000103.s002 (1.60 MB TIF)

Figure S3 Ventral striatum as the neural correlate of average prediction error, $\delta(t)$, across subjects for four different sub-tasks using softmax method. Neural activity corresponding to prediction errors generated from independently fitting the softmax reinforcement model to 4 sub-tasks (MS, RO, FR, PR). Activity in ventral striatum correlates with the magnitude of prediction error in each sub-task (MS, RO, FR, PR) (red: $p < 0.001$; yellow: $p < 0.005$, uncorrected) when reward contingencies vary slowly through time.

Found at: doi:10.1371/journal.pone.0000103.s003 (3.82 MB TIF)

ACKNOWLEDGMENTS

We thank Ann Harvey, Kristen Pfeiffer, and Nathan Apple for help conducting these experiments. We benefited from valuable discussion of the results and analyses with Leigh Nystrom, Jason Chein, Chess Stetson and Xu Cui.

Author Contributions

Conceived and designed the experiments: SM PM. Performed the experiments: PM JL. Analyzed the data: SM PM JL BK. Contributed reagents/materials/analysis tools: SM PM JL BK. Wrote the paper: SM PM JL BK.

REFERENCES

- Stephens DW, Krebs JR (1987) Foraging theory. Princeton NJ: Princeton University press. 262 p.
- Coase HR (1960) The problem of social cost. *J Law & Econ* 3: 1–23.
- Herrnstein RJ, Prelec D (1991) Melioration: A theory of distributed choice. *J Econ Perspect* 5(3): 137–156.
- Rescorla RA, Wagner AR (1972) A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Black AH, Prokasy WF, eds. *Classical conditioning II: Current research and theory*. New York: Appleton-Century-Crofts. pp. 64–99.
- Montague PR, Dayan P, Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J Neurosci* 16: 1936–1947.
- Balleine BW, Dickinson A (1998) Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology* 37: 407–419.
- Elliott R, Friston KJ, Dolan RJ (2000) Dissociable neural responses in human reward systems. *J Neurosci* 20: 6159–6165.
- Knutson B, Adams MC, Fong WG, Hommer D (2001) Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *J Neurosci* 21: RC159: 1–5.
- Montague PR, Berns GS (2002) Neural Economics and the biological substrates of valuation. *Neuron* 36: 265–284.
- O'Doherty J, Dayan P, Friston K, Critchley H, Dolan RJ (2003) Temporal difference models and reward-related learning in the human brain. *Neuron* 38: 329–337.
- Seymour B, O'Doherty J, Dayan P, Koltzenburg M, Jones AK, et al. (2004) Temporal difference models describe higher-order learning in humans. *Nature* 429: 664–667.
- O'Doherty J, Dayan P, Schultz J, Deichmann R, Friston K, et al. (2004) Dissociable roles of ventral and dorsal Striatum in instrumental conditioning. *Science* 304: 452–454.
- Haruno M, Kuroda T, Doya K, Toyama K, Kimura M, et al. (2004) A neural correlate of reward-based behavioral learning in caudate nucleus: a functional magnetic resonance imaging study of a stochastic decision task. *J Neurosci* 24: 1660–1665.
- Tanaka S, Doya K, Okada G, Ueda K, Okamoto Y, et al. (2004) Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nat Neurosci* 7(8): 887–893.
- Tricomi EM, Delgado MR, Fiez JA (2004) Modulation of caudate activity by action contingency. *Neuron* 41(2): 281–292.
- Galvan A, Hare TA, Davidson M, Spicer J, Glover G, et al. (2005) The role of ventral frontostriatal circuitry in reward-based learning in humans. *J Neurosci* 25: 8650–8656.
- Glaser J, Buchel C (2005) Formal learning theory dissociates brain regions with different temporal integration. *Neuron* 47: 295–306.

30. Kawagoe R, Takikawa Y, Hikosaka O (2004) Reward-predicting activity of dopamine and caudate neurons – a possible mechanism of motivational control of saccadic eye movements. *J Neurophysiol* 91: 1013–1024.
31. Berns GS, McClure SM, Pagnoni G, Montague PR (2001) Predictability modulates human brain response to reward. *J Neurosci* 21: 2793–2798.
32. McClure SM, Berns GS, Montague PR (2003) Temporal prediction errors in a passive learning task activate human striatum. *Neuron* 38: 339–346.
33. Daw ND, Niv Y, Dayan P (2005) Uncertainty-based competition between prefrontal and striatal systems for behavioral control. *Nat Neurosci* 8(12): 1704–1711.
34. Buchel C, Morris J, Dolan RJ, Friston KJ (1998) Brain systems mediating aversive conditioning: an event-related fMRI study. *Neuron* 20: 947–957.
35. Sanfey AG, Rilling JK, Aronson JA, Nystrom LE, Cohen JD (2003) The neural basis of economic decision-making in the Ultimatum Game. *Science* 300: 1755–1758.
36. Wicker B, Keysers C, Plailly J, Royet JP, Gallese V, et al. (2003) Both of us disgusted in My insula: the common neural basis of seeing and feeling disgust. *Neuron* 40: 655–664.
37. Aron AR, Fletcher PC, Bullmore ET, Sahakian BJ, Robbins TW (2003) Stop-signal inhibition disrupted by damage to right inferior frontal gyrus in humans. *Nat Neurosci* 6(2): 115–116.
38. Cools R, Clark L, Robbins TW (2004) Differential responses in human striatum and prefrontal cortex to changes in objects and rule relevance. *J Neurosci* 24(5): 1129–1135.
39. Greene JD, Sommerville RB, Nystrom LE, Darley JM, Cohen JD (2001) An fMRI investigation of emotional engagement in moral judgment. *Science* 293: 2105–2108.
40. McClure SM, Laibson DI, Loewenstein G, Cohen JD (2004) Separate neural systems value immediate and delayed monetary reward. *Science* 306: 503–507.
41. Frank MJ, Seeberger LC, O'Reilly RC (2004) By carrot or by stick: Cognitive Reinforcement learning in Parkinsonism. *Science* 306: 1940–1943.
42. Pasupathy A, Miller EK (2005) Different time course of learning-related activity in the prefrontal cortex and striatum. *Nature* 433: 873–876.
43. Gray JR (2004) Integration of emotion and cognitive control. *Curr Dir Psychol Sci* 13: 46–48.
44. Berns GS, Cohen JD, Mintum MA (1997) Brain regions responsive to novelty in the absence of awareness. *Science* 276: 1272–1275.
45. Montague PR, Hyman SE, Cohen JD (2004) Computational roles for dopamine in behavioral control. *Nature* 431: 760–767.
46. Clayton EC, Rajkowski J, Cohen JD, Aston-Jones G (2004) Phasic activation of monkey locus coeruleus neurons by simple decisions in a forced choice task. *J Neurosci* 24(44): 9914–9920.
47. Zink CF, Pagnoni G, Martin-Skurski ME, Chappelow JC, Berns GS (2004) Human striatal responses to monetary reward depend on saliency. *Neuron* 42(3): 509–517.
48. Yu AJ, Dayan P (2002) Acetylcholine in cortical inference. *Neural Netw* 15: 719–730.
49. Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ (2006) Cortical substrates for exploratory decisions in humans. *Nature* 441: 876–879.
50. Friston KJ, Holmes AP, Worsley K, Poline JP, Frith CD, et al. (1995) Statistical parametric maps in functional brain imaging: a general linear approach. *Hum Brain Mapp* 2: 189–210.