Implicit counterfactual effect in partial feedback reinforcement learning: behavioral and modeling approach

Zahra Barakchian[1]*, Abdol-hossein Vahabie[2,3]◕, Majid Nili Ahmadabadi[2]◕,

1 School of Cognitive Sciences, Institute for Research in Fundamental Sciences, Tehran, Iran
2 Cognitive Systems Laboratory, Control and Intelligent Processing Center of Excellence (CIPCE), School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran
3 Department of Psychology, Faculty of Psychology and Education, University of Tehran, Tehran, Iran

◕These authors contributed equally to this work.
* zbarakchian@ipm.ir

Abstract

Context by distorting values of options with respect to the distribution of available alternatives, remarkably affects learning behavior. Providing an explicit counterfactual component, outcome of unchosen option alongside with the chosen one (Complete feedback), would increase the contextual effect by inducing comparison-based strategy during learning. But It is not clear in the conditions where the context consists only of the juxtaposition of a series of options, and there is no such explicit counterfactual component (Partial feedback), whether and how the relativity will be emerged. Here for investigating whether and how implicit and explicit counterfactual components can affect reinforcement learning, we used two Partial and Complete feedback paradigms, in which options were associated with some reward distributions. Our modeling analysis illustrates that the model which uses the outcome of chosen option for updating values of both chosen and unchosen options, which is in line with diffusive function of dopamine on the striatum, can better account for the behavioral data. We also observed that size of this bias depends on the involved systems in the brain, such that this effect is larger in the transfer phase where subcortical systems are more involved, and is smaller in the deliberative value estimation phase where cortical system is more needed. Furthermore, our data shows that contextual effect is not only limited to probabilistic reward but also it extends to reward with amplitude. These results show that by extending counterfactual concept, we can better account for why there is contextual effect in a condition where there is no extra information of unchosen outcome.

Introduction

In everyday life, we frequently decide between options. Value of an option is usually learned via trial and error [1], and it is represented in multiple cortical and subcortical areas of the brain. Values of competing options interact with each other and consequently the context in which options are located can affect the representations [2]. Although the early studies on the contextual effects have been designed in the decision-making paradigm [3–8], a new trend has been formed recently to show that some of the behavioral biases come from contextual effects during value learning [9–11]. They showed that, in particular, in the paradigm in which the counterfactual outcomes pertaining to the chosen option were available (Complete feedback), subjects were strongly affected by the context, and this is mostly because they were induced to use a relative strategy. Although, it has been shown that there is a weaker contextual effect in the Partial version [9, 11], yet there is no global consensus about whether and how this effect happens.

Reinforcement learning is an incremental procedure that updates the option value via its prediction error [1]. This procedure is happening in the striatum where encodes action values [12–17], and is modulated by dopamine which encodes prediction errors [18]. Dopamine has opposing exciting and inhibiting effects on two distinct populations of striatal neurons called D1-SPNs and D2-SPNs (spiny projection neurons) respectively [19, 20]. Some reinforcement learning studies have shown opposing activities with similar strength in these two clusters during learning [21, 22]. Recent optical evidence suggests a model for Basal Ganglia that, it is the relative activity of these two clusters that represents an internal decision variable during decision making and learning [23–25]. For a good review on this issue, see [26]. Inspired by the opposing role of dopamine on D1- and D2-SPNs, while they encode two competing options' values, we proposed a simple reinforcement learning model called Opposing Learning model, in which the chosen prediction error in addition to updating the chosen option value (classic standard Q-learning), updates the unchosen option value, though in an opposing manner. This mechanism is consistent with diffusive nature of dopamine release and enables the model to endogenously encode the chosen and unchosen options' values relative to each other and consequently suggests having a contextual effect in the Partial feedback conditions too.

In the Complete feedback paradigm in which there are some exogenous factors that impose relativity on value learning procedure, the value learning strategy can be complex [9–11]. In these conditions, the main strategy might be to compare two presented outcomes and this would generate the regret and relief emotions. It has been shown that people tend to optimize their outcome difference, $\text{outcome}_{\text{factual}}$ - $\text{outcome}_{\text{counterfactual}}$ (i.e. minimize their regret and maximize their relief) [27–29]. In the Partial feedback paradigm, due to absence of regret and relief emotions, the main value learning strategy assumes to be the standard maximizing expected rewards. Interestingly, recent studies have illustrated that people are neither fully expected-reward optimizer nor fully outcome-difference optimizer, rather they are hybrid optimizers [9, 30], who use both of these strategies with different weights. The individual differences among people would depend on how much a person weighs each of these strategies. By adding a hybrid component to the basic OL model, we could extend the OL model for the Complete version too.

In this paper, we went beyond the standard definition of a counterfactual outcome, and focused on an uncommon subtle aspect of counterfactual role in value learning. This role is important in particular in the situations where there is no forgone outcome to trigger the comparison-based strategy explicitly. We used two types of feedback paradigms, with and without forgone outcomes. By using the chosen outcome as a counterfactual outcome for unchosen option we introduced a novel reinforcement learning model that could account for the contextual effect of the behavioral data better than previous related models. To see how contextual effect is different in two types of cortical and subcortical dominant behavior, we evaluated participants' behavior in two post-learning transfer phase and value estimation phase. We observed that participants behaved strongly biased in the transfer phase, while this bias was very weak in the estimation phase. This suggests that these two systems have different sensitivities to the contextual effect, such that subcortical system is more sensitive than cortical one.

To better dissociate the cortical and subcortical behavioral difference, we used reward amplitude rather the reward probability, because we assumed that complexity of reward amplitude can better engage the cortical parts of the brain in the estimation phase. Thus we could show that there is contextual effect also for reward amplitude.

Results

1 Results

Behavioral task

Two groups of participants performed two different versions of the instrumental learning tasks, the Partial feedback version, in which the factual outcomes were only provided, and the Complete feedback version, in which both factual and counterfactual outcomes were provided. Subjects were instructed to gain the most possible rewards during the task. Rewards were random independent numbers drawn from particular normal distributions. They observed two pairs of options (A_1, B) and (A_2, C), where A_1 and A_2 were associated with rewards from the same distribution of $\mathcal{N}(64, 13)$ and B and C were associated with rewards from two different distributions of $\mathcal{N}(54, 13)$ and $\mathcal{N}(44, 13)$ respectively. To conceal the task structure from the participants, although their associated values were equal, the images assigned to A_1 and A_2 were different. After the learning phase, they unexpectedly entered the post-learning transfer phase, in which all possible binary combinations of options (6 pairs) were presented to them (each combination 4 times), and they were asked to choose the option that was associated with the highest expected rewards in the preceding learning phase. With this design, if there is a bias in preferring A_2 over A_1, this transfer phase can reveal it. Similar designs were used in the context-dependent learning studies as well [9–11]. In order not to interfere with their previous learning, no feedbacks were provided in the transfer phase [9–11, 31, 32]. After each choice, they were asked to report their choice confidence on a scaled bar from 0 to 100. Finally, in the value estimation phase we asked the subjects to report their estimated expected value of each stimulus on a scaled bar ranging from 0 to 100.

Performance

To see whether the participants learned the options values during the task, at first, we calculated their performance in the learning phase which is the percentage in which they chose the advantageous option (the option with higher expected rewards). We observed that in both versions, the participants' performances were higher than 0.5. (Partial: performance $= 0.7613 \pm 0.1130$; Complete: performance $= 0.8823 \pm 0.0853$; Figure 3b). Consistent with the previous studies [9–11], we found that in the Complete version, the performance of the learning phase was significantly higher than that of the Partial version ($p = 4.5603e-07$, $tstat = 5.3522$, $df = 75$, one-tailed ttest), which means providing counterfactual outcomes facilitate learning. In addition to the learning phase, we also observed that subjects had higher performance in the transfer phase, such that participants significantly preferred the option with higher expected rewards in each combination (Partial: $p = 1.0577e-73$, $tstat = 23.4715$, $df = 348$; Complete: $p = 4.3483e-84$, $tstat = 24.7863$, $df = 418$; ttest). Additionally, the reported confidences for the most advantageous options were significantly higher than those with non-advantageous options (Partial: $p = 1.5970e-06$, $tstat = 4.9705$, $df = 173$; Complete: $p = 3.7111e-09$, $tstat = 6.1597$, $df = 208$; ttest). For these and all the following analyses, unanswered trials of the learning phase were excluded.

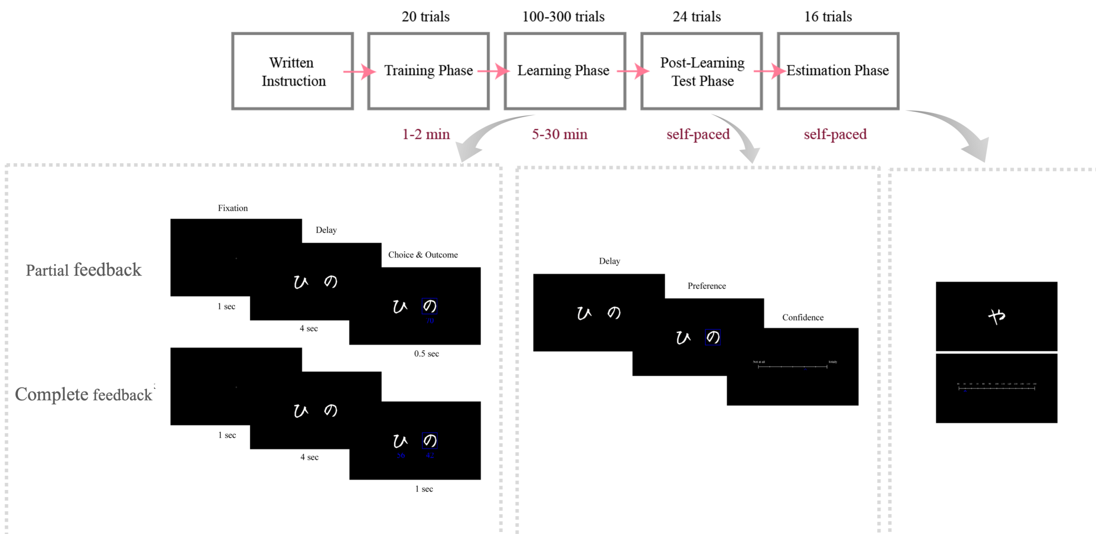


Fig 1. Behavioral Design. Time-lines of the Partial and Complete feedback versions. Subjects were instructed with written instructions and trained through 20 trials before beginning the main task. They learned two pairs of options in the Learning phase with trial and error. They transferred to the transfer phase after at least 100 and at most 300 trials, in which they were supposed to choose the most advantageous option between the two presented options, and report their choice confidence. In this phase, all possible binary combinations of options were presented.

Contextual effect

When through performance analysis, we made sure that participants learned the options' values during the learning phase, we turned to the transfer phase to see whether there is any contextual effect. Considering the first iterations of the participants' choices in the transfer phase, we found that participants' preferences between A_1 and A_2 have been significantly modulated by their distance from their paired options, such that despite having equal absolute values, participants preferred A_2 over A_1 (*transfer bias*) in both versions (Partial: $p = 0.04$, $ratio = 0.65$; Complete: $p = 0.01$, $ratio = 0.66$; binomial test) (Figure 3a). This trend remained when we consider all the four iterations of A_1 and A_2 though it loses significance (Partial: $p = 0.08$, $ratio = 0.63$; Complete: $p = 0.053$, $ratio = 0.64$; binomial test). This loss of significance might be due to strategy of balanced choice in subjects to reduce the risk for all four choices of no feedback.

To assure that the observed bias in the transfer phase belongs to the context-dependent value learning, and not to some other confounding factors, we probed which other factors could have affected the subjects' preference toward A_2. The observed bias might have occurred due to the fact that in the learning phase A_2 has been chosen more frequently than A_1. To test this possibility, we ran a logistic regression analysis to see whether the preference of A_2 over A_1 in the first (A_1, A_2) iteration of the transfer phase was due to the difference between the choice frequency of A_2 over A_1 in the learning phase. This analysis showed that the effect of choice frequency of A_2 over A_1 on the transfer bias is not significant for complete version and near significant for partial version (Partial: $p = 0.054$, $tstat = 1.92$; Complete: $p = 0.12$,

$tstat = 1.54$). Significant intercept of the regression confirms the transfer effect, even when choice frequency is controlled (Intercept, Partial: $p = 0.03$, $tstat = 2.15$; Complete: $p = 0.02$, $tstat = 2.20$).

Furthermore, we ran another logistic regression analysis to assess whether the different choice frequencies of options in the last trials of the learning phase (last 20 trials) have made the observed bias in the transfer phase. We again found no significant effect of late choice frequency on the transfer bias (Partial: $p = 0.56$, $tstat = -0.57$; Complete: $p = 0.29$, $tstat = 1.0473$) while intercepts remained near significant (Partial: $p = 0.06$, $tstat = 1.83$; Complete: $p = 0.03$, $tstat = 2.13$). The other alternative for transfer bias justification might be the amount of very small or very large rewards (rewards from the top or bottom of the reward distributions) that affected the transfer bias. Again, using logistic regression analysis, we separately tested whether sum of the observed rewards greater than $\mu + 2.5\sigma$ and less than $\mu - 2.5\sigma$ (μ and σ are the mean and standard deviation of the observed rewards, respectively) could affect the transfer bias. We found no significant effect of large and small rewards in either of the versions (large rewards: [Partial: $p = 0.40$, $tstat = 0.82$; Complete: $p = 0.62$, $tstat = 0.48$], Small rewards: [Partial: $p = 0.54$, $tstat = 0.60$; Complete: $p = 0.47$, $tstat = -0.71$]).

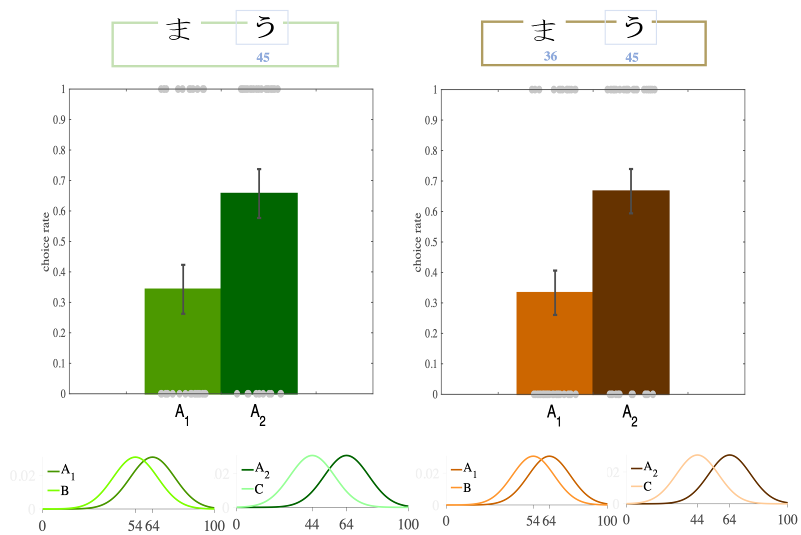


Fig 2. Transfer effect. In the transfer phase of both feedback versions, participants significantly preferred the option with higher relative value (A_2, dark green) between the two options with equal absolute values.

Value estimation

Considering the participants' first estimation in the average estimation phase, we found that participants almost precisely estimated the expected rewards of the most advantageous option based on their mean rewards, but the other non-advantageous options have been significantly underestimated (Figure 3c). This illustrates that when an option is chosen frequently, subjects could either track precisely the mean of its rewards or calculate its value at the moment of estimation. This result was also observed when we considered the participants' total estimation (the average of the four repetitions for each stimuli).

To test whether the observed contextual bias in the transfer phase, would also be

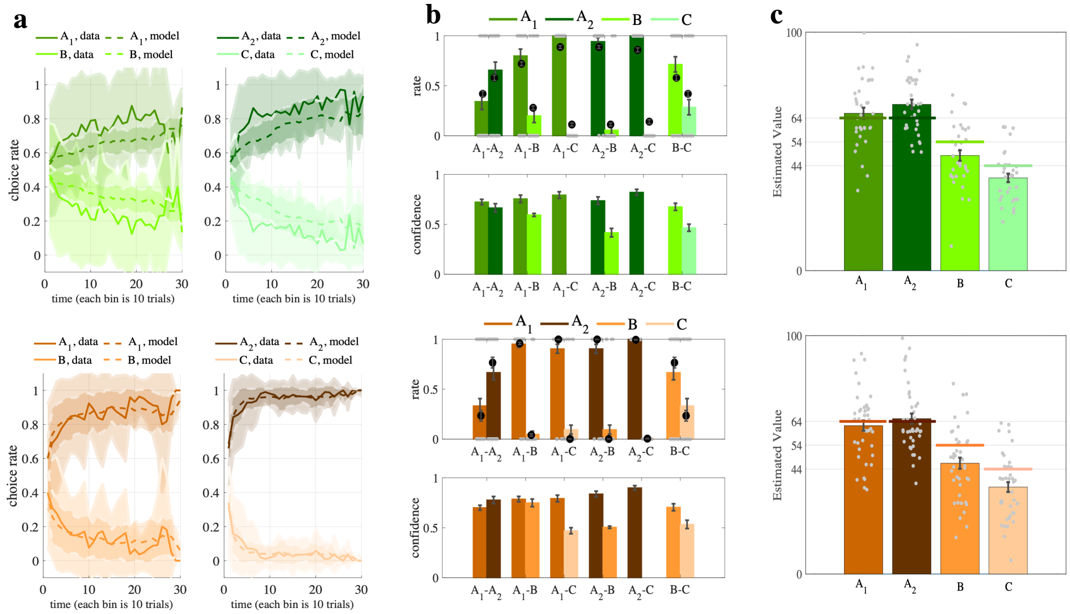


Fig 3. Behavioral results in the learning, transfer and estimation phases. a. The learning curves of two pairs of options in the learning phases of both versions show that participants learned to choose the advantageous options (A_1 in A_1B pair and A_2 in A_2C pair). The learning curve of the OL models in both versions also shows similar results. Each bin in the x-axis is the mean of the choices in 10 trials. The Partial version is green, and the Complete version is brown. Solid lines show the behavioral effect, and dashed lines show the model effect. **b.** The subjects' preferences in all 6 combinations (top), and corresponding confidences (down), with OL predictions (black dots). **c.**The value estimations of the subjects (colored bars) are very close to the real expected rewards of A_1, and A_2 options (colored lines). Partial version is green and Complete version is brown. Shadings are *SD* and error bars are *SEM*.

observed in the estimation phase, we ran a paired-ttest analysis on the estimated values. There was no significant difference between subjects' average estimation of A_1 and A_2 in both versions, yet there was a trend in overestimating A_2 compared to A_1 (Partial: $p = 0.25$, $tstats = -1.14$, $df = 34$; Complete: $p = 0.28$, $tstats = -1.08$, $df = 41$; paired-ttest). These results support the dual value-based system hypothesis, in which the subcortical system (BG) is responsible for stimulus-response association (the behavior that dominated in the transfer phase), and the cortical system (Frontal Cortex) is responsible for average reward computation (the behavior that dominated in the estimation phase) [33–37].

Comparison effect

When we observe the consequences of our decision, we compare the outcome of our decision with those alternative decisions we could have made. This comparison would trigger feelings of regret and relief if the outcome of our decision is either better or worse than those of other alternatives respectively. People naturally tend to avoid regret (approach the relief), and when one experiences regret (relief), she will likely switch to the other option (stay in the previous choice) [28, 29].

To test whether the outcomes difference of the previous trial of the same condition has influenced the switching behavior of the current trial in the learning phase, we used a hierarchical logistic regression analysis. In this analysis, we modeled the switching behavior of the subjects (1 if subject has switched, and 0 if subject has stayed on her previous choice), as a function of the outcomes difference the subject has experienced in the previous trial of the same condition, and also the difference of the expected values of the options in the current trial. The outcome difference in each trial was defined as the difference of the factual outcome and counterfactual outcome $\{r_{FC} - r_{CF}\}$ in the Complete version, and the difference of the factual outcome and the counterfactual value (expected rewards) $\{r_{FC} - V_{CF}\}$ in the Partial version. All regressors have been z-scored. While this analysis showed that there was a significant comparison effect in the Complete version, it showed no significance in the Partial version (Table 1). This means subjects tend to switch to other alternatives after experiencing regret and stay on their previous choice after experiencing relief, and this tendency is stronger in the Complete version compared to the Partial one. To investigate this effect more thoroughly, we performed a similar analysis on the logarithm of reaction times (logarithm). We observed again that, in the Complete version and not the Partial version, reaction times in each trial were significantly modulated by the outcome difference in the previous trial of the same condition, in a way that whenever the difference is smaller, the reaction time is slower, and vice versa (Table 1). This result is consistent with the post-error slowing phenomena that have been reported frequently in the decision-making literature [38, 39].

Table 1. Comparison effect of the subjects' switch behavior and reaction times.

Switch

	partial				*complete*			
Name	**Estimate**	**SE**	**tStat**	**pValue**	**Estimate**	**SE**	**tStat**	**pValue**
(Intercept)	-1.5528097	0.10551335	-14.716713	2.69E-48	-2.724902	0.17598005	-15.484153	2.28E-53
outcome difference	-0.0879529	0.0567055	-1.5510467	1.21E-01	-0.5462195	0.06292942	-8.6798744	4.68E-18
value difference	-1.123403	0.08767908	-12.812668	3.67E-37	-0.9158512	0.06505058	-14.079062	1.57E-44
condition	0.25688705	0.08761796	2.93189954	0.00337999	0.25104619	0.12809207	1.95988856	0.05004073

Reaction Time

	partial				*complete*			
Name	**Estimate**	**SE**	**tStat**	**pValue**	**Estimate**	**SE**	**tStat**	**pValue**
(Intercept)	-0.1164283	0.03073684	-3.7879077	0.00015321	-0.1211333	0.03585658	-3.3782727	0.00073263
outcome difference	0.01123051	0.00651389	1.72408744	0.08473669	-0.0164905	0.00526292	-3.1333433	0.00173402
value difference	-0.0699353	0.0101347	-6.9005836	5.64E-12	-0.0698999	0.01654412	-4.2250579	2.41E-05
condition	0.04191482	0.02390541	1.75336139	0.07958424	0.03658956	0.02364193	1.54765513	0.12174177

The hierarchical logistic regression and hierarchical simple regression analysis were performed on switch behavior and logarithms of reaction times of the subjects respectively. Along with the outcome difference as the main regressor, the current value differences between the two paired options and the condition type (A_1B, A_2C) were also included as a control regressor. The results illustrate that both current participants' choices and current reaction times were significantly influenced by the outcome differences of their previous choices in the Complete but not Partial version.

Opposing Learning model (OL)

Here, we introduce a novel reinforcement learning model, called **OL model**, adopted from the standard Q-learning model and inspired by the striatal mechanism. At first, we introduce the basic model for the Partial version, and then we extend this model for the Complete version.

Model description

This model is chosen-option centered in a way that value updating is done based on the prediction error of chosen option. Following the choice, the chosen prediction error simultaneously updates the chosen and unchosen values in an opposing manner (increasing and decreasing respectively). This mechanism is inspired by the opposing effect of dopamine on D1-SPNs and D2-SPNs neurons in the striatum, where they encode chosen and unchosen options respectively. The main reason to apply a single prediction error for updating of both options is that dopamine release is diffusive and so it is non-selective during release, thus, it will affect both D1 and D2-SPN neurons simultaneously.

$$Q_{ch} = Q_{ch} + \alpha_1 \delta_{ch}$$
$$Q_{un} = Q_{un} - \alpha_2 \delta_{ch}$$

where *ch* referred to *chosen* option, *un* referred to *unchosen* option, and $\delta_{ch} = r_{ch} - Q_{ch}$. Generally, when we refer to the OL model, we mean the OL model with two different learning rates, but in this paper, when we want to compare the two different versions of the OL model, $\alpha_1 = \alpha_2$, and $\alpha_1 \neq \alpha_2$, we call them OL_1, and OL_2 respectively. When the subject compares the options' values for making the choice, the decision is made by the softmax rule, $p(c) = \frac{1}{1+e^{\beta(Q_{un}-Q_{ch})}}$, where β is the inverse of the temperature parameter. The OL behavior is strongly dependent on the amount of α_2 relative to α_1. Based on α_2 in either of these three intervals: $0 \leq \alpha_2 < \alpha_1$, $\alpha_2 = \alpha_1$, or $\alpha_1 < \alpha_2 < 1$, the model generates a particular behavior.

OL contextual effect

In the OL model, the chosen and unchosen values are coupled, therefore their coding is not independent of each other, rather they are negatively correlated. Our simulation shows that the correlation between two paired options is proportionate to the following formula:

$$Corr(Q_1, Q_2) \approx -\frac{\alpha_2}{\alpha_1}$$

According to this formula, the amount of the correlation between Q_1 and Q_2 depends on the amount of unchosen learning rate. When α_2 changes from 0, where Q_1 and Q_2 are almost orthogonal (*corr* ≈ 0), to α_1, where Q_1 and Q_2 are almost fully correlated (*corr* ≈ -1), the encoding will change from almost fully absolute to almost fully relative (Figure 5a,b). Via simulating the experiment with typical agents of $\alpha_2 = 0$, $0 < \alpha_2 < \alpha_1$, and $\alpha_2 = \alpha_1$, we showed that we will have zero, moderate and large amount of contextual effect with never, temporary and permanent contextual effect, respectively (Figure 5c).

According to this formula, the amount of correlation between Q_1 and Q_2 depends on the amount of the unchosen learning rate. When α_2 changes from 0, where Q_1 and Q_2 are almost orthogonal (*corr* ≈ 0), to α_1, where Q_1 and Q_2 are almost fully correlated *corr* ≈ -1), the encoding will change from almost fully absolute to almost fully relative (Figure 5a,b). Through simulating the experiment with typical agents of $\alpha_2 = 0$, $0 < \alpha_2 < \alpha_1$, and $\alpha_2 = \alpha_1$, we showed that we will have zero, moderate and large amount of contextual effect with never, temporary and permanent contextual effect durations, respectively (Figure 5c, in the red box there is no contextual effect, in the yellow and green box there is a temporary moderate amount of contextual effect, and in the blue box there is a permanent large amount of contextual effect).

OL optimality

The inhibition role of the chosen prediction error on unchosen value would lead to an increase in the contrast between the competing options' values, and it leads to an increase in the performance, especially in an environment within a reasonable noise range (Figure 6a). To illustrate the performance change in the OL model, we did a simulation with a wide range of task settings, $\mu_2 - \mu_1 \in [1, 10]$, $\sigma = 1$, and a wide range

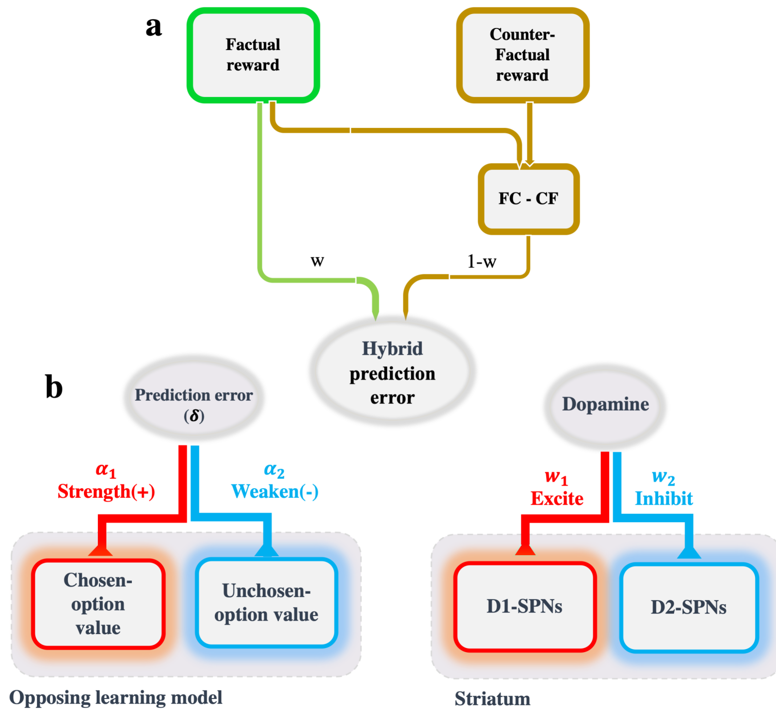


Fig 4. The schematic of the OL model and its extension. a. The comparison of the competing outcomes is a common strategy in the value learning strategy, particularly in situations where counterfactual outcomes are also provided along with factual outcomes. This comparison triggers the people's regret (relief) emotion which subsequently drives the avoidance (approach) action behavior. This tendency to minimize regret (and maximize relief) along with the tendency to maximize the expected rewards as a hybrid strategy that can account for the behavioral data is better than either of these strategies. The weights assigned to each strategy, absolute and relative, determine the amount of its effect on behavior. **b.** The idea behind the OL model comes from the opposing role of dopamine on two distinct populations of D1-SPN and D2-SPN neurons, which have been proposed to encode the chosen and unchosen options' values, by promoting the latter and inhibiting the former. Correspondingly, in the inspired model, chosen prediction error has an opposing role in updating the chosen and unchosen options' values, by strengthening the latter and weakening the former.

of parameters, $\alpha_1 \in [0, 1]$, $\alpha_2 \in [0, \alpha_1]$, and $\beta \in [0, 1]$ (for the full setting see the Methods). For the sake of simplicity, we performed all the simulation with the scaled Q-values (directly) and scaled β (inversely) so that σ was 1. By this scaling, the dynamics of the values will remain unchanged. Each of these simulations has been repeated 100 times and later averaged.

Our simulation analysis in the first step showed that the OL model as a reinforcement learning model has a better performance when the difference between competing options' values increases (Figure 1). This analysis also showed that when noise is in a reasonable range, with α_2, increasing, the performance will increase as well (relative to the SQL model, Figure 6a,b), and it means by embedding the α_2, inhibition mechanism in the model, we can have a more optimal learning behavior.

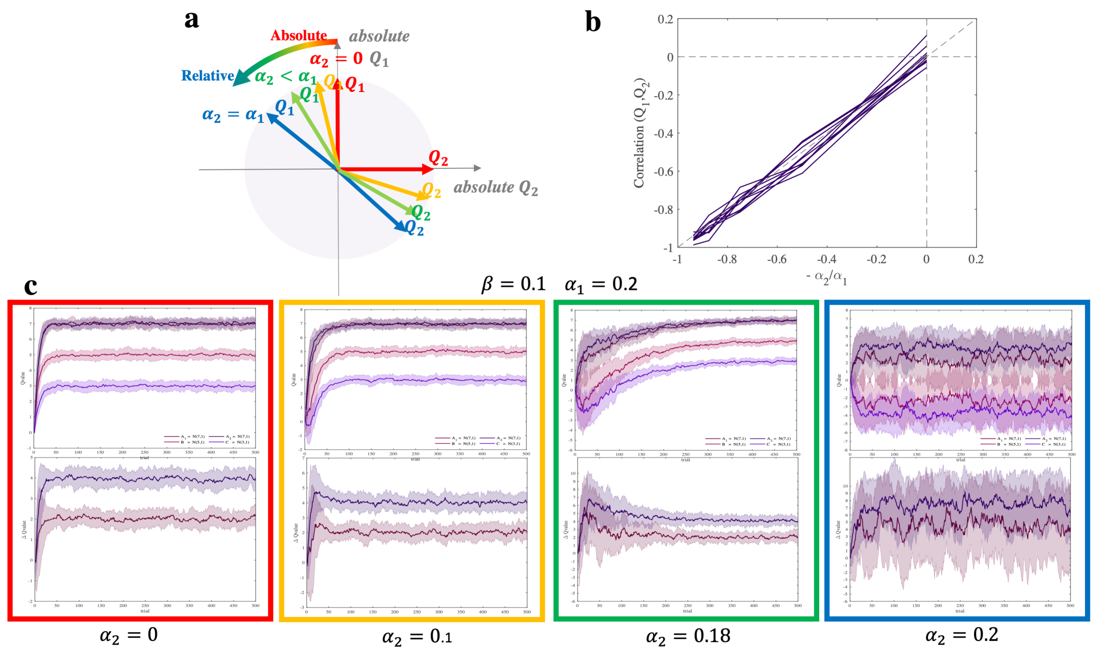


Fig 5. Correlation between two competing options' values estimated by the OL model. a. When $\alpha_2 = 0$, two estimated values are equal to their absolute values and they are orthogonal. But whenever α_2 gets closer to α_1, the estimated values in each pair become more correlated and each of them represents a stronger combination of the two absolute values. And when $\alpha_2 = \alpha_1$, estimated values are approximately fully correlated ($corr \approx -1$). **b.** The correlation between two paired options' values as a function of $-\alpha_2/\alpha_1$. **c.** The difference in the estimated values of A_1 and A_2 (contextual bias) emerges with increasing α_2. The q-values and their differences are in top and bottom parts of the figure respectively. The simulation has been done on two different pairs of options $[\mathcal{N}(7,1), \mathcal{N}(5,1)]$, and $[\mathcal{N}(7,1), \mathcal{N}(3,1)]$, with $\beta = 0.1$, $\alpha_1 = 0.2$, and four different $\alpha_2 = 0, 0.1, 0.18, 0.2$.

OL extension

The basic OL model introduced above, suggests the endogenous relative encoding in the Partial version. The main idea is the non-selective and diffusive behavior of dopaminergic signals on D1- and D2-SPN neurons. But in the Complete version there is another relativity inducing factor and that is to what extent factual outcomes are better or worse than the counterfactual outcomes. It has been shown that dopaminergic signals in the presence of counterfactual outcome differs from the standard prediction error, and it is the integration of reward and counterfactual prediction errors [30]. Furthermore, some studies have shown that by adding the outcome difference strategy to the learning procedure, the model can better account for the behavioral [9] or physiological [27] data. Therefore, we inserted the outcome difference component into the OL model to extend it for the Complete version (Figure 4b). It is worth mentioning that the outcome difference factor had a significant effect on the participants' switching behavior in the Complete version and not in the Partial version.

$$r_{abs} = r_{FC}$$
$$r_{rlt} = r_{FC} - r_{CF}$$

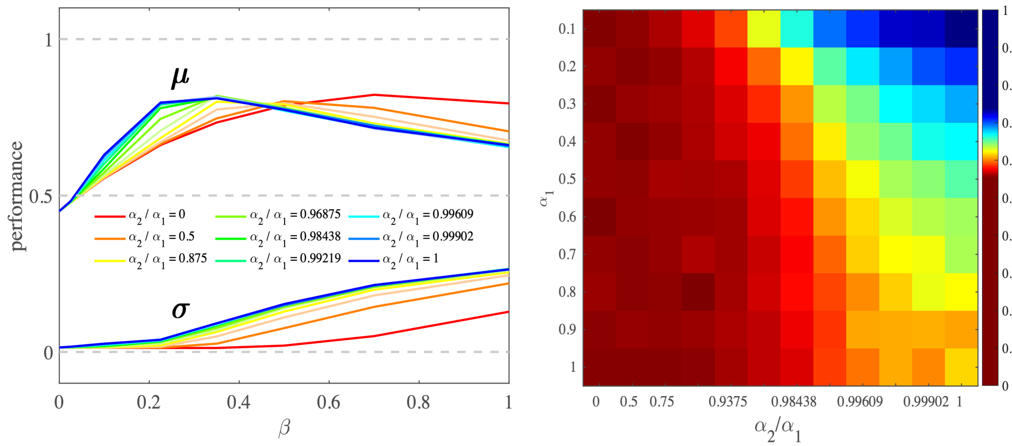


Fig 6. Performance comparison between OL and SQL model. a. As α_2/α_1 goes from 0 (SQL) to 1 (the OL_1) the peak of the performance shifts to the left, where the value of β is smaller, and also is more reasonable. In this β range. For higher α_2/α_1 the peak of performance has been reached in higher β that there is high variance in behavior. The performance has been obtained by averaging performance across all task settings and different ranges of α_2/α_1. **b.** This heat-map shows that by increasing α_2/α_1, performance will increase. This result comes from the task setting $[\mathcal{N}(10,1), \mathcal{N}(7,1)]$, and $\beta = 0.1$.

$$r_{hyb} = wr_{abs} + (1-w)r_{rlt}$$

$$\delta_{ch} = r_{hyb} - Q_{ch}$$

$$Q_{ch} = Q_{ch} + \alpha_1 \delta_{ch}$$
$$Q_{un} = Q_{un} - \alpha_2 \delta_{ch}$$

where w is the weight of absolute strategy. If the means of reward distributions of paired options are μ_1, and μ_2, and then their difference is $\mu_1 - \mu_2$, the means of the new reward distributions and their difference would be:

$$\mu_1' = \mu_1 + (1-w)\mu_2$$

$$\mu_2' = \mu_2 + (1-w)\mu_1$$

$$\mu_1' - \mu_2' = w(\mu_1 - \mu_2)$$

Using r_{hyb} in the prediction error formula seems as if we have two options with two new reward distributions, in a way that that their means get closer to each other, relative to when we use r_{FC}. Thus, this modification does not change the key OL behavior, and the extended-OL model still preserves all the above-mentioned properties. Therefore, by designing a proper prediction error, the OL model will have a good ability to be extended easily to a wide range of conditions.

Model comparison

Model fitting and model validation

In this part of the analysis, we compared the novel OL model with the related previously introduced models, in two ways: model-fitting and model-validation. We included the standard Q-learning model (SQL) as a benchmark, and the reference-point model (RP) [11], difference model [10], and hybrid model [9] as rivals in the model space. Almost all of the models had Partial and Complete versions. The OL model has two different versions, OL_1 where the chosen and unchosen options have the same learning rates, and OL_2 where they have different learning rates.

We did the fitting procedure for the learning phase of each subject and each model, and calculated their Bayesian exceedance probabilities. For the transfer phase, the negative log-likelihood were obtained by the likelihood that the model chooses the options that the subject has chosen in the transfer phase on its first iteration. Through model comparison, we found that the OL models especially OL_1 (for the Partial and Complete versions), had a better fitting criterion in the learning phase and also a better prediction criterion in the transfer phase (table 2).

In addition to model fitting analysis, we used model-validation analysis to test whether the OL model can generate the observed behavior. The simulation for each participant in each model was conducted by her best-fitted parameters, 100 times, and then were averaged. As expected, in the learning phase of both versions, agents' performances were higher than 0.5 (Partial: performance $= 0.6637 \pm 0.0627$; Complete: performance $= 0.8857 \pm 0.0639$; Figure 3b), and consistent with the behavioral results, the performance in the learning phase of the Complete version was significantly higher than that in the Partial version ($p = 4.4086e-25$, $tstat = 15.3079$, $df = 75$, one-tailed ttest). In addition to the learning phase, we also observed that the performance of the subjects was high in the transfer phase, such that participants significantly preferred the option with higher expected values (Partial: $p = 5.4079e-105$, $tstat = 31.8008$,$df = 348$; Complete: $p = 3.1818e-177$, $tstat = 49.5978$, $df = 418$; binomial test). We could also replicate the transfer effect (Figure 3a), in a way that agents preferred A_2 over A_1 in both feedback versions (Partial: $p = 0.04096$, $ratio = 0.65714$; Complete: $p = 6.8771e-05$, $ratio = 0.78571$; binomial test). This simulation analysis showed that the OL model could generate all key signatures of the behavioral data (Figure 3a,b).

Parameter recovery

To validate our model fitting, we probed the correlation between fitted and recovered parameters. For each best-fitted parameter, we performed parameter recovery for 100 distinct simulations and then averaged it. We found strong correlations between fitted and recovered parameters, ($corr \geq 0.9$) (Figure 7).

Discussion

The investigations of contextual effect on value learning have mostly focused on the putative role of counterfactual components in the Complete version. In this study, we showed that counterfactual components play an important role also in the Partial version where only factual outcomes are provided, and the counterfactual component here is the effect of chosen outcome on unchosen value. Inspired by the opposing role of dopamine on competing options' values in the striatum, we introduced a novel Opposing Learning model, in which the chosen prediction error, updates the competing options' values in an opposing manner. Unchosen value updating with chosen prediction error

Table 2. Model comparison: model fitting and model prediction.

Fitting (Learning Phase)

	SQL	RPA		Hyb	OL$_1$	OL$_2$
			Partial			
xp	$2e-05$	0		0	0.99998	0
pxp	$2.0047e-05$	$4.7129e-08$		$4.7129e-08$	0.99998	$4.7129e-08$
			Complete			
xp	0.001594	0	0.16604	0.000685	0.66409	$1e-06$
pxp	0.0024225	0.00083783	0.16591	0.0015188	0.66104	0.00083883

Prediction (Transfer Phase)

	SQL	RPA	Dif	Hyb	OL$_1$	OL$_2$
			Partial			
A_1A_2	0.69 ± 0.05	0.7 ± 0.06		0.72 ± 0.05	0.59 ± 0.04	0.64 ± 0.04
all	2.26 ± 0.14	2.24 ± 0.14		2.29 ± 0.13	2.27 ± 0.19	2.3 ± 0.19
			Complete			
A_1A_2	0.8 ± 0.07	1.12 ± 0.22	0.95 ± 0.18	1 ± 0.17	0.86 ± 0.12	0.86 ± 0.12
all	3.21 ± 0.51	3.41 ± 0.57	3.07 ± 0.52	3.41 ± 0.52	2.97 ± 0.54	2.99 ± 0.5

Fitting. Bayesian exceedance probability (xp) [40], and protected exceedance probability (pxp) [41] of the learning phase. **Prediction.** negative log likelihood (nll) of A_1A_2 and all 6 combinations of the transfer phase separately. $Mean \pm SEM$.

Table 3. Estimated parameters.

parameter	constraint	SQL	RPA	Dif	Hyb	OL$_1$	OL$_2$
				Partial			
β	$0 \leq \beta < \inf$	0.07 ± 0.03	0.12 ± 0.08		0.06 ± 0.04	0.02 ± 0.02	0.03 ± 0.02
α_1	$0 \leq \alpha_1 \leq 1$	0.25 ± 0.26	0.26 ± 0.27		0.37 ± 0.29	0.26 ± 0.2	0.32 ± 0.23
α_2	$0 < \alpha_2 \leq \alpha_1$		0.34 ± 0.3				0.21 ± 0.18
w	$0 \leq w \leq 1$				0.55 ± 0.37		
				Complete			
β	$0 \leq \beta < \inf$	0.12 ± 0.09	0.37 ± 0.24	0.37 ± 0.23	0.2 ± 0.15	0.11 ± 0.12	0.1 ± 0.1
α_1	$0 \leq \alpha_1 \leq 1$	0.14 ± 0.16	0.1 ± 0.12	0.09 ± 0.08	0.21 ± 0.15	0.22 ± 0.15	0.26 ± 0.14
α_2	$0 < \alpha_2 \leq \alpha_1$		0.11 ± 0.13				0.19 ± 0.16
α_3	$0 \leq \alpha_3 \leq 1$		0.35 ± 0.3				
w	$0 \leq w \leq 1$				0.28 ± 0.23	0.28 ± 0.17	0.32 ± 0.19

The estimated parameters for each model. $Mean \pm SD$.

will make the competing options' values correlated to each other which leads to the emergence of the contextual effect during learning. On the other hand, due to the inhibiting role of the prediction error in unchosen values, the contrast between options' values compared to the standard Q-learning model will increase, and this leads to higher performance in a reasonable exploration rate and more optimal learning than the standard way. This model could show the behavioral characteristics of the data and also by comparing it with the previous related models, it could better account for the data.

The majority of studies on instrumental learning paradigm used discrete rewards of 1 and 0 as gain and loss and subjects were supposed to estimate the probability of rewards for each option to maximize their payoffs [10, 11, 31]. But in the real world, we often experience continuous outcomes of our choices and are supposed to estimate their expected outcomes. Our secondary aim in this study was then to investigate the

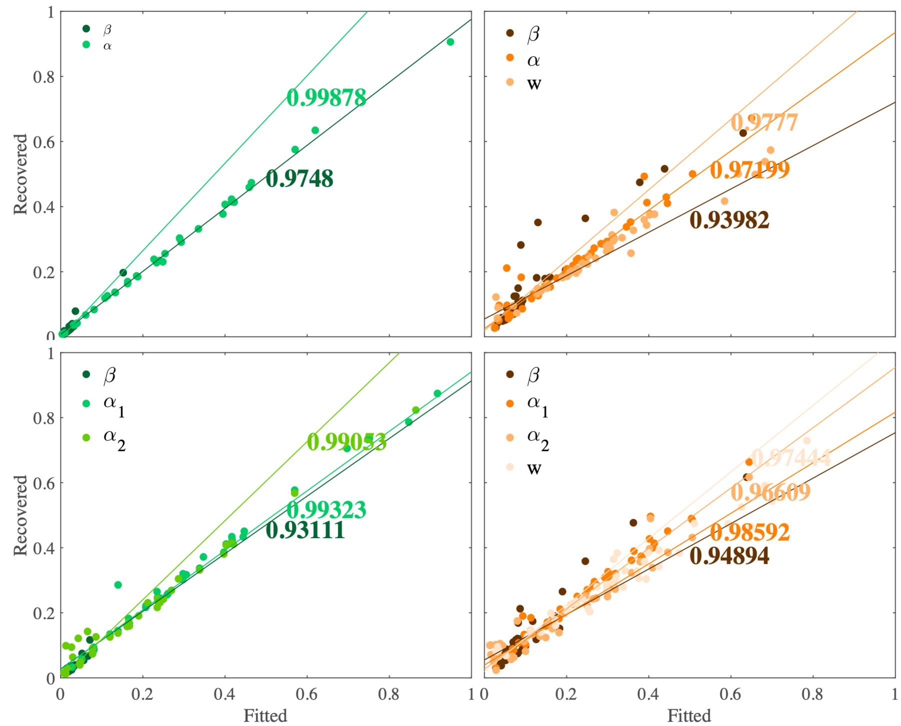


Fig 7. The correlations between fitted and recovered parameters. The OL_1 (top) and OL_2 (down) model for the Partial (left, green), and Complete (right, brown) versions. The recovery for each fitted parameter has been done 100 times and then averaged.

contextual effect in the paradigm with continuous reward amplitudes. We adopted previous instrumental learning tasks with novel reward designs, in which the stimuli were associated with some rewards drawn from specific normal distributions. With these complementary results, we could show that the contextual effect is not limited to probabilistic reward, but it extends to reward with amplitude.

Learning and decision making are two intermingled processes, and studying either of them cannot be separated from the other, as the recent evidence showed that some decision-making biases come from value learning that happens in a specific context [9–11]. The main neural underpinnings of these two processes are in the striatal circuitry in the subcortical part of the brain [42, 43]. A wide range of studies have shown a correspondence between the well-known reinforcement learning model, and striatum function [24, 26]. Dopamine is proposed to encode reward prediction errors [18, 44] and it reinforces the representations in the striatum [45], the region that has been proposed to encode options' values [12–17]. The main assumption in this model is that the chosen prediction error only affects the chosen value. But it has been proposed that the underlying function in the striatum relies on the opposing role of dopamine on two segregated populations of neurons which encode the competing options' values separately [21, 22, 25, 46]. This encouraged us to attempt to modify the standard Q-learning model to have a model more consistent with physiological evidence.

There are two direct and indirect pathways in the Basal Ganglia which have shown to have opposing roles; direct pathway promotes and indirect pathway inhibits options [24, 26]. These pathways originate from two distinct populations of neurons in the striatum, D1-SPNs, and D2-SPNs respectively, in which dopamine has an opposing influence on them, by stimulating D1-SPNs, and inhibiting D2-SPNs neurons [19, 20]. In associative learning studies, it has been shown that D1-SPNs and D2-SPNs neurons encode two opposing options' values of competing options in a two-forced choice operant learning task [21, 22, 25, 31, 46–48], in which D1-SPNs encodes the ongoing (chosen) option and D2-SPNs encodes its competing option. Being inspired by this evidence, we introduced a novel model in which chosen-related prediction error updates the chosen and unchosen value concurrently, but in an opposing manner by updating the latter and former in an increasing and decreasing manner respectively. The OpAL model have been previously introduced by Collins et al with a similar idea [48]. The main difference between OpAL and OL models is that OpAL uses reference-point mechanism explicitly, but in the OL model without explicit using of reference-point, it emerges during learning implicitly, and without adding complexity of reference point calculations, OL model predicts the behavior in a better manner.

The OL model having two concurrent associative learning for opposing actions has a good potential to explain the recent neural evidence. Several studies in different ways have shown that stimulation of D1-SPNs increases the approach behavior and decreases the avoid behavior, and stimulation of D2-SPNs increases avoid behavior and decreases approach behavior for the ongoing action. This evidence has also been shown by increasing and decreasing the amount of dopamine in the striatum [49], stimulating and inhibiting D1-SPNs and D2-SPNs by light [50], and removing the D1-SPNs and D2-SPNs activity by ablation experiment [51]. The relative activation of these two pathways encodes the internal variable of the underlying decision-making procedure [25], that can play the role of likelihood computation in the softmax rule [25] and make a bias towards the option with higher value [52, 53]. The specificity in these two pathways is similar and the amount and pattern of their activations are anti-correlated [23, 54, 55]. Similar kinds of reported evidence in decision-making paradigms have also been reported in the learning paradigms [50, 56, 57]. The opposing synaptic plasticity in these two pathways was also reported [58]. As has been shown in the Results Section, the OL model can potentially account for this evidence.

Due to being concurrently affected by chosen-related prediction error, competing options' values are encoded depending on each other. Indeed, this dependency appears as a correlation that is proportioned to $-\alpha_2/\alpha_1$. Whenever α_2 gets closer to α_1, their (absolute) correlation increases, such that when $\alpha_2 \approx 0$, the correlation is the least ($corr \approx 0$), and when $\alpha_2 = \alpha_1$ the correlation is the most ($corr \approx -1$). This correlation is also consistent with the physiological evidence which has shown that D1-SPNs and D2-SPNs neurons in the instrumental learning tasks have opposite activity with similar strength [21, 22]. Since in this model the competing options' values are anti-correlated, the OL estimated values depend on their paired options, and then this model generates the contextual effect. The amount of this contextual effect is proportioned to α_2/α_1. When $\alpha_2 = 0$, there is no contextual effect at all, when $0 < \alpha_2 < \alpha_1$, there is a moderate amount of contextual effect that is temporary and disappears over time (but in a long run). And when $\alpha_2 = \alpha_1$, there is the largest contextual effect that is permanent.

We showed that the OL model compared to its counterpart, the standard Q-learning model, has an advantage of being more optimal by having higher performance. Whenever α_2 gets closer to α_1, the performance in the environments with a reasonable amount of noise will increase, in a way that the more relative the model is, the higher is the performance. Improvement of performance is because of boosting the contrast between the options' values which leads to detect the superior options. Analogous to the OL model, there is also this kind of optimal behavior in the confirmation bias model. In this model, it is the asymmetric updating of positive and negative prediction errors for chosen and unchosen options' values that boosts the contrast between options' values [59].

It has been shown that people are not only affected by their factual rewards but also by their relative rewards that are the difference between the factual and counterfactual outcomes [27–29]. These relative outcomes are also encoded in the brain by dopamine [30, 60]. In particular, in conditions in which this comparison is available to participants, this effect is stronger and participants use the hybrid of absolute and relative strategies to learn and choose [9, 27]. In our behavioral analysis, we showed that the comparison effect is stronger in the Complete version than in the Partial one. This exogenous relativity is a different component compared to the endogenous relativity introduced by the OL model, then by inserting this factor into the model, we can expect to have better accounting for the behavioral data. As this model can be extended to any other well-defined prediction errors and preserve all its characteristics, we extend the OL model for the Complete version by inserting the outcome comparison part to it. This embedding could better explain the Complete version.

Substantial evidence demonstrates that two separate and parallel systems are involved in decision-making and learning, the Basal Ganglia and Frontal cortex, in which the Basal Ganglia plays a critical role for habitual behavior and the Frontal cortex plays a critical role in the goal directed behavior [61]. It is the weighted combination of these two systems that are involved in people's behavior. It has been shown that several factors modulate these weights [62–72]. Different amount of contextual effect in the learning, transfer and estimation phases are in line with this hypothesis. In each phase of the task, participants have different needs. In the learning phase, to gain more rewards, they need to know how much an option is better than its competing option. We expect to see that the learning phase strategy is reflected in the transfer phase, where they are supposed to continue to choose between pairs of options. Finally, in the estimation phase, in contrast to previous phases, they need to know the exact absolute values. According to these needs, we expect to have the most BG, the least FC weights in the learning phase, the modest BG, and FC weights in the transfer phase, and the least BG, the most FC weight in the estimation phase [73].

Taken together, in this paper we could show that we are affected by the context by the fine interaction of counterfactual outcomes. In the two-option learning tasks, we learn the value of each option relative to its alternative, even when we don't explicitly use the comparison strategy. On the other hand, although this contextual effect results in suboptimal decision-making outside the original context, it leads to an ecological advantage by gaining more rewards within the original context. Furthermore, and not surprisingly, people can access to both relative and absolute estimations of their options' values, and to use which of them depends on their needs and conditions. Like other contextual biases and irrationalities in the human behavior, this bias seems to have an advantage for people to use. Investigating the mechanism of these irrationalities helps us find a solution in conditions where advantages change into disadvantages, and it will be more critical when they change to disorder.

Materials and methods

Participants

Two groups of 41 and 47 subjects have participated for the Partial and Complete versions of our task respectively. We excluded 6 subjects from the Partial version and 5 subjects from the Complete version (2 and 3 subjects because they didn't learn the associations, and 4 and 2 subjects because their expected rewards for A_1, and A_2 were more than one, in Partial and Complete versions respectively, see below). After exclusion $N = 35$ subjects (age: 26 ± 6 ($mean \pm SD$), female: $n = 16$) and $N = 42$ subjects (age: 23 ± 5 ($mean \pm SD$), female: $n = 12$) remained for analysis in the Partial and Complete versions respectively. They were received their monetary rewards after they completed the task, according to their performances. They were all healthy volunteers that gave a written consent before starting the task. The study was approved by the local ethics committee.

Behavioral task

Two different cohorts of participants performed two different versions of instrumental learning tasks, which were adopted from previous studies [9–11]. The main structure of these two tasks was almost the same and included two consecutive phases of learning and post-learning transfer. The only difference was in the way feedbacks were provided to the subjects. In the Partial version, only the factual outcomes for chosen option were provided to the subjects, and in the Complete version, both the factual and counterfactual outcomes for chosen and unchosen options, respectively, were provided. Before the main task, subjects performed a short training session (20 trials) to be familiarized with the learning phase. The stimuli and the reward statistics of the training session were different from those of the main session. The stimuli were selected from the Japanese Hiragana alphabet.

The learning phase was made up of one session in which, in each trial two stimuli were presented on the screen, and participants were instructed to choose the option with higher expected rewards. This instrumental learning paradigm made participants to learn gradually by trial and error to choose the most advantageous option in each trial. The cues were shown to the subjects from two pairs of stimuli $\{A_1B, A_2C\}$, which means in each pair each stimulus was always presented with a similar stimulus. Each pair thus established a fixed context. These two contexts were pseudo-randomly interleaved across trials. The rewards of A_1 and A_2 stimuli were drawn from the same normal distribution of $\mathcal{N}(64, 13)$ and the rewards of B and C stimuli were drawn from a different normal distributions of $\mathcal{N}(54, 13)$ and $\mathcal{N}(44, 13)$, respectively. To control some confounding factors, rewards samples were drawn from the truncated distribution, which was in the $[\mu - 3\sigma, \mu + 3\sigma]$ ($[0, 100]$) interval. The parameters of the distributions were unknown to the subjects, and they were supposed to learn them. Although the reward statistics of A_1 and A_2 were the same, the images associated with them were different to conceal the task structure from the subjects.

The side of each stimulus on the screen, whether the right of the fixation point or the left, was also pseudo-randomized during the task, such that for the total number of trials for each context, in half of the trials a particular stimulus was presented on the right and in the other half, on the left. The subjects were asked to select their choices within a 4000 ms, otherwise, they missed that trial's reward, and the 'No Response' message was shown on the screen. Within each trial, the subjects chose their choice by pressing the left and right arrow keys for the left and right options respectively. Following the choice, the chosen option was surrounded with a blue square and the related outcomes were presented simultaneously on the screen. In the Partial version, the factual outcome was shown below the chosen option for 500 ms and in the Complete version, both the factual and counterfactual outcomes were shown below the chosen and unchosen options respectively for 1000 ms. In the Complete version, the information that subjects should process was two times the Partial version and in our pilot study, we found that having only 500 ms for observing the outcomes was not sufficient to process two continuous outcomes and so decreased the subjects' performance compared to the Partial version, therefore we doubled this time to 1000 ms. The next trial started after 1000 ms fixation screen. Each context was presented to the subjects at least in 50 trials and then two contexts consist of, at least 100 trials. After at least 100 trials, the task continued for each subject until the experienced mean of A_1 became almost equal to the experienced mean of A_2, (their difference became less than 1). If this condition was not met up in the 300th trial then the learning phase was stopped and this subject was excluded from the data. By this design, the number of trials always falls into the range of [100,300] and this number might be different for each subject.

Seamlessly after the learning phase, participants entered the post-learning transfer phase. They were not aware of the transfer phase until they completed the learning phase, in order not to use any memorizing strategy in the learning phase. In the transfer phase, all possible binary combinations of the stimuli (6 combinations) were presented to the participants and they were asked to choose the option with higher expected rewards. They were told that they will not only see the previously paired options in the learning phase but even the binary options which weren't paired in the preceding phase. Each combination was presented four times, giving a total $6 * 4 = 24$ trials that were presented in a pseudo-randomized order. This phase in contrast to the learning phase was self-paced (they were not force to choose in a limited time) and also no feedback was provided to the subjects, in order not to interfere with their learned values [9–11, 31, 32]. Following each choice, they had to report the confidence of their choice by using a scaled bar from 0 to 100 in which the leftmost side of the axis shows complete uncertain and the rightmost side shows complete certain. The confidence part was done by the mouse. After the transfer phase, subjects completed the estimation phase. In the estimation phase, stimuli were presented to the subjects one by one and they were asked to estimate its mean of rewards, using a scaled bar from 0 to 100. Each stimulus was repeated four times giving a total of $4 * 4 = 16$ trials which were presented pseudo-randomly. These trials were also self-paced and no feedback were provided to the subjects. The subjects were told their payoffs are based on the sum of rewards they would gain during the learning task. In the Complete version, subjects were notified that their total rewards are only based on the rewards of their choices. Although they were not paid in the transfer phase, they were encouraged to do as best as they can to answer correctly as if they would be paid. At the end of the task, their total rewards were shown on the screen.

Computational models

The Standard Q-Learning (SQL) Model

It is a common approach to compare the context-dependent learning models with the standard Q-learning model as a benchmark that plays the role of absolute learning model. In this model, the value of each option is only related to its own observed outcomes and not to other alternative outcomes.

$$\delta_{ch} = r_{ch} - Q_{ch}$$
$$Q_{ch} = Q_{ch} + \alpha\delta_{ch}$$

In the simplest form, it is only the chosen option which is updated following its outcomes observation, while in its extended form the unchosen options are also updated, but again with their own observed outcomes:

$$\delta_{ch} = r_{ch} - Q_{ch}$$
$$Q_{ch} = Q_{ch} + \alpha_1\delta_{ch}$$
$$\delta_{un} = r_{un} - Q_{un}$$
$$Q_{un} = Q_{un} + \alpha_2\delta_{un}$$

Which their learning rates can be the same or different ($\alpha_1 = \alpha_2$ or $\alpha_1 \neq \alpha_2$).

The Reference-Point (RP) Model

The idea of the reference-point (RP) model comes from the reference point phenomenon which is reported by behavioral and economic studies [74, 75]. According to this model, there is a distinct reference-point for each context that is obtained by its expected outcomes. Then, the relative outcome of each option is calculated in comparison to this reference-point. We implemented several forms of RP models considering the several forms of context reward [11]. The RPD, RPA, and RPM, when the contextual rewards, r_x, are considered to be direct r_{ch}, average of $(r_{ch} + Q_{un})/2$, and $\max(r_{ch}, Q_{un})$ respectively in the Partial version, and r_{ch}, $(r_{ch} + r_{un})/2$, and $\max(r_{ch}, r_{un})$ in the Complete version.

$$\delta_x = r_x - V_x$$
$$V_x = V_x + \alpha_1 \delta_x$$
$$\delta_{ch} = (r_{ch} - V_x) - Q_{ch}$$
$$Q_{ch} = Q_{ch} + \alpha_1 \delta_{ch}$$

where V_x is the value of the context, and Q_{ch} is the value of the chosen option. For the Complete version, we also update the unchosen options as below,

$$\delta_{un} = (r_{un} - V_x) - Q_{un}$$
$$Q_{un} = Q_{un} + \alpha_2 \delta_{un}$$

In the Complete version, we used different versions for RP. One which only updates the chosen value, and one which updates both options with the same and different learning rates.

The Difference (Dif) Model

Learning in a specific context in which a participant is supposed to maximize her rewards needs using a strategy in order to find a better option as soon as possible. The difference model is one of the models which gives a fast detection of the advantageous option by learning the relative value. In this model, the participants learn how much the superior option is better than the inferior one [10].

$$r_{rlt} = r_{FC} - r_{CF}$$
$$\delta = r_{rlt} - Q_{ch}$$
$$Q_{ch} = Q_{ch} + \alpha \delta$$

This model was only applied for the Complete version.

The Hybrid (Hyb) Model

It has been shown that people are not fully absolute or fully relative learners, rather they are hybrid learners in which their behaviors depend on how much they weigh either of these strategies [9].

$$r_{abs} = r_{FC}, \quad r_{rlt} = r_{FC} - r_{CF}$$
$$r_{hyb} = w r_{abs} + (1 - w) r_{rlt}$$
$$\delta = r_{hyb} - Q_{ch}$$
$$Q_{ch} = Q_{ch} + \alpha \delta$$

For the Partial version, we used the Q_{un} instead of r_{CF}.

The Opposing Learning (OL) Model

The OL model has been inspired by the opposing role of dopamine as prediction error on the chosen and unchosen options. In this model, the unchosen option is updated simultaneously with the chosen option and proportional to the chosen prediction error, but in an opposite manner.

$$\delta_{ch} = r_{ch} - Q_{ch}$$
$$Q_{ch} = Q_{ch} + \alpha_1 \delta_{ch}$$
$$Q_{un} = Q_{un} - \alpha_2 \delta_{ch}$$

In this model, the α_2 parameter controls the amount of contextual effect on the value learning procedures. For the Complete version, this model was extended to a version in which the counterfactual outcomes were considered in a hybrid manner.

$$r_{abs} = r_{FC}, \quad r_{rlt} = r_{FC} - r_{CF}$$
$$r_{hyb} = w r_{abs} + (1-w) r_{rlt}$$
$$\delta = r_{hyb} - Q_{ch}$$
$$Q_{ch} = Q_{ch} + \alpha_1 \delta$$
$$Q_{un} = Q_{un} - \alpha_2 \delta$$

Pure simulation procedure

The OL behavior has been examined in a wide range of task and parameter settings. Without loss of generality, we did the simulation with normalized settings such that we had $\sigma = 1$ in reward distributions. As an example, the normalized version of the setting of task $\mathcal{N}(\mu = 64, \sigma = 10)$, parameters of $\beta = 0.01$, and any α_1, α_2, changes to its normalized version of $\mathcal{N}(\mu = 6.4, \sigma = 1)$ (divide by 10), and parameters of $\beta = 0.1$ (multiply by 10), and the same α_1, α_2. The tasks settings covered 10 different pairs of options in which their relative values were covered $\{1, 2, ..., 10\}$ ($[\mu_1, \mu_2] \in \{[10, 9], [10, 8], ..., [10, 0]\}$, and $\sigma = 1$). The parameters settings covered a wide range of β: $\{0, 0.025, 0.05, 0.075, 0.1, 0.1025, ..., 0.4\} \cup \{0.5, 0.6, ..., 1\}$, α_1: $\{0.1, 0.2, ..., 1\}$, and α_2/α_1: $\{0, 0.5, 0.75, 0.875, 0.93, 0.96, 0.980.992, 0.996, 0.998, 0.999, 1\}$.

Fitting and simulation procedure

The data fitting was implemented by *fmincon* function of Matlab software (the MathWorks Inc., Natick, MA). The fittings have been done with several initial points to have higher probability in order to find a global optimum, rather than getting stuck on a local optimum. For obtaining the exceedance probabilities (xp) [40], and protected exceedance probabilities(pxp) [41] for the model-comparison part, and estimating parameters, we optimized maximum a posteriori (MAP) using weakly informative priors of $\beta(1.2, 1.2)$ for each parameter. It is worth noting that the range of options' values is in scale of 100, and so range of the β parameter will be in scale of much less than one, thus, the $\beta(1.2, 1.2)$ would be a proper prior in model fitting. The exceedance probability and protected exceedance probability have calculated based on [40, 41]. The simulation for each subject was done on its best fitted parameters for 100 repetitions, and then the representative behavior of this agent was obtained by averaging across its repetitions.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper.

References

1. Sutton RS, Barto AG. Reinforcement learning: An introduction. MIT press; 2018.
2. Louie K, Martino BD. Chapter 24 - The Neurobiology of Context-Dependent Valuation and Choice. In: Glimcher PW, Fehr E, editors. Neuroeconomics (Second Edition). second edition ed. San Diego: Academic Press; 2014. p. 455 – 476.
3. Summerfield C, Tsetsos K. Do humans make good decisions? Trends in cognitive sciences. 2015;19(1):27–34.
4. Tsetsos K, Moran R, Moreland J, Chater N, Usher M, Summerfield C. Economic irrationality is optimal during noisy decision making. Proceedings of the National Academy of Sciences. 2016;113(11):3102–3107.
5. Rigoli F, Chew B, Dayan P, Dolan RJ. Learning contextual reward expectations for value adaptation. Journal of cognitive neuroscience. 2018;30(1):50–69.
6. Rigoli F, Rutledge RB, Dayan P, Dolan RJ. The influence of contextual reward statistics on risk preference. NeuroImage. 2016;128:74–84.
7. Rigoli F, Friston KJ, Dolan RJ. Neural processes mediating contextual influences on human choice behaviour. Nature communications. 2016;7(1):1–11.
8. Rigoli F, Mathys C, Friston KJ, Dolan RJ. A unifying Bayesian account of contextual effects in value-based choice. PLoS computational biology. 2017;13(10):e1005769.
9. Bavard S, Lebreton M, Khamassi M, Coricelli G, Palminteri S. Reference-point centering and range-adaptation enhance human reinforcement learning at the cost of irrational preferences. Nature communications. 2018;9(1):1–12.
10. Klein TA, Ullsperger M, Jocham G. Learning relative values in the striatum induces violations of normative decision making. Nature communications. 2017;8:16033.
11. Palminteri S, Khamassi M, Joffily M, Coricelli G. Contextual modulation of value signals in reward and punishment learning. Nature communications. 2015;6(1):1–14.
12. Samejima K, Ueda Y, Doya K, Kimura M. Representation of action-specific reward values in the striatum. Science. 2005;310(5752):1337–1340.
13. Lau B, Glimcher PW. Value representations in the primate striatum during matching behavior. Neuron. 2008;58(3):451–463.

14. Kim H, Sul JH, Huh N, Lee D, Jung MW. Role of striatum in updating values of chosen actions. Journal of neuroscience. 2009;29(47):14701–14712.

15. Ito M, Doya K. Validation of decision-making models and analysis of decision variables in the rat basal ganglia. Journal of Neuroscience. 2009;29(31):9861–9874.

16. Ito M, Doya K. Distinct neural representation in the dorsolateral, dorsomedial, and ventral parts of the striatum during fixed-and free-choice tasks. Journal of Neuroscience. 2015;35(8):3499–3514.

17. Cai X, Kim S, Lee D. Heterogeneous coding of temporally discounted values in the dorsal and ventral striatum during intertemporal choice. Neuron. 2011;69(1):170–182.

18. Schultz W, Dayan P, Montague PR. A neural substrate of prediction and reward. Science. 1997;275(5306):1593–1599.

19. Shen W, Flajolet M, Greengard P, Surmeier DJ. Dichotomous dopaminergic control of striatal synaptic plasticity. Science. 2008;321(5890):848–851.

20. Surmeier DJ, Ding J, Day M, Wang Z, Shen W. D1 and D2 dopamine-receptor modulation of striatal glutamatergic signaling in striatal medium spiny neurons. Trends in neurosciences. 2007;30(5):228–235.

21. Shin JH, Kim D, Jung MW. Differential coding of reward and movement information in the dorsomedial striatal direct and indirect pathways. Nature communications. 2018;9(1):1–14.

22. Donahue CH, Liu M, Kreitzer A. Distinct value encoding in striatal direct and indirect pathways during adaptive learning. bioRxiv. 2018; p. 277855.

23. Markowitz JE, Gillis WF, Beron CC, Neufeld SQ, Robertson K, Bhagat ND, et al. The striatum organizes 3D behavior via moment-to-moment action selection. Cell. 2018;174(1):44–58.

24. Peak J, Hart G, Balleine BW. From learning to action: the integration of dorsal striatal input and output pathways in instrumental conditioning. European Journal of Neuroscience. 2019;49(5):658–671.

25. Tai LH, Lee AM, Benavidez N, Bonci A, Wilbrecht L. Transient stimulation of distinct subpopulations of striatal neurons mimics changes in action value. Nature neuroscience. 2012;15(9):1281.

26. Cox J, Witten IB. Striatal circuits for reward learning and decision-making. Nature Reviews Neuroscience. 2019;20(8):482–494.

27. Coricelli G, Dolan RJ, Sirigu A. Brain, emotion and decision making: the paradigmatic example of regret. Trends in cognitive sciences. 2007;11(6):258–265.

28. Coricelli G, Critchley HD, Joffily M, O'Doherty JP, Sirigu A, Dolan RJ. Regret and its avoidance: a neuroimaging study of choice behavior. Nature neuroscience. 2005;8(9):1255–1262.

29. Camille N, Coricelli G, Sallet J, Pradat-Diehl P, Duhamel JR, Sirigu A. The involvement of the orbitofrontal cortex in the experience of regret. Science. 2004;304(5674):1167–1170.

30. Kishida KT, Saez I, Lohrenz T, Witcher MR, Laxton AW, Tatter SB, et al. Subsecond dopamine fluctuations in human striatum encode superposed error signals about actual and counterfactual reward. Proceedings of the National Academy of Sciences. 2016;113(1):200–205.

31. Frank MJ, Seeberger LC, O'reilly RC. By carrot or by stick: cognitive reinforcement learning in parkinsonism. Science. 2004;306(5703):1940–1943.

32. Frank MJ, Moustafa AA, Haughey HM, Curran T, Hutchison KE. Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. Proc Natl Acad Sci U S A. 2007;104(41):16311–6. doi:10.1073/pnas.0706111104.

33. Ratcliff R, Frank MJ. Reinforcement-based decision making in corticostriatal circuits: mutual constraints by neurocomputational and diffusion models. Neural computation. 2012;24(5):1186–1229.

34. Frank MJ, Doll BB, Oas-Terpstra J, Moreno F. Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. Nature neuroscience. 2009;12(8):1062.

35. Hazy TE, Frank MJ, O'Reilly RC. Towards an executive without a homunculus: computational models of the prefrontal cortex/basal ganglia system. Philosophical Transactions of the Royal Society B: Biological Sciences. 2007;362(1485):1601–1613.

36. O'Reilly RC, Frank MJ. Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. Neural computation. 2006;18(2):283–328.

37. Frank MJ, Loughry B, O'Reilly RC. Interactions between frontal cortex and basal ganglia in working memory: a computational model. Cognitive, Affective, & Behavioral Neuroscience. 2001;1(2):137–160.

38. Notebaert W, Houtman F, Van Opstal F, Gevers W, Fias W, Verguts T. Post-error slowing: an orienting account. Cognition. 2009;111(2):275–279.

39. Jentzsch I, Dudschig C. Short Article: Why do we slow down after an error? Mechanisms underlying the effects of posterror slowing. Quarterly Journal of Experimental Psychology. 2009;62(2):209–218.

40. Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ. Bayesian model selection for group studies. Neuroimage. 2009;46(4):1004–1017.

41. Rigoux L, Stephan KE, Friston KJ, Daunizeau J. Bayesian model selection for group studies—revisited. Neuroimage. 2014;84:971–985.

42. O'Doherty J, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. science. 2004;304(5669):452–454.

43. Doya K. Complementary roles of basal ganglia and cerebellum in learning and motor control. Current opinion in neurobiology. 2000;10(6):732–739.

44. Glimcher PW. Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. Proceedings of the National Academy of Sciences. 2011;108(Supplement 3):15647–15654. doi:10.1073/pnas.1014269108.

45. Niv Y. Reinforcement learning in the brain. Journal of Mathematical Psychology. 2009;53(3):139–154.

46. Nonomura S, Nishizawa K, Sakai Y, Kawaguchi Y, Kato S, Uchigashima M, et al. Monitoring and updating of action selection for goal-directed behavior through the striatal direct and indirect pathways. Neuron. 2018;99(6):1302–1314.

47. Bariselli S, Fobbs W, Creed M, Kravitz A. A competitive model for striatal action selection. Brain research. 2019;1713:70–79.

48. Collins AG, Frank MJ. Opponent actor learning (OpAL): Modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. Psychological review. 2014;121(3):337.

49. Ryan MB, Bair-Marshall C, Nelson AB. Aberrant striatal activity in parkinsonism and levodopa-induced dyskinesia. Cell reports. 2018;23(12):3438–3446.

50. Yttri EA, Dudman JT. Opponent and bidirectional control of movement velocity in the basal ganglia. Nature. 2016;533(7603):402–406.

51. Lemos JC, Friend DM, Kaplan AR, Shin JH, Rubinstein M, Kravitz AV, et al. Enhanced GABA transmission drives bradykinesia following loss of dopamine D2 receptor signaling. Neuron. 2016;90(4):824–838.

52. Isoda M, Hikosaka O. Cortico-basal ganglia mechanisms for overcoming innate, habitual and motivational behaviors. European Journal of Neuroscience. 2011;33(11):2058–2069.

53. Hikosaka O, Nakamura K, Nakahara H. Basal ganglia orient eyes to reward. Journal of neurophysiology. 2006;95(2):567–584.

54. Klaus A, Martins GJ, Paixao VB, Zhou P, Paninski L, Costa RM. The spatiotemporal organization of the striatum encodes action space. Neuron. 2017;95(5):1171–1180.

55. Parker JG, Marshall JD, Ahanonu B, Wu YW, Kim TH, Grewe BF, et al. Diametric neural ensemble dynamics in parkinsonian and dyskinetic states. Nature. 2018;557(7704):177–182.

56. Kravitz AV, Tye LD, Kreitzer AC. Distinct roles for direct and indirect pathway striatal neurons in reinforcement. Nature neuroscience. 2012;15(6):816–818.

57. Lobo MK, Covington HE, Chaudhury D, Friedman AK, Sun H, Damez-Werno D, et al. Cell type–specific loss of BDNF signaling mimics optogenetic control of cocaine reward. Science. 2010;330(6002):385–390.

58. Shan Q, Ge M, Christie MJ, Balleine BW. The acquisition of goal-directed actions generates opposing plasticity in direct and indirect pathways in dorsomedial striatum. Journal of Neuroscience. 2014;34(28):9196–9201.

59. Lefebvre G, Summerfield C, Bogacz R. A normative account of confirmatory biases during reinforcement learning. bioRxiv. 2020;.

60. Lak A, Stauffer WR, Schultz W. Dopamine neurons learn relative chosen value from probabilistic rewards. Elife. 2016;5:e18044.

61. Doll BB, Shohamy D, Daw ND. Multiple memory systems as substrates for multiple decision systems. Neurobiology of learning and memory. 2015;117:4–13.

62. Daw ND, Niv Y, Dayan P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci. 2005;8(12):1704–11. doi:10.1038/nn1560.

63. Keramati M, Dezfouli A, Piray P. Speed/accuracy trade-off between the habitual and the goal-directed processes. PLoS computational biology. 2011;7(5):e1002055.

64. Otto AR, Raio CM, Chiang A, Phelps EA, Daw ND. Working-memory capacity protects model-based learning from stress. Proceedings of the National Academy of Sciences. 2013;110(52):20941–20946.

65. Wunderlich K, Smittenaar P, Dolan RJ. Dopamine enhances model-based over model-free choice behavior. Neuron. 2012;75(3):418–424.

66. Gillan CM, Otto AR, Phelps EA, Daw ND. Model-based learning protects against forming habits. Cognitive, Affective, & Behavioral Neuroscience. 2015;15(3):523–536.

67. Otto AR, Skatova A, Madlon-Kay S, Daw ND. Cognitive Control Predicts Use of Model-based Reinforcement Learning. Journal of Cognitive Neuroscience. 2015;27(2):319–333. doi:10.1162/jocn_a_00709.

68. Lucantonio F, Caprioli D, Schoenbaum G. Transition from 'model-based' to 'model-free' behavioral control in addiction: Involvement of the orbitofrontal cortex and dorsolateral striatum. Neuropharmacology. 2014;76:407 – 415. doi:https://doi.org/10.1016/j.neuropharm.2013.05.033.

69. Radenbach C, Reiter AMF, Engert V, Sjoerds Z, Villringer A, Heinze HJ, et al. The interaction of acute and chronic stress impairs model-based behavioral control. Psychoneuroendocrinology. 2015;53:268 – 280. doi:https://doi.org/10.1016/j.psyneuen.2014.12.017.

70. Smittenaar P, FitzGerald THB, Romei V, Wright ND, Dolan RJ. Disruption of Dorsolateral Prefrontal Cortex Decreases Model-Based in Favor of Model-free Control in Humans. Neuron. 2013;80(4):914 – 919. doi:https://doi.org/10.1016/j.neuron.2013.08.009.

71. Doll BB, Bath KG, Daw ND, Frank MJ. Variability in Dopamine Genes Dissociates Model-Based and Model-Free Reinforcement Learning. Journal of Neuroscience. 2016;36(4):1211–1222. doi:10.1523/JNEUROSCI.1901-15.2016.

72. Maier SU, Makwana AB, Hare TA. Acute Stress Impairs Self-Control in Goal-Directed Choice by Altering Multiple Functional Connections within the Brain's Decision Circuits. Neuron. 2015;87(3):621 – 631. doi:https://doi.org/10.1016/j.neuron.2015.07.005.

73. Gold JM, Waltz JA, Matveeva TM, Kasanova Z, Strauss GP, Herbener ES, et al. Negative symptoms and the failure to represent the expected reward value of actions: behavioral and computational modeling evidence. Archives of general psychiatry. 2012;69(2):129–138.

74. De Martino B, Kumaran D, Holt B, Dolan RJ. The neurobiology of reference-dependent value computation. J Neurosci. 2009;29(12):3833–42. doi:10.1523/JNEUROSCI.4832-08.2009.

75. Baucells M, Weber M, Welfens F. Reference-point formation and updating. Management Science. 2011;57(3):506–519.

Supporting information

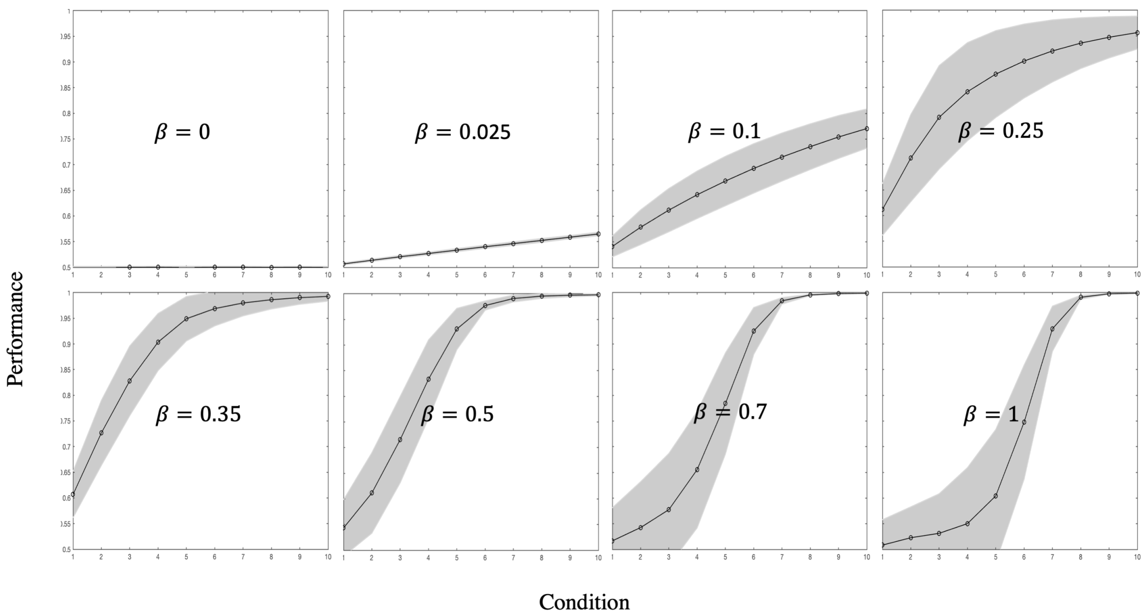
S1 Table. Model comparison with BIC.

	nll	BIC		
Partial				
	learning	**learning**	**learning + transfer(A_1A_2)**	**learning + transfer(all)**
SQL	88.18 ± 5.49	186.82 ± 11.05	188.21 ± 11.05	191.37 ± 11.03
RPD	87.17 ± 5.49	190.04 ± 11.11	191.48 ± 11.1	194.69 ± 11.1
RPA	87.69 ± 5.47	191.07 ± 11.06	192.51 ± 11.07	195.62 ± 11.06
RPM	87.18 ± 5.49	190.05 ± 11.11	191.48 ± 11.1	194.69 ± 11.1
Hyb	86.68 ± 5.48	189.05 ± 11.07	190.5 ± 11.08	193.7 ± 11.06
$\mathbf{OL_1}$	84.7 ± 5.49	179.86 ± 11.06	181.05 ± 11.06	184.45 ± 10.99
$\mathbf{OL_2}$	83.66 ± 5.37	183.01 ± 10.86	184.3 ± 10.86	187.73 ± 10.8

	nll	BIC		
Complete				
	learning	**learning**	**learning + transfer(A_1A_2)**	**learning + transfer(all)**
SQL	54.34 ± 4.98	119.24 ± 9.99	120.84 ± 10.01	125.72 ± 9.95
$\mathbf{QL_{21}}$	51.71 ± 4.99	113.98 ± 10.01	116.19 ± 9.98	121.69 ± 9.95
$\mathbf{QL_{22}}$	50.11 ± 5.01	116.05 ± 10.08	118.4 ± 10.03	124.06 ± 10
$\mathbf{RPA_1}$	51.71 ± 4.99	119.25 ± 10.03	121.18 ± 9.98	125.53 ± 9.9
$\mathbf{RPA_2}$	48.45 ± 4.99	118 ± 10.05	120.26 ± 9.98	124.96 ± 9.87
$\mathbf{RPM_1}$	51.71 ± 4.99	119.25 ± 10.03	120.84 ± 10	125.55 ± 9.92
$\mathbf{RPM_2}$	47.81 ± 5	116.73 ± 10.07	118.59 ± 10.03	124.54 ± 9.86
Dif	51.71 ± 4.99	113.98 ± 10.01	115.89 ± 9.95	120.18 ± 9.87
Hyb	48.94 ± 5	113.72 ± 10.05	115.73 ± 10.02	120.63 ± 9.91
$\mathbf{OL_1}$	47.98 ± 5	111.79 ± 10.05	113.52 ± 10.02	117.83 ± 9.93
$\mathbf{OL_2}$	47.53 ± 4.96	116.17 ± 9.99	117.9 ± 9.97	122.27 ± 9.88

BIC of three different parts, learning phase, learning and (A_1A_2) of the transfer phase, and learning and all 6 combinations of the transfer phase for model-space.

S1 Fig. An OL agent has higher performance when the distance between options values are higher.



The function of the performance changes with β as a variable. The conditions were covered 10 different pairs of options in which their relative values were covered $\{1, 2, ..., 10\}$ ($[\mu_1, \mu_2] \in \{[10, 9], [10, 8], ..., [10, 0]\}$, and $\delta = 1$). Performances were obtained with averaging across different α_1 and α_2.