Published in final edited form as:

Curr Biol. 2017 November 06; 27(21): 3375–3383.e3. doi:10.1016/j.cub.2017.09.051.

Caudate Microstimulation Increases Value of Specific Choices

Samantha R. Santacruz[1,2], **Erin L. Rich**[1,3], **Joni D. Wallis**[1,3], and **Jose M. Carmena**[1,2]

[1]Helen Wills Neuroscience Institute, University of California, Berkeley, CA, 94720 USA

[2]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, 94720 USA

[3]Department of Psychology, University of California, Berkeley, CA, 94720 USA

SUMMARY

Value-based decision-making involves an assessment of the value of items available and the actions required to obtain them. The basal ganglia are highly implicated in action selection and goal-directed behavior [1–4], and the striatum in particular plays a critical role in arbitrating between competing choices [5–9]. Previous work has demonstrated that neural activity in the caudate nucleus is modulated by task-relevant action values [6,8]. Nonetheless, how value is represented and maintained in the striatum remains unclear since decision-making in these tasks relied on spatially lateralized responses, confounding the ability to generalize to a more abstract choice task [6,8,9]. Here, we investigate striatal value representations by applying caudate electrical stimulation in macaque monkeys (n=3) to bias decision-making in a task that divorces the value of a stimulus from motor action. Electrical microstimulation is known to induce neural plasticity [10,11] and caudate microstimulation in primates has been shown to accelerate associative learning [12,13]. Our results indicate that stimulation paired with a particular stimulus increases selection of that stimulus, and this effect was stimulus-dependent and action-independent. The modulation of choice behavior using microstimulation was best modeled as resulting from changes in stimulus value. Caudate neural recordings (n=1) show that changes in value-coding neuron activity is stimulus value-dependent. We argue that caudate microstimulation can differentially increase stimulus values independent of action, and unilateral manipulations of value are sufficient to mediate choice behavior. These results support potential future applications of microstimulation to correct maladaptive plasticity underlying dysfunctional decision-making related to neuropsychiatric conditions.

RESULTS

Reward contingencies guide choices in a probabilistic reward choice task

To assess effects of microstimulation on the value of stimuli and actions, three rhesus macaques (Monkeys L, M, and P) were trained in a probabilistic reward choice task, in

Correspondence to Lead Contact: Jose M. Carmena (jcarmena@berkeley.edu).

Author Contributions

S.R.S. and J.M.C. conceived the study. S.R.S., E.L.R., J.D.W. and J.M.C. designed the experiments. S.R.S. performed the experiments and analysis. S.R.S., E.L.R., J.D.W. and J.M.C. wrote the manuscript.

which they learned to choose between colored targets associated with unique reward probabilities using different actions. In choice trials, the subject selects one of two presented colored targets (Figure 1A). In each session, new target colors were selected and arbitrarily assigned reward probabilities (STAR Methods). In each trial, targets were presented randomly on the left or right of the screen, requiring the subjects to execute different actions to make their choice. Therefore, subjects were required to learn the abstract association between the target color, not position, and its reward likelihood.

Subjects quickly learned the reward probability contingencies and selected the higher-value target color with greater frequency. No subject exclusively selected the higher-value target color after initial learning, but rather their decision policies tended to track local fluctuations in reward (Figure 1B). To verify that subjects indeed developed a decision-policy independent of spatial location, we calculated the likelihood of higher-value target color selection after initial learning in sham sessions conditioned on target presentation side (Figure 1C), and found there was no significant difference in position choice ($F_{1,18} = 2.818$, $p = 0.110$ for Monkey L; $F_{1,22} = 4.225$, $p = 0.052$ for Monkey M; $F_{1, 24} = 0.139$, $p = 0.712$ for Monkey P; one-way ANOVA). This indicates that all subjects successfully acquired a value representation associated with target color that was used to guide their decision-making policies independent of action.

High-frequency microstimulation biases target selection in free-choice trials

In each session, trials where organized in 3 blocks. The first block, Block A, consisted of 100–150 free-choice and instructed trials, randomly interleaved, over which reward contingencies were learned. In Block B, the subject completed 100 instructed trials where target colors were presented with equal probability. During microstimulation sessions, stimulation was applied during the center hold period of instructed trials to a particular target color in Block B. The third block, Block A’, was identical to Block A except that all instructed trials were to the target paired with stimulation and stimulation was delivered during these trials.

High-frequency microstimulation was delivered in the anterior caudate (Figure 1D) during the center hold period of instructed trials to the low-value (two-color task) or the medium-value target (three-color task) during Blocks B and A’. Stimulation was administered during instructed trials so that it was selectively associated with one of the two targets. Microstimulation had a significant impact on the monkeys’ decision-making policy ($F_{2,28} = 6.515$, $p = 0.005$ for Monkey L; $F_{2,33} = 7.811$, $p =0.002$ for Monkey M; $F_{2,37} = 4.784$, $p = 0.014$ for Monkey P; one-way ANOVA). All subjects showed an increased preference for the lower-value target when microstimulation was delivered relative to sham stimulation ($p < 0.01$ for Monkey L, $p < 0.05$ for Monkey M, $p < 0.05$ for Monkey P; post-hoc Tukey’s HSD test, Figure 2A).

To further characterize this effect and its time course, we examined whether this tendency was more pronounced on choice trials immediately following an instructed trial with stimulation. We found that all subjects maintained a significantly increased preference for the lower-value target for 5 or more consecutive free-choice trials following a trial with stimulation. For all monkeys, the main effect of stimulation was significant ($F > 4.9$, $p <$

0.01 in all three subjects), but latency and the interaction were not (two-way ANOVA, Figures 2D–F). Consistently we found that microstimulation paired with a lower-value target resulted in an increased preference which persisted for multiple successive choices. These data suggest that caudate microstimulation can impact ongoing decision-making processes in behaving animals.

Caudate microstimulation increases stimulus, not action, value

To verify that stimulation did not simply inject noise into decision-making processes, we analyzed control data from Monkeys L and M in which the stimulation was paired with the higher-value target. For Monkey L, this involved performing an additional set of experiments in which stimulation was paired with the high-value target (n = 11 sessions), while for Monkey M we considered the choice behavior on trials in which the low-value and medium-value target colors were presented together, making the medium-value the higher-value color in this context. Stimulation paired with the higher-value target significantly decreased the probability of selecting the lower-value target color relative to stimulation paired with the lower-value target color ($p < 0.05$; post-hoc Tukey's HSD test, Figure 2A). This demonstrates that caudate stimulation does not add noise to learned associations and supports the hypothesis that stimulation increases stimulus value in a target-specific manner.

We also looked at whether the action associated with choices following an instructed trial with stimulation in Block A' (Figure 2B) and whether receiving a reward changed the effects of stimulation on the subsequent choice (Figure 2C), and hence performed a two-way MANOVA analysis using target location and reward during the trial with stimulation as the independent variables, and target location and target color choice on the subsequent trial as the dependent variables. Stimulation was equally effective regardless of target location and did not preferentially bias the subjects toward selecting a target on a particular side (main effect of target side: $F_{2,36} = 0.571$, p 0.45 for Monkey L; $F_{2,43} = 0.278$, $p = 0.759$ for Monkey M; $F_{2,60} = 0.507$, $p = 0.48$ for Monkey P; two-way MANOVA). We also found no significant difference in the fraction of free-choice trials in which the lower-value target color was selected following a rewarded or unrewarded stimulation trial (main effect of reward: $F_{2,36} = 2.776$, $p = 0.1$ for Monkey L; $F_{2,43} = 1.040$, $p = 0.362$ for Monkey M; $F_{2,60} = 0.627$, $p = 0.43$ for Monkey P; two-way MANOVA, Figure 2C). This demonstrates that the change in stimulus value caused by microstimulation was not significantly modulated by reward and, therefore, we conjecture it is unlikely to operate by directly changing prediction error signaling.

Stimulation modulates value updates in a stimulus-specific manner

The mechanism by which stimulation introduces the observed bias in the decision-making process remains to be determined. To address this, we fit the subjects' behavior using a RL approach known as Q-learning [14]. The standard Q-learning algorithm consists of a value update function which includes a learning rate parameter, α, which dictates how much the subject "learns" from error. The values are used in a decision rule to determine the probability of selecting a given stimulus. This rule contains an inverse temperature parameter, β, which indicates how random choice behavior is (STAR Methods). Using this model and the ML parameter fits, we find that the learning rate is unchanged, but that the β

parameter changes across conditions (Figures 3C,D). A lower β value corresponds to greater exploratory behavior, which is how this model explains the increase in lower-value target selection when stimulation is paired with this target. Conversely, a higher β value corresponds to greater exploitation, which is how the model explains a decrease in lower-value target selection when stimulation is paired with the higher-value target.

Although this change in β does reflect the behavioral trends, it is counterintuitive that stimulation would induce opposite effects in the exploration-exploitation tradeoff simply by changing which stimulus it is coupled with. The regular Q-learning model is limited in how it can explain the effects of stimulation since it does not explicitly account for what is essentially a new input in the decision making process. Thus, we propose four candidate models that explicitly include an additional parameter, λ, to explore different ways in which stimulation could modify the stimulus value updates or decision rule (STAR Methods; Figure 3A). We computed the Bayesian Information Criterion (BIC) for each model [15,16]. The BIC includes a penalty for models with more parameters, thus ensuring that our proposed models are not quantified to be better by simply being more complex. We found that the multiplicative Q parameter candidate model is the best fit (Figure 3B, F). The average BIC values were found to be significantly different ($F_{4,55} = 2.736$ for Monkey L, $F_{4,55} = 2.61$ for Monkey M, $F_{4,85} = 2.520$ for Monkey P; one-way ANOVA), with the average BIC value for the model with a multiplicative Q parameter significantly smaller than for all other candidate models ($p < 0.05$; post-hoc Tukey’s HSD).

We find the ML parameter values for the best-fitting model which includes a multiplicative parameter in the value update equation. The learning rate, α, and the multiplicative gain parameter, λ, differ depending on whether stimulation is paired with the lower- or higher-value target, whereas the β remains fixed (Figures 3C–E). These changes suggest that changes in behavior are reflected in changes the value update process rather than directly through the decision policy, which is consistent with the stimulus-specific behavioral effects observed.The difference the multiplicative gain parameter, λ, likely reflects the fact that there is a different amount of modulation required on the perceived value of the lower-value and higher-value target color options in order to sway decision-making.

Caudate neurons encoding stimulus value are modulated by microstimulation

We recorded neural activity in the caudate nucleus of Monkey M during the choice task and isolated individual phasically active neurons (PANs; n = 131 for stimulation sessions, n = 135 for sham sessions; Figure 4A). To determine the neural representation of stimulus values in this population, we performed multiple linear regression of neural firing rates around the colored targets’ presentation time with stimulus values derived from the Q-learning model as regressors. Additionally, we included choice (i.e. chosen-value) and motor covariates, namely reaction time and movement time, in the regression. Roughly the same percentage of neurons exclusively co-varied with the stimulus value of only one of the stimulus colors, though many neurons co-varied with the values of multiple stimuli (Figure 4B).

We hypothesized that value-coding during the deliberation hold time was modulated by stimulation. We found that spiking activity of value-coding neurons changed following stimulation and that these changes related to stimulus value (Figure 4C). Firing rates in the

400 ms following target presentation typically were increased in neurons that co-varied with Q_{med}, as shown for a representative neuron in Figure 4D ($F_{1,232} = 4.899$, $p = 0.028$, ANCOVA with stimulus value as a covariate), whereas there was little to no change for neurons that did not encode value, as shown for non-value coding neuron in Figure 4E ($F_{1,174} = 0.281$, $p = 0.597$, ANCOVA). The amount of change tended to be stimulus value-dependent and this was true of peak firing rates as well (Figure 4F). To compare firing rate changes between stimulation and sham conditions, we pooled all Q_{med}-coding neurons from these sessions and examined the difference in peak firing rates after stimulation was administered (Block A') relative to before stimulation was administered (late trials in Block A). For larger stimulus values, we found significant increasing differences in peak firing rate during stimulation sessions (Figure 4G; main effects of stimulation condition and stimulus value: $p < 0.05$; interaction effect: $F_{5,69} = 4.98$, $p < 0.01$; two-way ANOVA). This stimulus value-dependent effect supports the best fitting Q-learning model that indicates that stimulation modulates stimulus value.

DISCUSSION

Here we have shown that caudate microstimulation can alter value-based choice behavior and that this causal bias results from a selective change in the subjective value of a target stimulus. The effects cannot be explained by changes in action values, stimulus-action associations, nor stimulus-reward associations. Unilateral striatal manipulations can impact lateralized motor movements [17,18] and striatal medium spiny neurons encode actions for such movements [19–21]. In this work, we developed a choice task where optimal decisions were not spatially lateralized and asked how microstimulation would alter behavior. This approach allowed us to precisely decouple whether stimulation drove an action-specific or stimulus-specific behavioral responses. Indeed, we found that the latter was the case. For all three subjects we found that target location during stimulation did not significantly impact subsequent stimulus or action choice.

In previous studies, the action of choosing a target on the left or right side of the screen was associated with reward, in contrast to our task design. As such, we hypothesize that stimulus values may also be represented during the pre-response hold period if they are used to guide the upcoming action and thus that stimulation delivered coincident with this activity could modulate striatal encoding of stimulus value. Previous work conjectured that high-frequency stimulation of the caudate might lead to long-term potentiation (LTP) in particular corticostriatal synapses [13]. In striatal slice preparations it has been demonstrated that LTP is governed in part by dopaminergic activity [22,23]. Given that high-frequency stimulation can enhance dopamine release [24,25], caudate stimulation may result in LTP of particular corticostriatal synapses, positively reinforcing stimulus-specific activity.

The results also demonstrate that the change in behavior due to microstimulation was independent of reward. However, there is a question of whether microstimulation itself is perceived as rewarding. Stimulation of several brain regions, including the orbitofrontal cortex, amygdala, nucleus accumbens, ventral tegmental area, and lateral hypothalamus, can be perceived as rewarding and reinforce behavior [26–29]. However, in a similar choice task as used here there was no bias for a particular choice for trials in which reward alone versus reward with caudate microstimulation was delivered, suggesting that caudate stimulation itself does not increase saliency [13]. Since mediation of choice behavior through caudate stimulation was not significantly modulated by reward, it follows that caudate stimulation is neither perceived as rewarding nor causes biased decision making by directly modulating response to reward.

The use of an RL framework enabled us to differentiate between different sources of value and decision biases that could account for choice behavior changes. Animals respond to changing reward contingencies by altering their behavior and repeating actions with rewarding outcomes [14,30,31]. We found that stimulation preferentially biased the subjects towards the target it was paired with in the absence of any reward contingency change. The model that best explained this bias indicated that stimulation changed stimulus value. Our behavioral evidence indicates that these effects persisted for multiple trials, supporting the notion that it is a modification of an underlying representation, rather than a transient biasing effect. Indeed, the neural data shows that there are changes in the neural activity of value-coding neurons and that these changes are stimulus value-dependent.

Overall, we have demonstrated that high-frequency stimulation delivered to the caudate can modulate decision-making processes, and we speculate that the mechanism by which this is accomplished involves changes in striatal value representations of stimuli. An inability to appropriately evaluate stimuli and use these values to inform decisions lies at the core of neuropsychiatric disorders like anxiety, depression and addiction. Our results suggest that electrical stimulation may offer a novel therapeutic approach to help regulate these valuations of actions in patients with impaired decision-making abilities. Hence, our results not only demonstrate how electrical stimulation can change choice behavior but are also suggestive of future therapies for neuropsychiatric diseases.

STAR Methods

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Jose M. Carmena (jcarmenaberkeley.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Three male rhesus macaques were used in these experiments. All experiments were performed in compliance with the regulations of the Animal Care and Use Committee at the University of California, Berkeley. The three subjects weighed roughly 11.5 kg, 9.1 kg, and 10.2 kg at the time of the study, and were approximately 7 – 9 years of age at the time of the study. All subjects were healthy and housed in pairs.

METHOD DETAILS

Surgery—Three rhesus macaques were implanted with recording chambers. For Monkeys L and P, we used standard methods for acute neurophysiology that have been described in detail elsewhere (Lara, Kennerley, & Wallis, 2009). For Monkey M, we used a custom semichronic microdrive array to record from moveable single microelectrodes that were chronically implanted (Gray Matter Research, Bozeman, MT). Chamber positions were calculated based on images obtained from 1.5-T magnetic resonance imaging (MRI) scans of each subject's brain. Monkey P was implanted bilaterally with cylindrical chambers centered above lateral orbital regions, allowing access to the anterior caudate. Monkeys L and M were implanted unilaterally with a custom-machined chamber enabling access to anterior and posterior caudate.

Stimulation—Microstimulation pulse trains consisted of biphasic pulses with no inter-pulse interval and a cathodal leading phase. Each phase was 200 us in duration and the pulse frequency was 200 Hz. Stimulation was either constant-current, with a current amplitude in the range of 10 – 50 uA, or constant-voltage, with an amplitude of 1 V. These stimulation parameters are consistent with previous studies using electrical stimulation in non-human primates (Afraz et al., 2006; Amemori & Graybiel, 2012; Ditterich et al., 2003; Hanks et al., 2006; Nakamura & Hikosaka, 2006; Williams & Eskandar, 2006). Stimulation trains lasted 1000 ms and coincided with the center-hold period of the task. This design ensured that the subject had negligible movement during the stimulation epoch. Two Platinum-Iridium microelectrodes (Alpha Omega; Microprobes) with an impedance in the range of 100 – 300 kOhm were used to administer stimulation in a bipolar manner. For Monkeys L and P, on each experimental day electrodes were lowered manually using custom-built microdrives to a target depth in the head of the caudate. For Monkey M, electrodes were positioned in the caudate chronically.

Behavioral Task—Three macaque monkeys were trained in a probabilistic reward free-choice joystick task. Briefly, the subjects were trained to use a joystick to control a cursor on a computer screen and select colored circular targets that each had a probability of reward associated with the color of the target. The joystick was affixed to the front of the primate chair and subjects were free to use either hand at any point in the task to control the joystick. This task consisted of two types of trials: (1) free-choice trials and (2) instructed trials. In free-choice trials, the subject was trained to hold the cursor at the center target for 1000 ms. During this center-hold period, two peripheral circular targets of different colors were simultaneously shown on the screen. At the end of the hold, cued by the removal of the center target from the screen, the subject freely moved the cursor to one of the two peripheral targets to select it by holding the cursor inside the target for 1000 ms. During instructed trials, only one peripheral target was presented. In a given session, two or three colors from a set of 12 colors were chosen and arbitrarily assigned values. Two subjects (Monkeys L and P) were trained in a task with two targets and one subject (Monkey M) was trained in a task with three target colors. If two colors were used, then each was assigned to be either the low-value (40% reward probability) or high-value (80% reward probability) target color. With three colors, each was assigned to be the low-value (35% reward probability), medium-value (60% reward probability), or high-value (85% reward probability) target color. During training each subject was tested with different color-pairings in the probabilistic reward choice task and the same task with a reward-contingency reversal, to ensure that the colors were distinguishable to the subject. Reward schedules for each target were pseudo-random across blocks of 100 trials, meaning that the reward assignments were random but a fixed number of rewards were allocated over the block. For example, an 80% reward probability corresponded to 80 trial indices in a block of 100 trials being uniformly selected without replacement to have a reward associated with the respective target in that trial. This pseudo-random reward schedule ensured that the empirical reward likelihood for small numbers of trials was close to the true reward probability. On trials with microstimulation administered, the reward schedule remained unchanged for the associated target so that the subject's experience of reward for the target was no different than on trials without intervention. When administered, stimulation was delivered coincident with the center-hold period and lasted the duration of the hold. A trial was considered to be successful if the subject completed the 1000 ms center-hold followed by holding at a peripheral target for 1000 ms within a 10 s period. If a reward was scheduled to be allocated with the selected peripheral target, a custom-programmed Arduino triggered a solenoid reward system to deliver a small amount of juice to the subject. The same trial was repeated up to 10 times until it was successfully completed and the subject advanced to the next trial.

Behavioral Manipulations—Since we aimed to preferentially increase the value of one of two alternative targets, stimulation was only delivered during instructed trials to that target. After an initial learning block (Block A) of 100 or 150 trials, depending on whether two or three colors were used, consisting of 70% free-choice trials with 30% instructed trials randomly-interleaved, we exposed the subject to a priming block (Block B) of all instructed trials. In Block B, the subject was equally likely to be instructed to any target color. On trials where the instruction was to the low-value target in the two-color task or the medium-value target in the three-color task, stimulation was administered during the center-hold period. The causal effects of stimulation were then assayed in a third block, Block A', of which 70% of trials were free-choice and 30% were instructed to the target paired with stimulation. Microstimulation was also paired with the hold period of the instructed trials in Block A'. If stimulation had no effect on the value of the target, we would expect that the forced-exploration of both targets during Block B would result in a similar or stronger preference for the high-value target during Block A'. If stimulation did change the value of the target, we would expect to see increased preference for the low-value target and this is indeed what we find. Control experiments with one subject were also performed in which microstimulation was paired with the high-value target instead of the low-value target and the opposite behavioral bias was found. In sham sessions, an identical block structure was utilized but stimulation was not administered during Blocks B and A'. Monkey L performed 11 sham sessions and 12 stimulation sessions, Monkey M performed 12 sham sessions and 12 stimulation sessions, and Monkey P performed 14 sham sessions and 18 stimulation sessions.

Behavior Data Models—Analyses were performed in Python with custom-written routines utilizing publicly available Python libraries. Only sessions in which the subject learned well during Block A and completed at least 100 trials in Block A' were included in the analysis. Monkey L completed 12 sessions with constant-current stimulation, 11 control sessions (constant-current stimulation associated with the high-value target), and 11 sham sessions. Monkey M performed 12 sham sessions and 12 stimulation sessions. Monkey P completed 18 sessions with constant-current stimulation, 11 sessions with constant-voltage stimulation, and 14 sham sessions. Q-learning, a model-free RL algorithm, was used to fit the subjects' free-choice behavior and modified to included explicit parameters modeling the effects of stimulation on decision-making. The learning rate, α, determines how much the value of a choice is updated by new information in each time step, and the inverse temperature, β, indicates how much expected rewards affect the probability of selecting a stimulus. The standard Q-learning algorithm (Sutton & Barto, 1998) consists of the following value update equations:

$$Q(t)=Q(t-1)+\alpha\delta(t), \quad [1]$$

$$\delta(t)=r(t)-Q(t-1). \quad [2]$$

Using a soft-max decision rule, the probability of selecting the lower-value (a_{LV}) target over the higher-value (a_{HV}) target is:

$$P(a_{\mathrm{LV}}(t)|Q_{\mathrm{LV}}(t),Q_{\mathrm{HV}}(t))=\frac{1}{1+\exp(\beta[Q_{\mathrm{HV}}(t)-Q_{\mathrm{LV}}(t)])}. \quad [3]$$

The variables $Q_{LV}(t)$ and $Q_{HV}(t)$ represent the values of the lower-value and higher-value targets at time t, respectively. The parameter α is the learning rate and the parameter β is known as the inverse temperature. When compared with alternative approaches, this set of equations will be referred to as the "regular" model. Maximum likelihood (ML) estimates for the model parameters were found for sham, stimulation, and control sessions. Parameters were found for Block A' using the ML parameters fit separately from behavior in Block A as initial estimates. For sham sessions, behavior was modeled with the standard Q-learning algorithm with average accuracy of 85.7 ± 0.6%, 84.2% ± 0.6%, and 88.4 ± 0.7% for Monkeys L, M, and P, respectively.

We tested four modified versions of this framework that explicitly include a parameter for stimulation. The first two approaches included modifications to the value update equations in either an (1) additive or (2) multiplicative manner. A parameter, λ, in both cases captures the magnitude of the stimulation effect on value and the term $S(t)$ is a binary indicator of stimulation being administered on trial t. In the first case, $Q(t-1) \rightarrow Q(t-1) + \lambda S(t)$ is used to replace the term for the value at time $t-1$. In the second case, $Q(t-1) \rightarrow Q(t-1) + \lambda Q(t-1)S(t)$ is used to replace the previous value term. This updated equation equals $Q(t-1)$ when there is no stimulation ($S(t) = 0$) and $(1 + \lambda)Q(t-1)$ when there is stimulation ($S(t) = 1$).

The second set of approaches considered alternatively supposed that the decision rule was directly affected by stimulation. Again, we consider both additive and multiplicative effects of stimulation. An additive parameter in the decision rule was modeled as

$$P(a_{\mathrm{LV}}(t)|Q_{\mathrm{LV}}(t),Q_{\mathrm{HV}}(t))=\frac{1}{1+\exp(\beta[Q_{\mathrm{HV}}(t)-Q_{\mathrm{LV}}(t)]+\lambda S(t))}. \quad [4]$$

Alternatively, a multiplicative scaling factor in the decision rule induced by stimulation was modeled as

$$P(a_{\mathrm{LV}}(t)|Q_{\mathrm{LV}}(t),Q_{\mathrm{HV}}(t))=\frac{1-(1-\lambda)\mathrm{S(t)}}{1+\exp(\beta[Q_{\mathrm{HV}}(t)-Q_{\mathrm{LV}}(t)])}. \quad [5]$$

For the standard and modified Q-learning strategies, all parameters used to model the subjects' behavior were maximum likelihood estimates. The value update equation was updated on both instructed and free choice trials, but the decision rule was only simulated for free-choice trials. The ability of each of these models to fit the behavioral data was assessed using the Bayesian Information Criterion (BIC), which is defined as (Neath & Cavanaugh, 2012; Schwarz, 1978)

$$\mathrm{BIC}=-2\ \ln\ \hat{L}+n\ \ln\ T,$$

where $\hat{L}$ is the maximum of the likelihood function, n is the number of parameters in the model, and T is the number of trials.

Neural Data Analysis—In analyzing the neural activity, our goal was to determine which neurons co-varied with stimulus value and how the activity of these neurons changed with stimulation. We used multiple linear regression to determine how the activity of all well-isolated units co-varied with value, as well as movement variables and choice. The responses of individual neurons were fit using the following multiple linear regression,

$$\mathbf{y}=\beta_1\mathbf{C}+\beta_2\mathbf{MT}+\beta_3\mathbf{RT}+\beta_4\mathbf{Q_{low}}+\beta_5\mathbf{Q_{med}}+\beta_6\mathbf{Q_{high}},$$

where **y** is the firing rate in the window [0,400) ms following target presentation, **C** is the chosen target color, **MT** is the movement time, and **RT** is the reaction time. The variables $\mathbf{Q_{low}}$, $\mathbf{Q_{med}}$, and $\mathbf{Q_{high}}$ represent the dynamic stimulus values estimates for the low-value, medium-value, and high-value target colors as determined by the best-fitting Q-learning model. Statistical significance of regressors was determined using incremental F-statistic with a significance level of 0.05. Neurons were classified as "value" neurons if their activity co-varied with any Q value, regardless of whether they also co-varied with other regressors. Choice, movement time, and reaction time neurons were categorized as exclusively co-varying with the associated regressor. Neurons that did not significantly co-varying with any of these regressors were labeled as non-responsive.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analyses were parametric and were performed using IBM SPSS, as well as custom-written Python routines. Python code utilized the numpy, scipy, statsmodels and sklearn libraries which are publically available. For behavioral analyses, one-way ANOVA, two-way ANOVA, and two-way MANOVA statistical tests were performed. For these analyses, the sample sizes correspond to the number of behavioral sessions. For neural analyses, an ANCOVA and two-way ANOVA were performed. In this case, the sample sizes correspond to the number of trials considered in the analysis. For all analysis of variance statistics, the F-test value is reported with the first subscript indicating the between groups degrees of freedom and the second subscript indicating the within groups degrees of freedom. Tukey's HSD was used for post-hoc statistical analysis only when F-tests were associated with significant p-values ($p < 0.05$).

Acknowledgments

This work was supported by NIDA R21 DA041791, NIMH R01 MH097990 to J.D.W., and the Defense Advanced Research Projects Agency (W911NF-14- 2-0043) to J.D.W. and J.M.C.

References

1. Gremel CM, Costa RM. Orbitofrontal and striatal circuits dynamically encode the shift between goal-directed and habitual actions. Nat. Commun. 2013; 4:2264. [PubMed: 23921250]
2. Hikosaka O, Nakamura K, Nakahara H. Basal Ganglia Orient Eyes to Reward. J Neurophysiol. 2006; 95:567–584. [PubMed: 16424448]
3. Redgrave P, Rodriguez M, Smith Y, Rodriguez-Oroz MC, Lehericy S, Bergman H, Agid Y, DeLong MR, Obeso JA. Directed and habitual control in the basal ganglia : implications for Parkinson ' s disease. Nat. Rev. Neurosci. 2010; 11:760–772. [PubMed: 20944662]
4. Yin HH, Knowlton BJ. The role of the basal ganglia in habit formation. Nat. Rev. Neurosci. 2006; 7:464–476. [PubMed: 16715055]
5. Kravitz AV, Tye LD, Kreitzer AC. Distinct roles for direct and indirect pathway striatal neurons in reinforcement. Nat. Neurosci. 2012; 15:816–818. [PubMed: 22544310]
6. Lau B, Glimcher PW. Value representations in the primate striatum during matching behavior. Neuron. 2008; 58:451–63. [PubMed: 18466754]
7. 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. J. Neurosci. 2014; 34:9196–9201. [PubMed: 25009253]
8. Samejima K, Ueda Y, Doya K, Kimura M. Representation of action-specific reward values in the striatum. Science. 2005; 310:1337–40. [PubMed: 16311337]
9. Tai L-H, Lee AM, Benavidez N, Bonci A, Wilbrecht L. Transient stimulation of distinct subpopulations of striatal neurons mimics changes in action value. Nat. Neurosci. 2012; 15:1281–9. [PubMed: 22902719]
10. Jackson A, Mavoori J, Fetz EE. Long-term motor cortex plasticity induced by an electronic neural implant. Nature. 2006; 444:56–60. [PubMed: 17057705]
11. Madhavan R, Chao ZC, Potter SM. Plasticity of recurring spatiotemporal activity patterns in cortical networks. Phys. Biol. 2007; 4:181–193. [PubMed: 17928657]
12. Nakamura K, Hikosaka O. Facilitation of saccadic eye movements by postsaccadic electrical stimulation in the primate caudate. J. Neurosci. 2006; 26:12885–12895. [PubMed: 17167079]
13. Williams ZM, Eskandar EN. Selective enhancement of associative learning by microstimulation of the anterior caudate. Nat. Neurosci. 2006; 9:562–8. [PubMed: 16501567]
14. Sutton, RS., Barto, AG. Reinforcement Learning. Cambridge, MA: MIT Press; 1998.

15. Neath AA, Cavanaugh JE. The Bayesian information criterion : background, derivation, and applications. WIREs Comput. Stat. 2012; 4:199–203.

16. Schwarz G. Estimating the dimension of a model. Ann. Stat. 1978; 6:461–464.

17. Kravitz AV, Freeze BS, Parker PR, Kay K, Thwin MT, Deisseroth K, Kreitzer AC. Regulation of Parkinsonian motor behaviors by optogenetic control of basal ganglia circuitry. Nat. Lett. 2010; 466:622–626.

18. Schwarting RKW, Huston JP. The unilateral 6-hydroxydopamine lesion model in behavioral brain research. Analysis of functional deficits, recovery and treatments. Prog. Neurobiol. 1996; 50:275–331. [PubMed: 8971983]

19. Kim H, Sul JH, Huh N, Lee D, Jung MW. Role of striatum in updating values of chosen actions. J. Neurosci. 2009; 29:14701–14712. [PubMed: 19940165]

20. Kubota Y, Liu J, Hu D, DeCoteau WE, Eden UT, Smith AC, Graybiel AM. Stable encoding of task structure coexists with flexible coding of task events in sensorimotor striatum. J. Neurophysiol. 2009; 102:2142–2160. [PubMed: 19625536]

21. Thorn CA, Atallah H, Howe M, Graybiel AM. Differential Dynamics of Activity Changes in Dorsolateral and Dorsomedial Striatal Loops during Learning. Neuron. 2010; 66:781–795. [PubMed: 20547134]

22. Hernandez-Lopez S, Bargas J, Surmeier DJ, Reyes A, Galarraga E. D1 Receptor Activation Enhances Evoked Discharge in Neostriatal Medium Spiny Neurons by Modulating an L-Type Ca2+ Conductance. J. Neurosci. 1997; 17:3334–3342. [PubMed: 9096166]

23. Wickens JR, Reynolds JNJ, Hyland BI. Neural mechanisms of reward-related motor learning. Curr. Opin. Neurobiol. 2003; 13:685–690. [PubMed: 14662369]

24. Borland LM, Shi G, Yang H, Michael AC. Voltammetric study of extracellular dopamine near microdialysis probes acutely implanted in the striatum of the anesthetized rat. J. Neurosci. Methods. 2005; 146:149–158. [PubMed: 15975664]

25. Cragg SJ, Hille CJ, Greenfield SA. Dopamine Release and Uptake Dynamics within Nonhuman Primate Striatum In Vitro. J. Neurosci. 2000; 20:8209–8217. [PubMed: 11050144]

26. Mora F, Avrith DB, Phillips AG, Rolls ET. Effects of satiety on self-stimulation of the orbitofrontal cortex in the rhesus monkey. Neurosci. Lett. 1979; 13:141–145. [PubMed: 119182]

27. Rolls ET, Burton MJ, Mora F. Neurophysiological analysis of brain-stimulation reward in the monkey. Brain Res. 1980; 194:339–357. [PubMed: 6770964]

28. Bichot NP, Heard MT, Desimone R. Stimulation of the nucleus accumbens as behavioral reward in awake behaving monkeys. J. Neurosci. Methods. 2011; 199:265–272. [PubMed: 21704383]

29. Arsenault JT, Rima S, Stemmann H, Vanduffel W. Role of the primate ventral tegmental area in reinforcement and motivation. Curr. Biol. 2014; 24:1347–1353. [PubMed: 24881876]

30. Dayan P, Balleine BW. Reward, Motivation, and Reinforcement Learning. Neuron. 2002; 36:285–298. [PubMed: 12383782]

31. Wise RA. Dopamine, learning and motivation. Nat. Rev. Neurosci. 2004; 5:1–12.

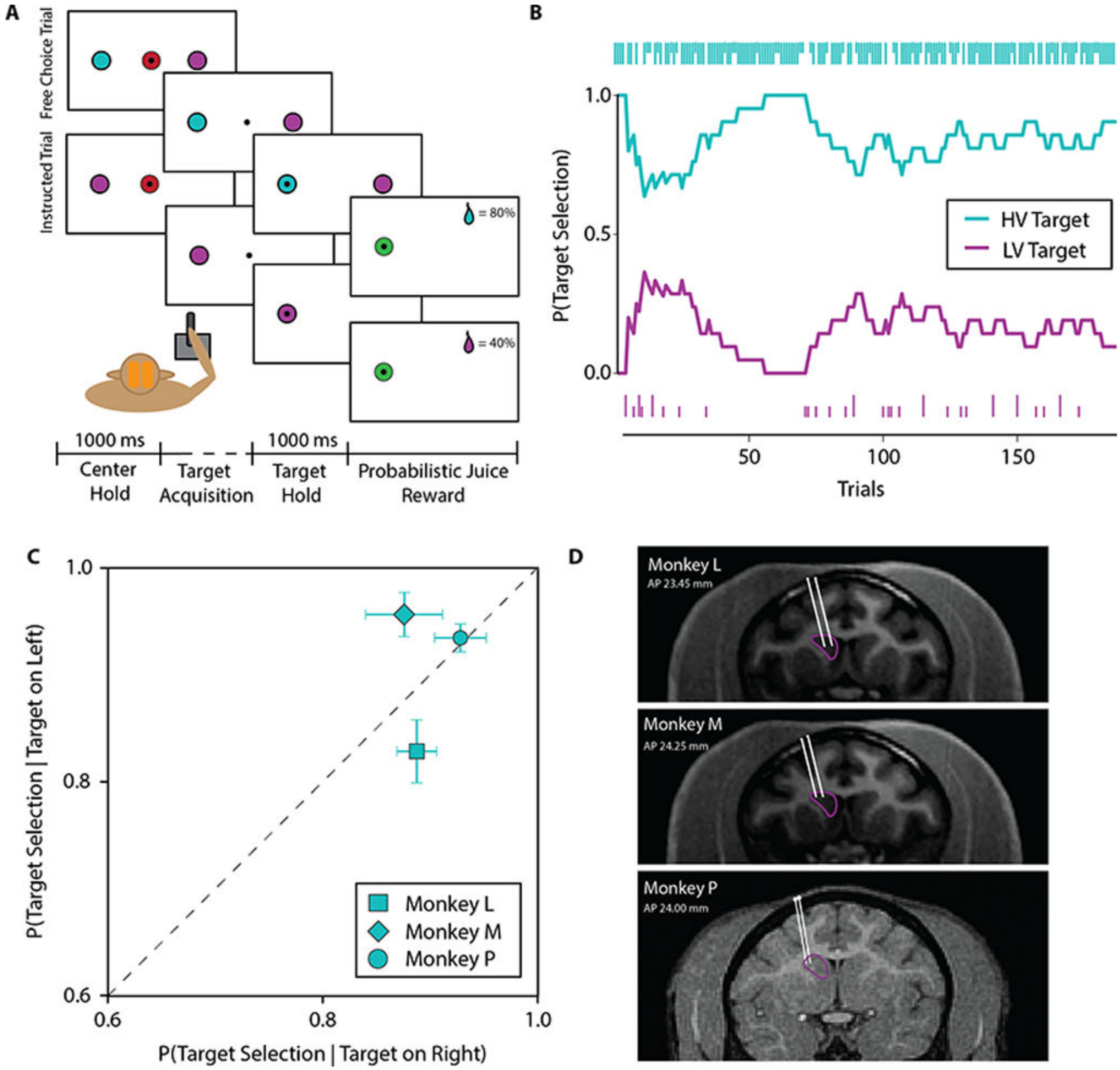

Figure 1. Experimental setup and behavioral task

(A) Cartoon depicting the two different trial types encountered by the subject in the probabilistic reward choice task. Note that the target colors randomly alternate sides of presentation, so that the subjects must learn to associate color, not spatial location, with reward probability. **(B)** Representative choice behavior during free-choice trials. The main plot shows the empirical probability of selecting each target over a sliding window of 20 trials. The small bars on the top and bottom portions of the screen indicate whether a reward was give or not when each target was chosen. Short bars indicate the absence of reward and long bars indicate presence of reward. **(C)** Conditional probabilities of selecting the higher-value (HV) target given that it was presented on either the left or right side of the screen. Selection agnostic to spatial location would lie on the identity line, shown as a dashed line in the plot. Results shown are from sham sessions for all subjects. **(D)** Microelectrode positions superimposed on MR images for each subject. The caudate is outlined in magenta and microelectrode trajectories are marked in white.

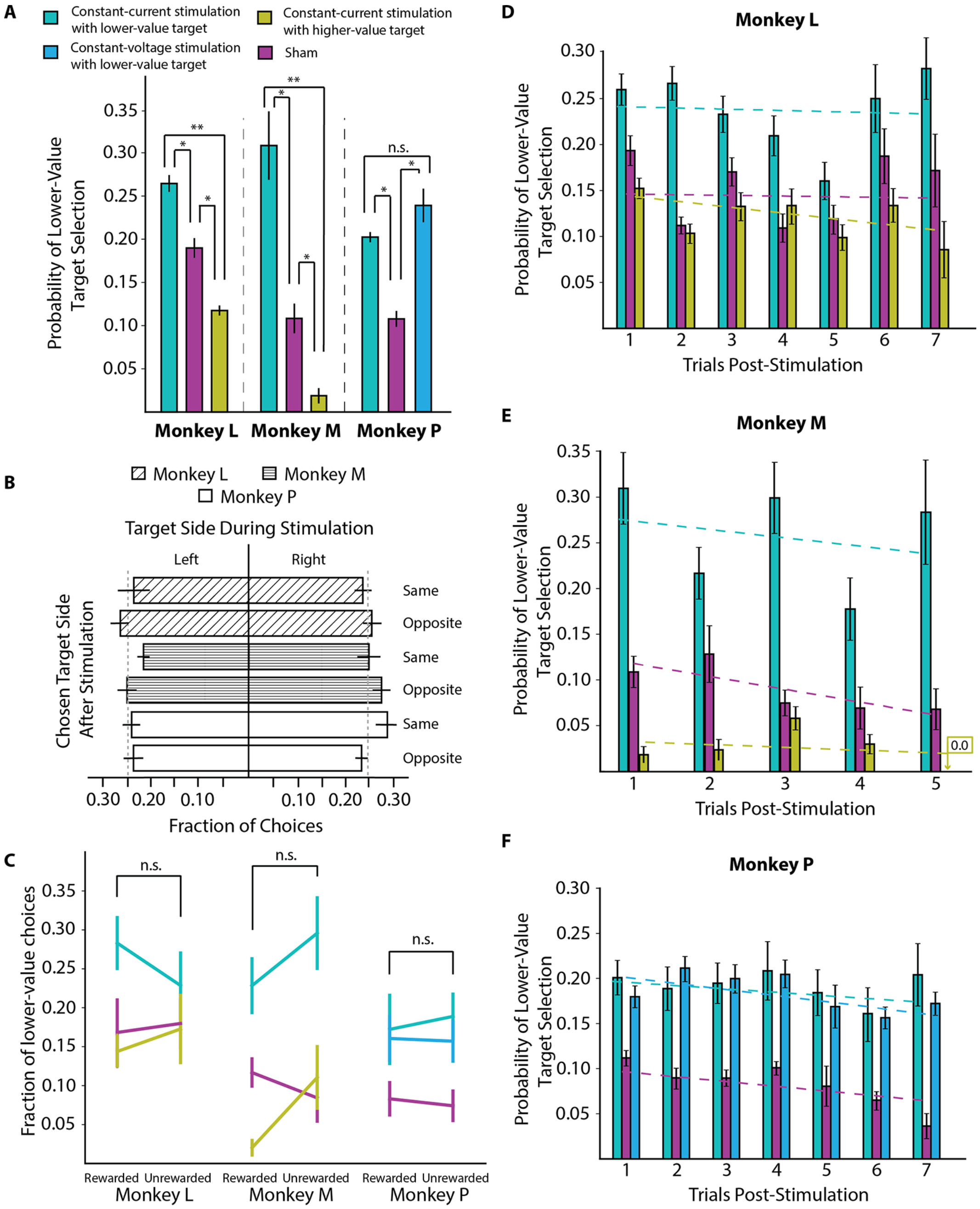

Figure 2. Microstimulation results

(A) The probability of selecting the lower-value target on free-choice trials. **(B)** Fraction of times of target presentation on a given side during the instructed trial with stimulation and the selection of a target on the same or opposite side in the subsequent free-choice trial. **(C)** Fraction of lower-value target choices on free-choice trial following a stimulation trial that was either rewarded or unrewarded. **(D)** The probability of selecting the lower-value target on free-choice trials aligned to their latency following the forced-choice trial with stimulation for Monkey L (interaction effect: $F_{8,135} = 0.507$, p = 0.849, main effect of stimulation condition: $F_{2,135} = 4.959$, p = 0.008, main effect of trial latency from stimulation: $F_{4,135} = 1.535$, $p = 0.196$; two-way ANOVA), **(E)** for Monkey M (interaction effect: $F_{8,161} = 0.470$, $p = 0.876$, main effect of stimulation condition: $F_{2,161} = 21.798$, $p < 0.001$, main effect of trial latency from stimulation: $F_{4,161} = 0.453$, $p = 0.770$), and **(F)** for Monkey P (interaction effect: $F_{8,185} = 0.281$, $p = 0.972$, main effect of stimulation condition: $F_{2,185} = 12.125$, $p < 0.001$, main effect of trial latency: $F_{4,185} = 0.490$, $p = 0.743$; two-way ANOVA). Significant differences are indicated as: n.s. (not significant), ** ($p < 0.01$), and * ($p < 0.05$).

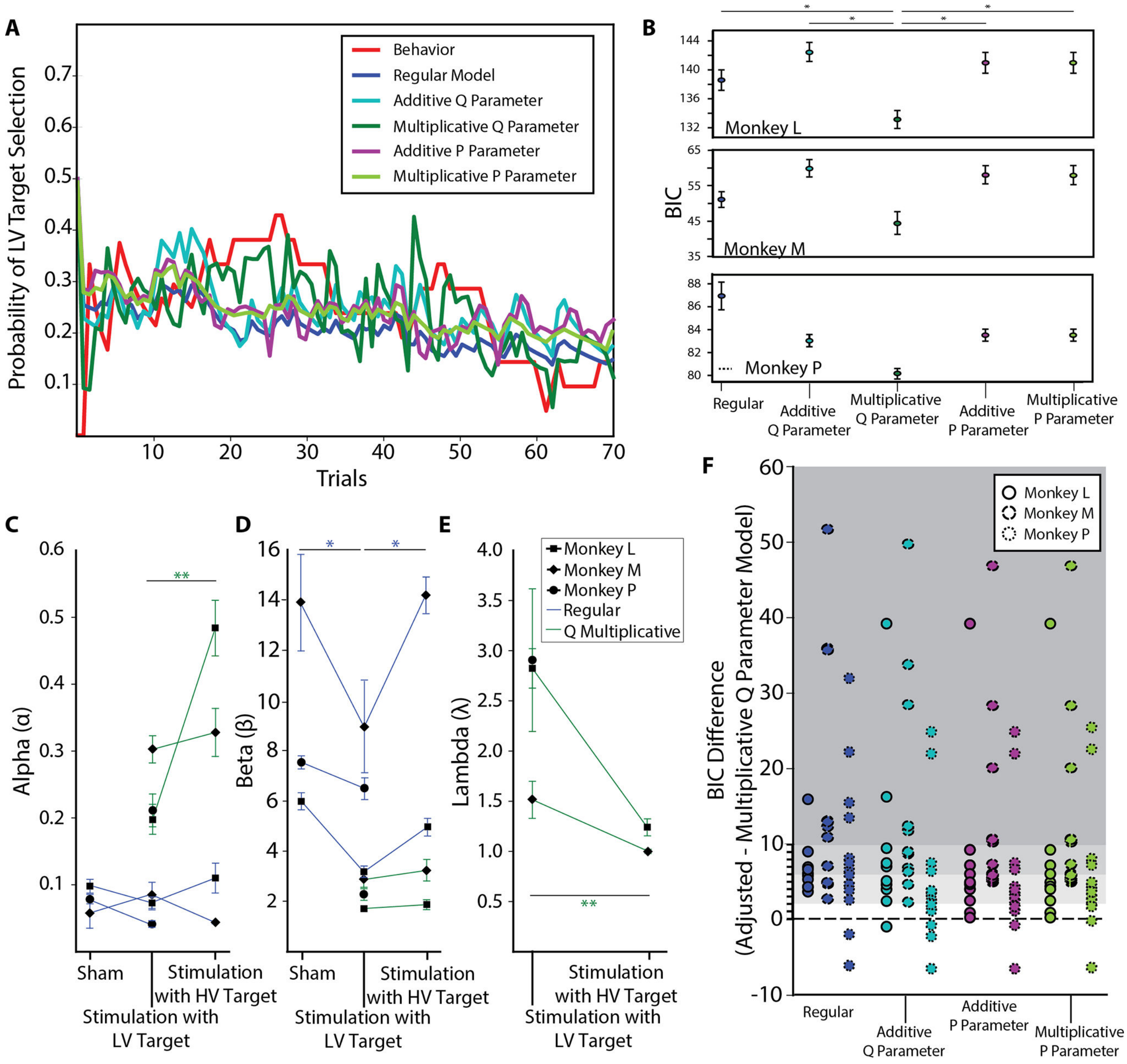

Figure 3. Computational model fitting

(A) Representative model fits for the regular and adjusted Q-learning algorithms with a soft-max decision rule. Results are plotted along with the raw behavior of the subject which is averaged over a sliding window of 20 trials. **(B)** Session-averaged BIC values for the regular and adjusted Q-learning candidate models. **(C)–(E)** Average Q-learning parameters averaged across sessions. The inverse temperature, *β*, was significantly different across conditions for the regular Q-learning model (main effect of stimulation condition: $F_{2,27} = 6.247$, $p < 0.01$ for Monkey L; $F_{2,21} = 9.563$, $p< 0.01$ for Monkey M; $F_{2,30} = 7.379$, $p < 0.01$ for Monkey P; one-way MANOVA). **(F)** The difference between BIC values per session for the various adjusted models (including the regular unadjusted model) and the model with the value update equation modified to include a multiplicative parameter capturing the effect of stimulation. Gray shadings indicate preference for the multiplicative Q parameter modification, with a BIC difference in the range 2 – 6 indicating a positive preference, 6 – 10 indicating a strong preference, and > 10 indicating a very strong preference. Significant differences are indicated as: n.s. (not significant), ** ($p < 0.01$), and * ($p < 0.05$) using post-hoc Tukey's HSD.

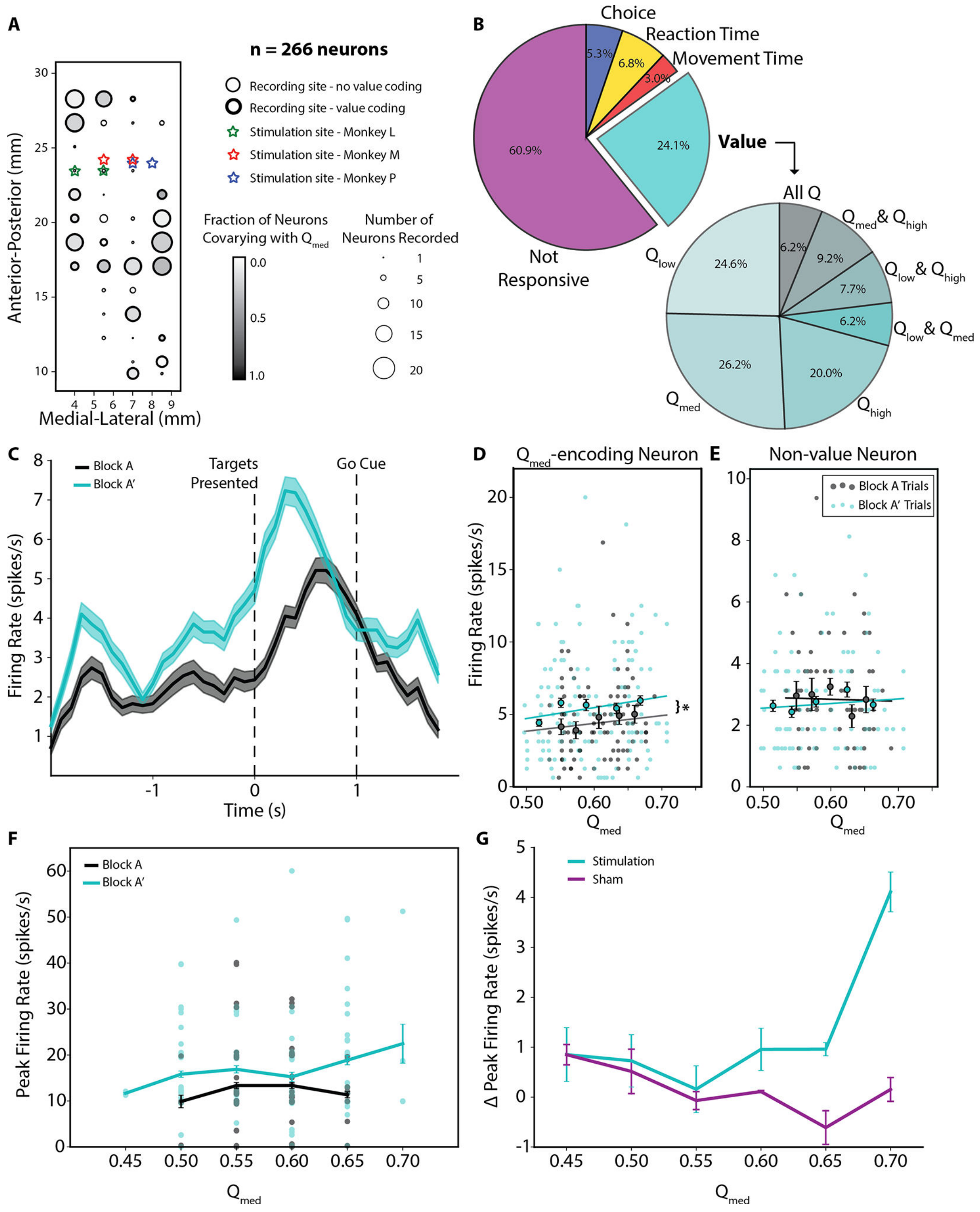

Figure 4. Neural correlates of value changes

(A) Recording locations for Monkey M and stimulation locations for all subjects. A totally of 266 task-related caudate neurons were recorded. Medial-lateral coordinate values are presented from midline, while Anterior-Posterior coordinates are presented relative to the interaural line. The marker size indicates the number of neurons sampled per site, while the marker outline indicates if any neuron recorded at the location significantly co-varied with stimulus value. The shading of the marker indicates the proportion of neurons that particularly co-varied with the value Q_{med}. **(B)** Pie chart on left shows the neurons (n = 266) categorized into five main types based on the linear regression analysis. The pie chart on the right expands upon the number of Value neurons to demonstrate the frequency in which neurons were responsive to three different values, $Q_{low,}$ Q_{med}, and Q_{high}, and combinations thereof. **(C)** Average firing rate of a representative Q_{med}-value coding caudate neuron during Blocks A and A'. Activity is taken only from trials in which Q_{med} was associated with the lower-value target, i.e. when the medium-value and high-value targets were presented together. Only the last 100 trials in Block A, after initial learning, are considered so that firing rate changes are not dominated by effects of learning. **(D)** The firing rate during picture onset as a function of the modeled value Q_{med} is shown for the same representative neuron. Each marker represents the per trial firing rate in the window [0,400) ms from when the targets are presented, while the lines represent the linear fit of firing rate as a function of value given by the linear regression. The slopes of the linear regression fits were not significantly different ($m_{Block\ A} = 6.279$, $m_{Block\ A'} = 8.573$, t-value = 1.263, p = 0.21), but there was a significant difference of 0.962 in the y-intercept (t-value = 2.213, p = 0.028). This suggests that there is a significant increase in firing rate during Block A' for all Q-values. Circles with error bars represent the trial-averaged firing rates for each of 5 equally populated stimulus value bins. Again, only the last 100 trials in Block A are used for comparison so that firing rate changes are not dominated by effects of learning. **(E)** Similar data as shown in Figure 2D for a representative non-value coding neuron from the same recording session. The linear regression coefficients were not significant ($p > 0.05$ for both blocks), indicating firing rate was not significantly modulated by stimulus value in either block. **(F)** The peak firing rate during picture onset as a function of the modeled value is shown for the same representative stimulus value-coding neuron for Blocks A (late trials only) and A'. **(G)** The difference in peak firing rate between Blocks A' and A averaged across all value-coding neurons (n = 64). The peak firing rate was taken from the window [0,400) ms from target presentation. Only the last 100 trials in Block A are used for comparison so that firing rate changes are not dominated by effects of learning. A two-way ANOVA finds that there are significant main effects (stimulation condition: $F_{1,69} = 4.17$, $p < 0.05$; stimulus value: $F_{5,69} = 2.53$, $p < 0.05$), as well as a significant interaction effect between stimulation condition and value ($F_{5,69} = 4.98$, $p < 0.01$).

KEY RESOURCES TABLE

REAGENT or RESOURCE	**SOURCE**	**IDENTIFIER**
Antibodies		
Bacterial and Virus Strains		
Biological Samples		
Chemicals, Peptides, and Recombinant Proteins		
Critical Commercial Assays		
Deposited Data		
Experimental Models: Cell Lines		
Experimental Models: Organisms/Strains		
Nonhuman primate (Rhesus macaque)	UC Davis California National Primate Research Center	
Oligonucleotides		
Recombinant DNA		
Software and Algorithms		

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Python (Analysis of behavior and neural data; reinforcement learning model implementation)	Python Software Foundation	https://www.python.org
SPSS (Analysis of behavior and neural data)	IBM	https://www.ibm.com/analytics/us/en/technology/spss/
Other		