Behavioral/Systems/Cognitive

Dopamine-Mediated Reinforcement Learning Signals in the Striatum and Ventromedial Prefrontal Cortex Underlie Value-Based Choices

Gerhard Jocham,[1,2] Tilmann A. Klein,[1,4] and Markus Ullsperger[1,3]

[1]Max Planck Institute for Neurological Research, Cognitive Neurology Research Group, D-50931 Cologne, Germany, [2]FMRIB Centre, University of Oxford, John Radcliffe Hospital, Headington, Oxford OX3 9DU, United Kingdom, [3]Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behaviour, 6500 HE Nijmegen, The Netherlands, and [4]Max Planck Institute for Human Cognitive and Brain Sciences, 04103 Leipzig, Germany

A large body of evidence exists on the role of dopamine in reinforcement learning. Less is known about how dopamine shapes the relative impact of positive and negative outcomes to guide value-based choices. We combined administration of the dopamine D_2 receptor antagonist amisulpride with functional magnetic resonance imaging in healthy human volunteers. Amisulpride did not affect initial reinforcement learning. However, in a later transfer phase that involved novel choice situations requiring decisions between two symbols based on their previously learned values, amisulpride improved participants' ability to select the better of two highly rewarding options, while it had no effect on choices between two very poor options. During the learning phase, activity in the striatum encoded a reward prediction error. In the transfer phase, in the absence of any outcome, ventromedial prefrontal cortex (vmPFC) continually tracked the learned value of the available options on each trial. Both striatal prediction error coding and tracking of learned value in the vmPFC were predictive of subjects' choice performance in the transfer phase, and both were enhanced under amisulpride. These findings show that dopamine-dependent mechanisms enhance reinforcement learning signals in the striatum and sharpen representations of associative values in prefrontal cortex that are used to guide reinforcement-based decisions.

Introduction

The ability to learn about the relative value of various available options and to make choices between alternatives on the basis of these associative values is a hallmark of adaptive, goal-directed behavior. The neuromodulator dopamine, particularly in the striatum, has long been known to be crucial for reinforcement learning (Wise, 2004). Striatal dopamine is important for learning to both approach rewarding outcomes and avoid aversive outcomes (Salamone, 1994; Salamone and Correa, 2002). However, it is less clear how learning about the values of stimuli or actions translates into later choice behavior in new situations that require decisions on the basis of these learned values. A computational model of basal ganglia dopamine function suggests that dopamine in the striatum facilitates learning from positive outcomes through its action on D_1 receptors on striatonigral neurons, whereas learning from negative outcomes is assumed to be mediated by decreased dopamine transmission via D_2 receptors on striatopallidal neurons (Frank, 2005). This model has been supported by data from patients suffering from Parkinson's disease (PD), which results from a profound depletion of dopamine in the dorsal striatum (Frank et al., 2004). Similarly, individuals with a genetically driven reduction in D_2 receptor density were impaired at learning from negative but not positive outcomes (Klein et al., 2007). Notably, a central aspect of the tasks used in these studies is that they involve a reinforcement learning phase, where subjects learn the values of various stimuli, and a later transfer phase. In this transfer phase, subjects' decisions are guided by the associative values previously acquired. It is not clear whether the observed results represent solely an effect of dopamine on reinforcement learning or choice behavior. Thus, it could be argued that dopaminergic genes or PD-induced dopamine depletion acted on value-based choice behavior (i.e., behavior in the transfer phase), either instead of or in addition to action on reinforcement learning. The absence of behavioral differences during the initial learning phase in these studies is consistent with this interpretation (Frank et al., 2004; Klein et al., 2007). The ventromedial prefrontal cortex (vmPFC) is a key structure for value-based decisions. It has been implicated in encoding representations of expected value (O'Doherty, 2004; Blair et al., 2006; Padoa-Schioppa and Assad, 2006; Gläscher et al., 2009), and its integrity is crucial to human value-based choices (Bechara et al., 1994, 2000). While there is an abundance of literature on the role of dopamine in reinforcement learning, little is known about the role of this neuromodulator in signaling learned values in the vmPFC and in value-based choices.

Our aim was as follows: (1) to investigate for the first time the role of dopamine D_2 receptors in reinforcement learning

Received July 26, 2010; revised Nov. 3, 2010; accepted Dec. 1, 2010.

This work was supported by a grant from the Deutsche Forschungsgemeinschaft (JO-787/1-1) to G.J. We are grateful to Theo O. J. Gründler for help with task programming and recruiting participants.

Correspondence should be addressed to Gerhard Jocham at the above address. E-mail: gjocham@fmrib.ox.ac.uk.

DOI:10.1523/JNEUROSCI.3904-10.2011

Copyright © 2011 the authors 0270-6474/11/311606-08$15.00/0

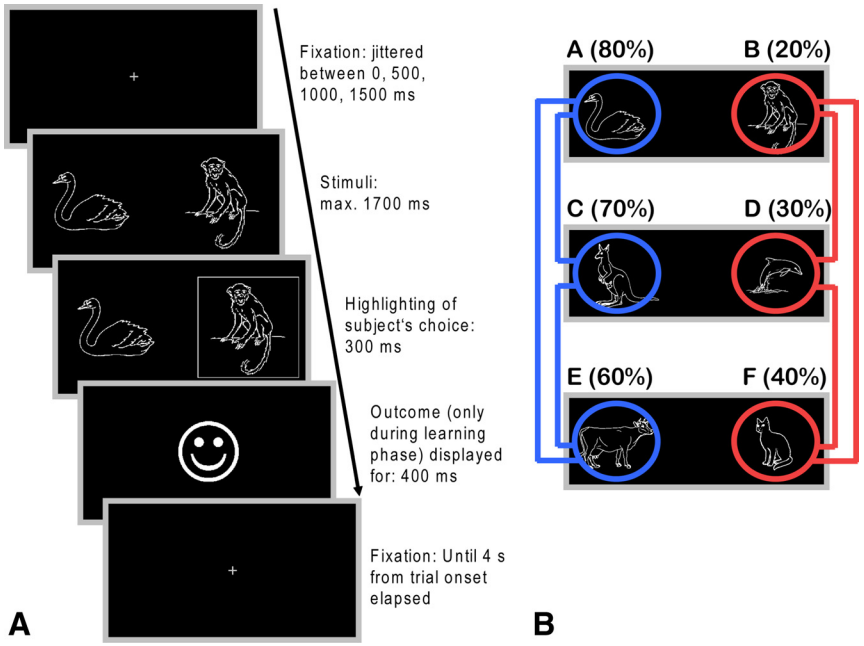

Figure 1. Sequence of stimulus events within a trial of the reinforcement learning and choice task. ***A***, Following selection of one of the two stimuli, the choice was visualized to the subject by a white frame (presented for 300 ms) around the corresponding symbol. This was immediately followed by positive or negative feedback, according to the task schedule. ***B***, In the subsequent choice phase, symbols were rearranged to yield 12 novel combinations of symbols. In addition, the three pairs from the learning phase were also presented. Trials were identical to those from the learning phase, with the exception that no outcome was presented. Of particular interest in this phase were so-called win–win trials (highlighted in blue) and lose–lose trials (highlighted in red) in which two symbols associated with a very high or very low probability of reinforcement, respectively, were combined.

signals in this kind of task with functional magnetic resonance imaging (fMRI) under a direct pharmacological challenge; and (2) to test whether dopaminergic modulation, in addition to possibly affecting reinforcement learning, also impacts representations of learned value in the vmPFC.

Materials and Methods

Subjects. Eighteen volunteers participated in the fMRI study. We included only male subjects to avoid menstrual cycle-dependent interactions between the dopaminergic system and gonadal steroids (Becker et al., 1982; Becker and Cha, 1989; Creutz and Kritzer, 2004; Dreher et al., 2007). One participant dropped out for reasons unrelated to the study, and another was excluded because of excessive head motion. The resulting 16 subjects (age: 26.13 ± 0.85 years, body weight: mean 78.75 ± 1.85 kg, mean ± SEM) were included in further data analysis. All participants gave written informed consent to the procedure, which had been approved by the local ethics committee of the Medical Faculty of the University of Cologne (Cologne, Germany).

Drug administration. Each subject received once, in a double-blind fashion, a pill that contained either placebo, the D_2-selective antagonist amisulpride (200 mg), or the D_2 receptor agonist pramipexole (0.5 mg) on separate occasions, which were separated by at least 1 week to assure complete washout of the drug before the next measurement. The order of treatments was balanced across subjects. Subjects were informed about the drugs' pharmacological properties, their general clinical use, and possible adverse effects before inclusion in the study. Exclusion criteria included history of neurological or psychiatric illness, drug abuse, and use of psychoactive drugs or medication in the 2 weeks before the experiment. In addition, subjects were instructed to abstain from alcohol and any other drugs of abuse during the entire course of the study. Because pramipexole induced a number of nonspecific effects, among them a suppression of visually evoked activity in the checkerboard control task, the data from this condition were not further analyzed and are not reported here.

Study procedure. After informed consent was obtained, subjects first completed a full version of the reinforcement learning and choice task (as described below in the next section) without fMRI scanning. This was done to avoid initial learning effects in the first fMRI session. For the two measurements with fMRI and medication (or placebo), the procedure was as follows. After arrival, volunteers completed a visual analog scale (Bond and Lader, 1974) to assess subjective effects of the drugs, such as sedation. Thereafter, measurements of heart rate and blood pressure were obtained. Next, a first venous blood sample was obtained for determination of baseline (undrugged) levels of prolactin. This was followed by administration of drug or placebo. fMRI measurements began about 150 min after drug administration, which is approximately the time at which amisulpride reaches peak blood levels. Approximately 15 min before the start of fMRI measurements (immediately before positioning in the scanner), subjects' heart rate and blood pressure were again controlled, they filled out the visual analog scales a second time, and a second blood sample for determination of prolactin levels under drug or placebo was collected. In the scanner, subjects first completed the learning and choice task, which lasted slightly less than 40 min. This was followed by a flashing checkerboard task that served to investigate nonspecific effects of the drugs on hemodynamic activity. Because neuronal dopamine receptors are sparse in the occipital cortex (Lidow et al., 1991; Hall et al., 1996), drug effects on activity induced by visual stimulation are likely not mediated by neuronal dopamine receptors. Thereafter, field maps were acquired for later B0 unwarping of the functional images. This was followed by a 60 trial forced choice reaction time task (still in the scanner, but without MRI measurements) to control for possible drug-induced psychomotor retardation. Thereafter, participants left the scanner room and completed the trail-making task (versions A and B). This was done to assess nonspecific drug effects on attention. Finally, subjects' heart rate and blood pressure were again measured and, if they felt well, subjects were paid out and released. Rating scores from the analog scales are shown in Table S1, results from the trail-making task are in Table S2, and heart rate and blood pressure results are in Table S3 (each available at www.jneurosci.org as supplemental material). Change in prolactin levels are shown in Table S4. After the final session, participants were debriefed about the purpose of the study and about the order in which they had received drug or placebo, respectively.

Reinforcement learning and choice task. The task that we used was adapted from the probabilistic selection task developed by Frank et al. (2004) (Fig. 1). The reinforcement learning and choice task consisted of two phases: an initial reinforcement learning phase and a subsequent transfer phase. During the learning phase, subjects were presented with three pairs of symbols that were probabilistically associated with reward. In each pair, one symbol was always "better" (i.e., higher reward probability) than the other, but the differences in reward probability were unequal across the three pairs: The probabilities for pairs AB, CD, and EF were 80/20, 70/30, and 60/40%, respectively. Symbol pairs were presented in random order 120 times each, thus totaling 360 trials. For each pair of symbols, the side (left/right) of presentation on the screen was pseudorandomized such that each symbol appeared on each side in half of the trials. New sets of six symbols were used on each session. Mean trial duration was 4 s. Additionally, 36 null events of the same duration were randomly interspersed. Subjects had to indicate their choice with the index finger of the left or right hand. On each trial, a central fixation cross (duration randomly jittered between 0, 500, 1000, and 1500 ms) was displayed. This was followed by presentation of the symbols, which remained on screen until the subject responded or 1700 ms elapsed. A

subject's choice was confirmed by a white frame around the corresponding symbol, which remained on screen (together with the symbols) for 300 ms. Immediately thereafter, the outcome (a smiling face indicating a reward of 0.01 Euro or a frowning face for no reward) was revealed. The cumulative reward was paid out at the end of the experiment. Trials in the transfer phase were identical to those in the learning phase, with the exception that, to prevent new learning, no outcome was revealed. Subjects were however informed that they would also receive 0.01 Euro for each "correct" choice made in this phase. In the transfer phase, subjects were confronted with the three symbol combinations from the learning phase in addition to all 12 possible novel symbol combinations. Each of these 15 symbol combinations was shown 12 times. Additionally, 18 null events of the same duration were randomly interspersed. We analyzed the number of correct choices on each symbol pair over the entire course of the learning phase and in bins of 20 trials for each pair. In the transfer phase, we computed the percentage of correct choices of the better symbol in difficult-to-decide trials: win–win trials (AC, AE, and CE) and lose–lose trials (BD, BF, and DF). These should reveal whether participants had learned more detailed value representations from positive or negative outcomes. Furthermore, we calculated the percentage of correct choices in choose A trials (AC, AD, AE, and AF) and avoid B trials (BC, BD, BE, and BF).

Forced choice reaction time task. This task was administered to test whether the drug caused psychomotor retardation. A fixation cross was presented centrally. On each of the 60 trials, a symbolic square button was presented horizontally to the left or right (30 left, 30 right in randomized order) of the fixation cross. Trials were separated by 1500 ms between the response and the onset of next stimulus. Subjects' were instructed to respond with the corresponding index finger as fast and as accurate as they could. Reaction times on this task were not affected by amisulpride relative to the drug-free condition (Table S5, available at www.jneurosci.org as supplemental material).

Trail-making task. This pencil–paper task requires subjects to connect, in ascending order, numbers with straight lines (part A) or connect letters and numbers in ascending order, alternating between letters and numbers (part B). This served as a test for attention and visuomotor speed. No differences were observed between the treatment conditions (Table S2).

Checkerboard stimulation. After an initial fixation of 10 s [equivalent to five volumes, repetition time (TR): 2 s], a checkerboard flashing at a frequency of 8 Hz was presented for 20 s. This was followed by a 20 s rest block during which only the dark gray central fixation cross was presented. Rest and stimulation periods alternated until six blocks of each were completed. The checkerboard task served to assess potential nonspecific drug effects on local blood flow. Because the occipital cortex is virtually devoid of dopamine receptors, any drug-induced changes in activity evoked by the visual stimulation reflect effects of dopamine on non-neuronal dopamine receptors. We found no differences between amisulpride and placebo in visually evoked activity. As an additional control measure, we also contrasted right- and left-handed responses during the learning phase of the reinforcement learning and choice task. Again, no differences in motor-related activity were found between drug and placebo.

Additional behavioral analyses. We analyzed whether further behavioral parameters were influenced by amisulpride. Win–stay behavior was defined as choosing the same symbol again on the next trial after having received a reward the last time this symbol was chosen. Accordingly, lose–shift behavior was defined as choice of the alternative symbol on the next trial after having received no reward the last time this symbol was chosen. Exploratory choices were defined as choices of the symbol with the lower Q value in each pair. Furthermore, reaction times for various trial types were analyzed. For the learning phase, reaction times for all three pairs of symbols (AB, CD, and EF) were calculated separately. For the transfer phase, reaction times were calculated for choose A and avoid B trials and for win–win and lose–lose trials These various trial types were analyzed overall and separated according to correct and incorrect choices. The results are shown in supplemental Table S5.

Reinforcement learning model. A standard action–value learning model (Sutton and Barto, 1998; Jocham et al., 2009) was fitted to subjects' behavior in the reinforcement learning phase. For each of the six stimuli, A to F, the model estimates an action value, $Q(A)$ to $Q(F)$, on the basis of the sequence of choices made and the outcomes experienced by the subject. These values are initialized with zero and are then updated on each trial where the respective stimulus was chosen according to the following rule: $Q_{t+1}(A) = Q_t(A) + \alpha * \delta_t$. The prediction error δ on trial t is the difference between the actual and the expected outcome: $\delta_t = r_t - Q_t(A)$, where r_t is the reward on trial t, which is either one or zero. The learning rate α scales the impact of the prediction error, i.e., the degree to which the prediction error is used to update action values. The probability of the model for selecting one particular stimulus from a pair, for instance A from the AB pair, is given by the softmax rule, which is a probabilistic choice rule: $p_t(A) = \exp(Q_t(A)/\beta)/[\exp(Q_t(A)/\beta) + \exp(Q_t(B)/\beta)]$. The parameter β reflects the subject's bias toward either exploratory (i.e., random choice of one response) or exploitatory (i.e., choice of the response with the highest Q value) behavior. Both α and β are free model parameters that are fit to subjects behavior such as to maximize the model's likelihood for the choices that were actually made by the subject. Iterations were run across both parameters from 0.001 to 1 for α and from 0.001 to 3 for β with a step size of 0.001. Thus, α and β can take values ranging from 0.001 to 1 or 3, respectively. The best fitting parameters are those that yield the highest probability of the model for the response that was actually made by the subject on any given trial. This is calculated by the log likelihood estimate: $LL = \log(\Pi_t P\_Ch_t)$. P_Ch_t is the probability of the model to make the choice that was actually made by the subject on trial t. The LL was 128.61 ± 14.08 (mean ± SEM) for the placebo and 120.73 ± 11.3 for the amisulpride condition (difference not significant).

Acquisition and analysis of fMRI data. Data acquisition was performed at 3 T on a Siemens Magnetom Trio equipped with a standard birdcage head coil. Thirty slices (3 mm thickness, 0.3 mm interslice gap) were obtained parallel to the anterior commissure–posterior commissure line using a single-shot gradient echo-planar imaging (EPI) sequence [TR: 2000 ms; echo time (TE): 30 ms; bandwidth: 116 kHz; flip angle: 90°; 64 × 64 pixel matrix; field of view (FOV): 192 mm] sensitive to blood oxygen level-dependent (BOLD) contrast. A high resolution brain image (three-dimensional reference dataset) was recorded from each participant in a separate session using an modified driven equilibrium Fourier transform sequence. For B0 unwarping of the EPI images, field maps were acquired using a gradient echo sequence (TR: 1260 ms; TE: 5.20, 9.39 and 15.38 ms; flip angle: 60°; 128 × 128 pixel matrix, FOV: 210 mm) of the same geometry as the EPI images. Analysis of fMRI data was performed using tools from the Functional Magnetic Resonance Imaging of the Brain Software Library (Smith et al., 2004). Functional data were motion corrected using rigid body registration to the central volume (Jenkinson et al., 2002). Geometric distortions in the EPI images were corrected using the field maps and an n-dimensional phase unwrapping algorithm (Jenkinson, 2003). Low-frequency signals were removed using Gaussian-weighted lines 1/100 Hz high-pass filter. Spatial smoothing was applied using a Gaussian filter with 6 mm full width at half maximum. Slice time acquisition differences were corrected using Hanning windowed sinc interpolation. EPI images were registered with the high resolution brain images and normalized into standard [Montreal Neurological Institute (MNI)] space using affine registration (Jenkinson and Smith, 2001). A general linear model was fitted into prewhitened data space to account for local autocorrelations (Woolrich et al., 2001).

For analysis I, a general linear model was set up to investigate activity related to rewards and punishments and activity related to the trial-by-trial amplitude of the reward prediction error in the learning phase. To account for the main effect of outcomes, two regressors modeling positive and negative outcomes at feedback onset were included. A third regressor consisted of the amplitude of the (signed) reward prediction error, also modeled at the time of the outcome.

For analyses IIa and b (transfer phase), we investigated which areas showed higher activity during win–win compared with lose–lose trials (a) and during choose A compared with avoid B trials (b), respectively. The respective trial types were modeled at stimulus onset and contrasted against each other. Trial types that fell into neither class were modeled as events of no interest. Analyses IIIa and b investigated which areas of the

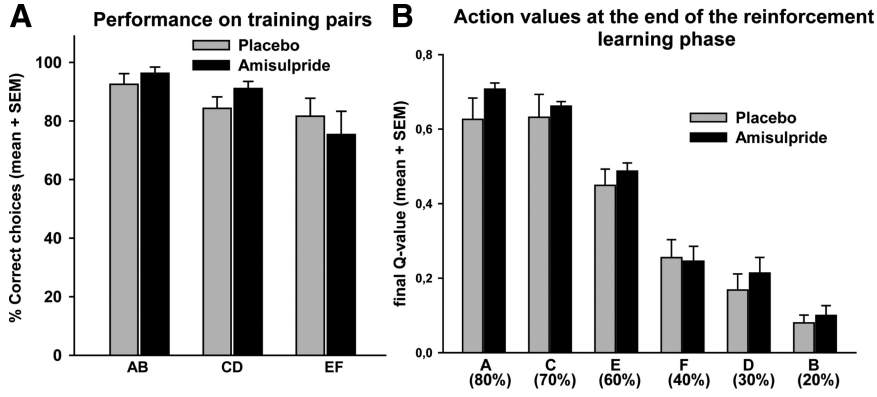

Figure 2. ***A***, Performance on probe trials of the transfer phase. The three symbol pairs previously presented in the learning phase were also administered in the transfer phase, but in the absence of an outcome. This served as a measure of how well the initial discrimination had been learned. ***B***, Action values for the six symbols at the end of the learning phase were estimated by the reinforcement learning algorithm.

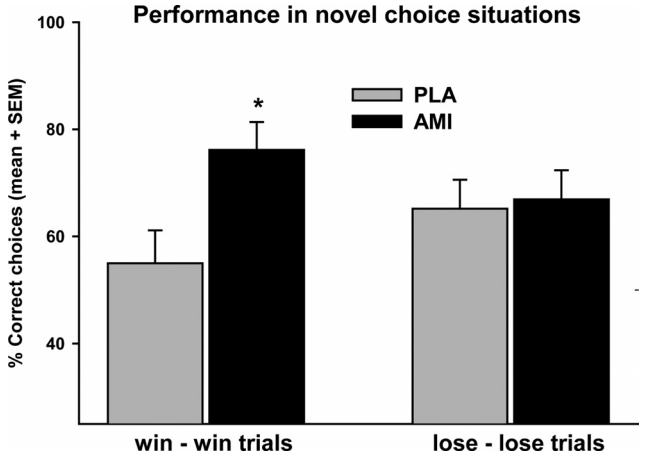

Figure 3. Percentage of correct choices of the better symbol on win–win (AC, AE, and CE) and lose–lose trials (BD, BF, and DF) in the transfer phase. *$p < 0.05$, paired t-test against placebo. AMI, Amisulpride; PLA, placebo.

brain covaried with the learned value of the stimuli in the transfer phase. Using the model-derived action values at the end of the learning phase, we constructed regressors that contained (a) for each trial the relative value of the stimulus that was actually chosen, $V_{\text{relchosen}} = V_{\text{chosen}} - V_{\text{unchosen}}$, or (b) each trial's learned value, V_{state}, which is the action values for both stimuli weighted by the model's probability (softmax choice probability) of choosing them. An additional regressor containing the onsets of the symbols was included in the general linear model to account for the main effect of stimulus presentation. Further analyses were conducted to assess potential nonspecific drug effects on hemodynamic activity (visually evoked activity in the checkerboard paradigm and motor-related activity as assessed by the contrast between left hand and right hand responses during the learning phase). All analyses were first performed separately for both drug conditions to detect patterns of activation. Subsequently, paired t-tests were performed to assess differences in brain activity between the two treatment conditions. Results are reported on the whole-brain level. Unless stated differently, we used a relatively liberal statistical threshold of $p < 0.001$, uncorrected. For correlations with behavior, parameter estimates were extracted from the peak coordinates of these contrasts.

Results

Amisulpride effects on reinforcement learning and on choices in the transfer phase

Subjects learned to reliably choose the better option from all three pairs of symbols over the course of the six learning blocks ($p < 0.001$ for main effect of block for all three symbol pairs; Fig. S1, available at www.jneurosci.org as supplemental material). Learning was generally not affected by amisulpride. There was neither an effect of drug nor a drug by block interaction in any of the three symbol pairs (all ps > 0.21). Furthermore, performance on probe trials in the transfer phase (the original pairs from the learning phase) was not modulated by amisulpride (Fig. 2*A*). A number of other behavioral parameters did not differ between conditions (see supplemental results for additional behavioral analyses and results). We fitted a reinforcement learning model to subjects' data from the learning phase. Neither the learning rate α (placebo: 0.14869 ± 0.02387; amisulpride: 0.1444 ± 0.02013, mean ± SEM) nor the temperature β (placebo: 0.2017 ± 0.0241; amisulpride: 0.2147 ± 0.01485) differed between the two treatments ($p > 0.65$). The action values at the end of the reinforcement learning phase showed a decrease with decreasing reinforcement probability (main effect of symbol: $F_{(5,75)} = 97.62$, $p < 0.001$) (Fig. 2*B*). Thus, the best options had also acquired the highest action values. However, none of these action values differed between placebo and amisulpride (no effect of group, no symbol × drug interaction, $p > 0.38$).

In the transfer phase, amisulpride significantly improved performance on win–win trials but not on lose–lose trials. Repeated-measures ANOVA yielded an interaction of drug × trial type ($F_{(1,15)} = 4.54$, $p = 0.05$). *Post hoc* paired t-test revealed that this was due to an improvement of win–win performance under amisulpride ($p = 0.012$), whereas lose–lose performance was unchanged ($p > 0.85$) (Fig. 3). Interestingly, participants' performance under placebo was above chance level only in lose–lose trials ($p = 0.006$), but not in win–win trials ($p = 0.214$, one sample t-test). We also analyzed the trial classification scheme originally used (choose A and avoid B trials) (Frank et al., 2004). We found exactly the same pattern, namely improved choose A performance under amisulpride (amisulpride vs placebo: $p = 0.023$) and no effects on avoid B performance ($p > 0.73$, preplanned comparisons). As expected from the larger differences in reward probabilities, both choose A performance and avoid B performance were significantly above chance level regardless of treatment ($p < 0.001$) (supplemental Fig. S2, available at www.jneurosci.org as supplemental material).

To rule out that the amisulpride-induced improvement in win–win and choose A performance results from interference with a different form of learning, namely extinction, we analyzed the time course of choose A and win–win performance by splitting the transfer phase into three blocks of equal length. Subjects in the transfer phase responded under extinction conditions, as no immediate outcome was presented (even though subjects received a reward for every correct choice at the end of the transfer phase). We found no evidence of extinction: Both choose A performance and win–win performance were remarkably stable across the three blocks. There was a main effect of treatment ($F_{(1,15)} = 7.79$, $p < 0.015$) but no effect of block ($F_{(2,30)} = 1.09$, $p > 0.35$) nor a treatment × block interaction ($F_{(2,30)} = 0.29$, $p > 0.75$) for win–win performance. Similar results were obtained for choose A performance.

BOLD activity in the reinforcement learning and transfer phases

During the learning phase, rewarding outcomes engaged the striatum, in particular the ventral striatum (Fig. 4). In addition, there was a small focus of activity related to reward receipt in the vmPFC under placebo, which was more pronounced under amisulpride (MNI $x = -8$, $y = 47$, $z = -11$, 244 mm^3, z-max = 4.04, paired t-test) (Fig. 4). Analysis of the parameter estimates from the peak coordinate of this difference showed that this effect was primarily driven by a more pronounced signal decrease in response to negative outcomes under amisulpride. Preplanned comparisons showed that there was stronger signal decrease to negative outcomes under amisulpride ($p = 0.044$) and no effect on reward responses ($p > 0.75$) (see Fig. 6, left).

BOLD responses correlated with the model-derived reward prediction error were found in the striatum, in particular in mid-ventral parts of the caudate nucleus as well as in the bilateral lateral prefrontal cortex and the cingulate cortex. Reward prediction error coding in the striatum was enhanced under amisulpride (MNI $x = 9$, $y = 5$, $z = -1$, 48 mm^3, z-max = 3.43). Notably, reward-related responses in the striatum and vmPFC neither correlated with performance during the learning nor the transfer phase. In contrast, prediction error responses in the striatum (parameter estimates again extracted from the peak coordinate of the difference between drug and placebo) were predictive of later performance on win–win trials in the transfer phase ($r = 0.320$, $p = 0.037$) (see Fig. 7, left).

In the transfer phase, we first found that the vmPFC showed higher activity during choose A trials than during avoid B trials in the amisulpride, but not in the placebo condition. Paired t-test also revealed a difference between amisulpride and placebo (MNI $x = 6$, $y = 59$, $z = 0$, 1165 mm^3, z-max = 3.82 and MNI $x = -8$, $y = 56$, $z = -10$, 176 mm^3, z-max = 4.19, respectively) (Fig. 5*A*). The parameter estimates from this peak coordinate (Fig. 6, right) suggest that this was due to a nonsignificant enhancement of both the signal increase to choose A trials and the signal decrease to avoid B trials under amisulpride ($p = 0.11$ and $p = 0.217$, respectively, preplanned comparisons). Because the vmPFC has been implicated in encoding the expected value of stimuli (O'Doherty, 2004; Blair et al., 2006; Padoa-Schioppa and Assad, 2006; Gläscher et al., 2009) and because no outcome is revealed, we reasoned that the observed difference in vmPFC activity could only be due to differences in the trials learned values.

We first found that activity in the vmPFC correlated with the trials' overall learned value, V_{state} in the amisulpride condition (MNI $x = -9$, $y = 43$, $z = -12$, 59 mm^3, z-max = 3.36) and at a lower threshold in the placebo condition (MNI $x = 8$, $y = 61$, $z = -19$, z-max = 2.31). Paired t-test confirmed that activity in the vmPFC correlated stronger with V_{state} in the amisulpride compared with the placebo condition (MNI $x = -16$, $y = 39$, $z = -8$, z-max = 3.15) (Fig. 5*B*). This is consistent with the poor performance of placebo-treated volunteers in win–win trials. In fact, the magnitude of vmPFC activity related to V_{state} predicted subjects' performance on win–win trials: Higher activity was associated with better performance ($r = 0.355$, $p = 0.023$) (Fig. 7, right) across subjects in both conditions. In addition, we found that activity in the posterior putamen correlated with $V_{relchosen}$ (pla-

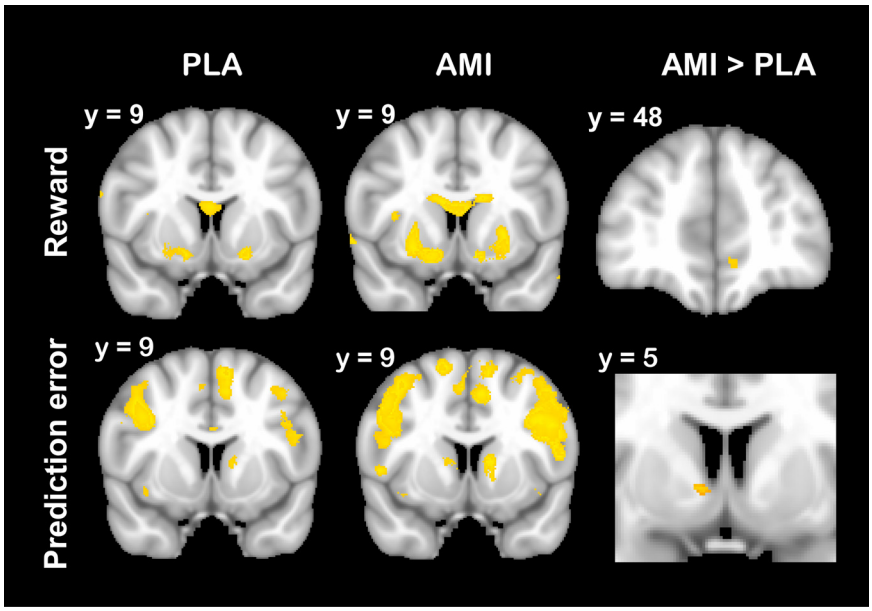

Figure 4. Signal change related to the receipt of a reward (top) and to reward prediction errors (bottom). Signal change in the striatum and ventromedial prefrontal cortex was found in both the placebo (PLA; left) and amisulpride condition (AMI; middle). Amisulpride increased both reward-related signal change in the ventromedial prefrontal cortex and prediction error-related signal change in the striatum compared with placebo (right). Images are thresholded at $z > 3.09$.

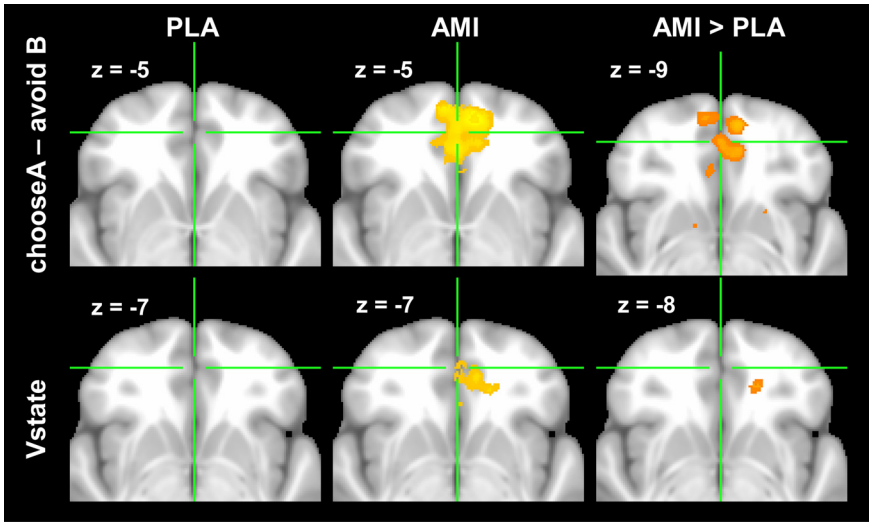

Figure 5. Signal change related to choose A versus avoid B trials (top) and to the learned value of the symbols on each trials (bottom). Amisulpride (AMI) increased the effect of choose A versus avoid B trials (right, top) and the activity related to the symbols' learned value (right, bottom) in the ventromedial prefrontal cortex. Images are thresholded at $z > 2.3$ for display purposes. The green crosshairs are positioned at $x = 0$, $y = 48$ for comparison with the effect of amisulpride on reward processing shown in the upper right-hand panel of Figure 4. PLA, Placebo.

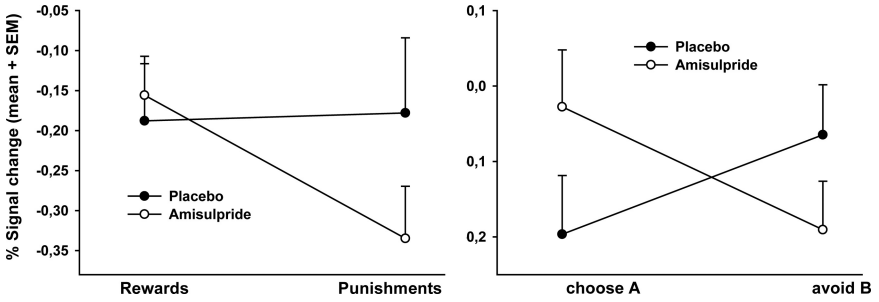

Figure 6. Percentage of signal change in response to rewards and nonrewards (left) and to choose A and avoid B trials (right) in the ventromedial prefrontal cortex during the learning and transfer phases, respectively. Signal change was extracted from the peak coordinate of the difference between amisulpride and placebo in the respective contrast. The analysis shows that the drug-induced difference in reward responses is primarily due to a stronger signal decrease to nonrewarding outcomes in the amisulpride condition. The increased response in the choose A versus avoid B contrast is driven by a nonsignificant enhancement of both the signal increase to choose A trials and the signal decrease to avoid B trials.

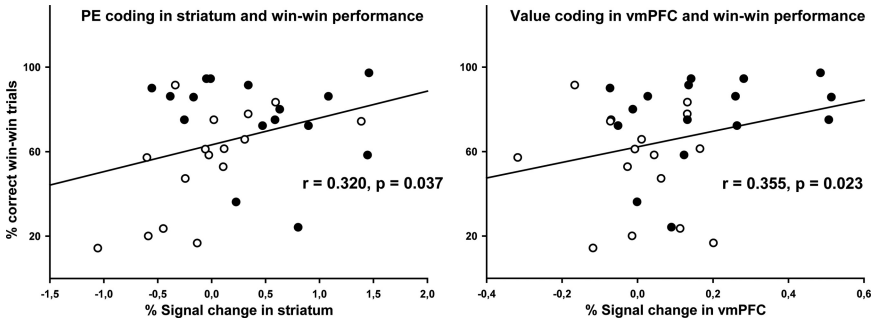

Figure 7. Signal change related to reward prediction errors in the striatum during the learning phase (left) and to the symbols' learned values in the ventromedial prefrontal cortex during the transfer phase (right) correlated with correct choices on win–win trials. White, Placebo; black, amisulpride.

cebo: MNI $x = 26$, $y = 3$, $z = -10$, 232 mm^3 and MNI $x = -27$, $y = -8$, $z = 9$, 44m^3 z-max = 3.71; amisulpride: MNI $x = 28$, $y = -12$, $z = 4$, 1185 mm^3, MNI $x = -24$, $y = -5$, $z = -7$, 914 mm^3, and MNI $x = -30$, $y = -15$, $z = 7$, 861 mm^3, z-max = 4.34) (Fig. S3, available at www.jneurosci.org as supplemental material). Identical patterns of activation were found when we repeated the analysis using V_{chosen} instead of $V_{relchosen}$ as parametric regressor. Putaminal activity related to $V_{relchosen}$ also correlated with subjects' performance on win–win trials in both the placebo ($r = 0.456$, $p = 0.038$) and amisulpride condition ($r = 0.578$, $p = 0.009$). However, striatal activity related to $V_{relchosen}$ did not differ between treatment conditions.

Discussion

This study reveals that low-dose antagonism of dopamine D_2 receptors can profoundly affect value-based choice behavior without significant effects on prior reinforcement learning. Indeed, learning to choose the more frequently rewarded of two symbols was not influenced by the drug. In accordance with this, model-estimated action values at the end of the learning phase did not differ between treatment conditions. In contrast, behavior during the transfer phase, where no performance feedback was given, was markedly altered by amisulpride. Performance on trials that required decisions between two highly reward-associated symbols (win–win trials) was improved by amisulpride. No effect was found on decisions between two very low reward-associated symbols (lose–lose trials). The same pattern was found when we compared choose A trials (novel combinations of the best symbol with all other symbols) and avoid B trials (novel combinations of the worst symbol with all other symbols): Again, amisulpride enhanced participants' ability to select the best option A but had no effect on avoiding the least fruitful option B. These effects of amisulpride on choice behavior were predicted by drug effects on brain activity during both the learning and the transfer phase. Reward prediction error (PE)-related signals in the striatum were enhanced by amisulpride, as was reward-related activity in the vmPFC. While reward-related activity in the striatum and vmPFC showed no correlation with performance in the transfer phase, PE-related BOLD activity in the striatum was predictive of later performance on win–win trials. In the transfer phase, vmPFC coded the expected value of each trial, i.e., the learned value of the two symbols on each trial, weighted by the agent's current choice policy. This signal was also predictive of performance on win–win-trials and was enhanced by amisulpride. Thus, both the enhanced coding of PE during the learning phase and of learned value during the transfer phase under amisulpride might be accountable for participants' improved performance under this treatment.

Our finding that vmPFC dynamically tracks the options' learned values dovetails well with a number of studies ascribing a central role to this brain area in valuation (O'Doherty et al., 2004; Blair et al., 2006; Plassmann et al., 2007; Hare et al., 2008; Chib et al., 2009; Gläscher et al., 2009). As our study cannot dissociate precisely which component of value computation is represented here, we do not make any specific strong claims on the exact nature of the value computation being encoded in vmPFC. Nonetheless, in line with others, we show that vmPFC represents learned value, and we extend this by showing that such representations of learned value carry a dopaminergic component, as this signal is enhanced under amisulpride. Given that blockade of dopamine D_2 receptors is well known to perturb reinforcement learning (at least at higher doses) (Salamone, 2002), it might at first glance appear puzzling that we find an enhancement of value coding under amisulpride. However, it is likely that the low dose we used led to an increase in firing of dopamine neurons and subsequently increased dopamine release in target areas because of a primary blockade of autoreceptors (Di Giovanni et al., 1998). While this is a plausible explanation for the enhanced striatal PE coding, this mechanism cannot account for the effects observed in the vmPFC. Unlike mesostriatal dopamine neurons, mesocortically projecting dopamine neurons appear to be devoid of autoreceptors (Kilts et al., 1987). We think that it is more likely that amisulpride enhanced value signals in the vmPFC by shifting the balance of D_1 versus D_2 receptor activation in favor of preferential D_1 receptor activation. An influential account of dopaminergic modulation in the PFC suggests that under preferential D_2 receptor stimulation, multiple inputs may gain access to the network, none of which is particularly strongly represented however. In contrast, under preferential D_1 receptor stimulation, the net

increase in inhibition causes PFC networks to be more robust against the impact of disturbing inputs while strong inputs are stabilized and actively maintained (Seamans and Yang, 2004). Thus, it is plausible that amisulpride treatment enhanced value signals by allowing representations of learned value to be stabilized and shielded from the impact of disturbing noise. Our finding that the enhanced response on choose A compared with avoid B trials under amisulpride was due both to an amplification of the response to choose A trials and a suppression of avoid B responses lends support to an interpretation within this framework. Alternatively, one might argue that the effects observed in the vmPFC may represent a remote effect that is triggered from a different brain structure. However, we think that this is unlikely. First, there was no area other than the vmPFC itself that showed higher activity in the transfer phase under amisulpride. Second, amisulpride has a distinctive binding profile. Despite the low density of dopamine D_2 receptors in the prefrontal cortex, it has been shown that, at low doses, amisulpride primarily occupies cortical and limbic rather than striatal receptors (Xiberas et al., 2001; Bressan et al., 2003).

It is noteworthy that amisulpride facilitated difficult decisions only when choices had to be made between two highly reward-associated symbols, but not when choosing between two very low reward-associated symbols. Because vmPFC activity covaried with V_{state}, and V_{state} is very low on lose–lose and avoid B trials, one would expect low value-related vmPFC activity in these trials. Given that the net effect of shifting the balance between preferential D_1 and D_2 receptor activation is an amplification of strong inputs, it is well conceivable that the value signal elicited during lose–lose trials is not pronounced enough to be enhanced through this mechanism, thereby allowing only trials with a high V_{state}, i.e., win–win or choose A trials, to benefit from this effect. It should be noted that this tentative mechanism cannot account for the at chance-level performance on win–win trials under placebo.

In addition to the signal in the vmPFC related to the options' learned values, we found that activity in the posterior putamen correlated with the learned value of the stimulus that was ultimately chosen. This is consistent with the manner in which the problem of action selection is thought to be resolved in the dorsal striatum (Mink, 1996; Redgrave et al., 1999). Striatal circuitry is configured such that the most salient input gains access to the motor outputs at a winner-take-all basis, while the nonselected responses are suppressed. This is consistent with what we find, since signal change in the dorsal striatum positively correlated with the learned value (i.e., "saliency") of the chosen option. This striatal coding of action value was also correlated with subjects' performance. However, the striatal signal related to the value of the chosen option was not modulated by administration of amisulpride.

The finding that amisulpride did not affect reinforcement learning seems at odds with a study that found that the D_2 antagonist haloperidol attenuated reinforcement learning and striatal prediction error coding (Pessiglione et al., 2006). This discrepancy may either be caused by differences in the drugs pharmacological profiles or it may be a dose-related effect. In fact, work in progress in our lab using a dose of amisulpride (400 mg) that yielded plasma levels sufficient to achieve blockade of a substantial proportion of striatal D_2 receptors also blunted both reinforcement learning and striatal signals.

The observed behavioral effects of low-dose dopamine D_2 receptor antagonism appear consistent with the framework suggested by Frank (2005). However, our findings go beyond existing knowledge and indicate that, in addition to learning from positive and negative outcomes, dopamine is also involved in choice behavior on the basis of learned values, i.e., after learning has already taken place. It may appear puzzling that amisulpride affected choice behavior only in the transfer phase. One might argue that the learning phase also involves choices, and these choices should be equally affected by amisulpride. However, the learning phase involves direct comparisons between two symbols for which the outcome is always presented. This is clearly different from the transfer phase, where subjects are confronted with entirely new choice situations and where they never receive any feedback about whether their choices were correct, i.e., where they have no opportunity to learn from their choices.

Together, our findings show that antagonism of D_2 receptors improved performance on choices between two highly reward-associated options and concurrently enhanced prediction error coding in the striatum and tracking of learned value in the vmPFC. We speculate that by blocking autoreceptors and thereby enhancing striatal prediction error coding and by shifting the balance of D_1 versus D_2 receptor activation toward preferential D_1 activation in the prefrontal cortex, amisulpride helped to stabilize representations of learned value and thereby improved choice performance.

References

Bechara A, Damasio AR, Damasio H, Anderson SW (1994) Insensitivity to future consequences following damage to human prefrontal cortex. Cognition 50:7–15.

Bechara A, Tranel D, Damasio H (2000) Characterization of the decision-making deficit of patients with ventromedial prefrontal cortex lesions. Brain 123:2189–2202.

Becker JB, Cha JH (1989) Estrous cycle-dependent variation in amphetamine-induced behaviors and striatal dopamine release assessed with microdialysis. Behav Brain Res 35:117–125.

Becker JB, Robinson TE, Lorenz KA (1982) Sex differences and estrous cycle variations in amphetamine-elicited rotational behavior. Eur J Pharmacol 80:65–72.

Blair K, Marsh AA, Morton J, Vythilingam M, Jones M, Mondillo K, Pine DC, Drevets WC, Blair JR (2006) Choosing the lesser of two evils, the better of two goods: specifying the roles of ventromedial prefrontal cortex and dorsal anterior cingulate in object choice. J Neurosci 26:11379–11386.

Bond AJ, Lader MH (1974) The use of analogue scales in rating subjective feelings. Br J Med Psychol 47:211–218.

Bressan RA, Erlandsson K, Jones HM, Mulligan R, Flanagan RJ, Ell PJ, Pilowsky LS (2003) Is regionally selective D_2/D_3 dopamine occupancy sufficient for atypical antipsychotic effect? An in vivo quantitative [^{123}I]epidepride SPET study of amisulpride-treated patients. Am J Psychiatry 160:1413–1420.

Chib VS, Rangel A, Shimojo S, O'Doherty JP (2009) Evidence for a common representation of decision values for dissimilar goods in human ventromedial prefrontal cortex. J Neurosci 29:12315–12320.

Creutz LM, Kritzer MF (2004) Mesostriatal and mesolimbic projections of midbrain neurons immunoreactive for estrogen receptor beta or androgen receptors in rats. J Comp Neurol 476:348–362.

Di Giovanni G, Di Mascio M, Di Matteo V, Esposito E (1998) Effects of acute and repeated administration of amisulpride, a dopamine D_2/D_3 receptor antagonist, on the electrical activity of midbrain dopaminergic neurons. J Pharmacol Exp Ther 287:51–57.

Dreher JC, Schmidt PJ, Kohn P, Furman D, Rubinow D, Berman KF (2007) Menstrual cycle phase modulates reward-related neural function in women. Proc Natl Acad Sci U S A 104:2465–2470.

Frank MJ (2005) Dynamic dopamine modulation in the basal ganglia: a neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism. J Cogn Neurosci 17:51–72.

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

Gläscher J, Hampton AN, O'Doherty JP (2009) Determining a role for ventromedial prefrontal cortex in encoding action-based value signals during reward-related decision making. Cereb Cortex 19:483–495.

Hall H, Farde L, Halldin C, Hurd YL, Pauli S, Sedvall G (1996) Autoradiographic localization of extrastriatal D_2-dopamine receptors in the human brain using [^{125}I]epidepride. Synapse 23:115–123.

Hare TA, O'Doherty J, Camerer CF, Schultz W, Rangel A (2008) Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. J Neurosci 28:5623–5630.

Jenkinson M (2003) Fast, automated, *N*-dimensional phase-unwrapping algorithm. Magn Reson Med 49:193–197.

Jenkinson M, Smith S (2001) A global optimisation method for robust affine registration of brain images. Med Image Anal 5:143–156.

Jenkinson M, Bannister P, Brady M, Smith S (2002) Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17:825–841.

Jocham G, Neumann J, Klein TA, Danielmeier C, Ullsperger M (2009) Adaptive coding of action values in the human rostral cingulate zone. J Neurosci 29:7489–7496.

Kilts CD, Anderson CM, Ely TD, Nishita JK (1987) Absence of synthesis-modulating nerve terminal autoreceptors on mesoamygdaloid and other mesolimbic dopamine neuronal populations. J Neurosci 7:3961–3975.

Klein TA, Neumann J, Reuter M, Hennig J, von Cramon DY, Ullsperger M (2007) Genetically determined differences in learning from errors. Science 318:1642–1645.

Lidow MS, Goldman-Rakic PS, Gallager DW, Rakic P (1991) Distribution of dopaminergic receptors in the primate cerebral cortex: quantitative autoradiographic analysis using [^{3}H]raclopride, [^{3}H]spiperone and [^{3}H]SCH23390. Neuroscience 40:657–671.

Mink JW (1996) The basal ganglia: focused selection and inhibition of competing motor programs. Prog Neurobiol 50:381–425.

O'Doherty JP (2004) Reward representations and reward-related learning in the human brain: insights from neuroimaging. Curr Opin Neurobiol 14:769–776.

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

Padoa-Schioppa C, Assad JA (2006) Neurons in the orbitofrontal cortex encode economic value. Nature 441:223–226.

Pessiglione M, Seymour B, Flandin G, Dolan RJ, Frith CD (2006) Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. Nature 442:1042–1045.

Plassmann H, O'Doherty J, Rangel A (2007) Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. J Neurosci 27:9984–9988.

Redgrave P, Prescott TJ, Gurney K (1999) The basal ganglia: a vertebrate solution to the selection problem? Neuroscience 89:1009–1023.

Salamone JD (1994) The involvement of nucleus accumbens dopamine in appetitive and aversive motivation. Behav Brain Res 61:117–133.

Salamone JD (2002) Functional significance of nucleus accumbens dopamine: behavior, pharmacology and neurochemistry. Behav Brain Res 137:1.

Salamone JD, Correa M (2002) Motivational views of reinforcement: implications for understanding the behavioral functions of nucleus accumbens dopamine. Behav Brain Res 137:3–25.

Seamans JK, Yang CR (2004) The principal features and mechanisms of dopamine modulation in the prefrontal cortex. Prog Neurobiol 74:1–58.

Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TE, Johansen-Berg H, Bannister PR, De Luca M, Drobnjak I, Flitney DE, Niazy RK, Saunders J, Vickers J, Zhang Y, De Stefano N, Brady JM, Matthews PM (2004) Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 23 [Suppl 1]:S208–S219.

Sutton RS, Barto AG (1998) Reinforcement Learning: An Introduction. Cambridge, MA: MIT.

Wise RA (2004) Dopamine, learning and motivation. Nat Rev Neurosci 5:483–494.

Woolrich MW, Ripley BD, Brady M, Smith SM (2001) Temporal autocorrelation in univariate linear modeling of FMRI data. Neuroimage 14:1370–1386.

Xiberas X, Martinot JL, Mallet L, Artiges E, Canal M, Loc'h C, Mazière B, Paillere-Martinot ML (2001) In vivo extrastriatal and striatal D_2 dopamine receptor blockade by amisulpride in schizophrenia. J Clin Psychopharmacol 21:207–214.