Behavioral/Cognitive

Differential Modulation of Reinforcement Learning by D2 Dopamine and NMDA Glutamate Receptor Antagonism

Gerhard Jocham,[1,2,4] **Tilmann A. Klein,**[5,6] **and Markus Ullsperger**[2,3,7]

[1]Faculty of Economics and Management, [2]Center for Behavioral Brain Sciences, [3]Department of Neuropsychology, Otto-von-Guericke-University Magdeburg, D-39106 Magdeburg, Germany, [4]Max Planck Institute for Neurological Research, D-50931 Cologne, Germany, [5]Max Planck Institute for Human Cognitive and Brain Sciences, D-04103 Leipzig, Germany, [6]Day Clinic for Cognitive Neurology, University Hospital Leipzig, D-04103 Leipzig, Germany, and [7]Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, 6525 EZ Nijmegen, The Netherlands

The firing pattern of midbrain dopamine (DA) neurons is well known to reflect reward prediction errors (PEs), the difference between obtained and expected rewards. The PE is thought to be a crucial signal for instrumental learning, and interference with DA transmission impairs learning. Phasic increases of DA neuron firing during positive PEs are driven by activation of NMDA receptors, whereas phasic suppression of firing during negative PEs is likely mediated by inputs from the lateral habenula. We aimed to determine the contribution of DA D2-class and NMDA receptors to appetitively and aversively motivated reinforcement learning. Healthy human volunteers were scanned with functional magnetic resonance imaging while they performed an instrumental learning task under the influence of either the DA D2 receptor antagonist amisulpride (400 mg), the NMDA receptor antagonist memantine (20 mg), or placebo. Participants quickly learned to select ("approach") rewarding and to reject ("avoid") punishing options. Amisulpride impaired both approach and avoidance learning, while memantine mildly attenuated approach learning but had no effect on avoidance learning. These behavioral effects of the antagonists were paralleled by their modulation of striatal PEs. Amisulpride reduced both appetitive and aversive PEs, while memantine diminished appetitive, but not aversive PEs. These data suggest that striatal D2-class receptors contribute to both approach and avoidance learning by detecting both the phasic DA increases and decreases during appetitive and aversive PEs. NMDA receptors on the contrary appear to be required only for approach learning because phasic DA increases during positive PEs are NMDA dependent, whereas phasic decreases during negative PEs are not.

Key words: **dopamine; glutamate; NMDA; prediction error; reinforcement learning; striatum**

Introduction

Learning to select those stimuli or actions that result in rewards and to avoid those that result in punishments is crucial for adaptive success. Formal learning theory posits that learning is driven by prediction errors (PEs), the difference between experienced and expected outcome (Rescorla and Wagner, 1972; Sutton and Barto, 1998). If an action leads to an outcome that is better than expected, this results in a positive PE, and the action's value is increased. If the outcome is worse than expected, a negative PE results and the value of the action is decreased. The firing pattern of midbrain dopamine (DA) neurons in a number of species, including humans, has been shown to reflect PE (Schultz et al., 1997; Hollerman and Schultz, 1998; Satoh et al., 2003; Bayer and Glimcher, 2005; Roesch et al., 2007; Takahashi et al., 2009; Zaghloul et al., 2009). It is commonly assumed that this dopaminergic PE is then broadcast to target structures of the midbrain DA system, in particular the basal ganglia. There, it is thought to act as a teaching signal by modifying plasticity at corticostriatal synapses (Schultz et al., 1997; Shen et al., 2008). Activity correlated with PE has also been found in a vast number of human neuroimaging studies, where it is usually observed in areas that receive the densest DA projections, such as striatum and prefrontal cortex (O'Doherty et al., 2004; Pessiglione et al., 2006, 2008; Jocham et al., 2009, 2011; Valentin and O'Doherty, 2009). More recently, human functional magnetic resonance imaging (fMRI) studies have also found that PE correlates directly in the DA neuron containing midbrain substantia nigra and ventral tegmental area (D'Ardenne et al., 2008; Klein-Flügge et al., 2011). Striatal PE correlates with learning success (Schönberg et al., 2007), and precise temporal modulation of DA neuron firing using optogenetics indicates a causal role for PEs in learning (Steinberg et al., 2013). Evidence accrues that positive and negative PEs are generated by different mechanisms. NMDA glutamate receptors located on midbrain DA neurons appear essential for the phasic bursts during positive PEs (Overton and Clark, 1992; Wang et al., 2011). In contrast, inputs from the lateral

Received Feb. 24, 2014; revised Aug. 18, 2014; accepted Aug. 19, 2014.

Author contributions: G.J., T.A.K., and M.U. designed research; G.J. and T.A.K. performed research; G.J. analyzed data; G.J., T.A.K., and M.U. wrote the paper.

This work was supported by a grant from the Deutsche Forschungsgemeinschaft (DFG JO-787/1-1) and by the federal state of Saxony-Anhalt, Project: Center for Behavioral Brain Sciences to G.J., M.U. was supported by a grant from the Deutsche Forschungsgemeinschaft (SFB 779 A12). We thank Jushan Bai and Pierre Perron for making the Matlab code for their breakpoint test available online.

The authors declare no competing financial interests.

Correspondence should be addressed to Gerhard Jocham, Center for Behavioral Brain Sciences, Faculty of Economics, Otto-von-Guericke-Universität Magdeburg, Gebäude 22, Universitätsplatz 2, D-39106 Magdeburg, Germany. E-mail: jocham@ovgu.de.

DOI:10.1523/JNEUROSCI.0757-14.2014

habenula, which is excited by aversive events (Ullsperger and von Cramon, 2003; Matsumoto and Hikosaka, 2009; Bromberg-Martin et al., 2010) seem to drive suppression of DA neuron firing during negative PEs (Matsumoto and Hikosaka, 2007). Our present study was aimed to determine the different roles of NMDA and DA receptors in instrumental approach and avoidance learning. We hypothesized that NMDA receptor antagonism would impair learning to approach appetitive stimuli by interfering with the generation of positive PEs, while having no effect on learning to avoid aversive stimuli. In contrast, we expected that blockade of DA D2 receptors would interfere with both approach and avoidance learning, because those receptors are sensitive to both increases and decreases in extracellular DA levels associated with rewards and punishments, respectively. We addressed this hypothesis by scanning healthy human volunteers with fMRI while they performed an instrumental approach and avoidance learning task under the influence of either placebo or antagonists of D2-like and NMDA receptors.

Materials and Methods

Participants. Twenty-four healthy male participants [age, 25.46 ± 0.53 years; body weight, 80.5 ± 2.6 kg (mean ± SEM)] participated in the experiment. We excluded female 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). We excluded two participants because they performed a substantial number of misses on ≥1 of the three sessions (≥40 misses on 270 trials; see Results), which prevented reliable calculation of learning curves. All analyses reported here are from the final sample of $n = 22$ volunteers. 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).

Study procedure. Each subject received the DA D2-class receptor antagonist amisulpride (400 mg), the glutamate NMDA receptor antagonist memantine (20 mg), or placebo in a double-blind cross-over design. Sessions were separated by ≥2 weeks to ensure complete washout of the drug before the next session. The order of treatments was balanced across subjects. Volunteers 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, history of 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.

After arrival, subjects first 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. This was followed by administration of a pill that contained either drug or placebo. fMRI measurements began 240 min after administration of drug or placebo, which is approximately the time at which those drugs reach peak blood levels. Fifteen minutes before the start of fMRI measurements (immediately before positioning in the scanner), subjects' heart rate and blood pressure were again controlled and they received a few practice trials on the learning task. In the scanner, subjects first completed the reinforcement learning task, which lasted slightly <40 min. 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 slowing. Thereafter, participants left the scanner room and completed the Trail Making Test Part B. This was done to assess nonspecific drug effects on attention. Then they filled out the visual analog scales a second time. Finally, subjects' heart rate and blood pressure were again measured and, if they felt well, subjects were paid and released. 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. We analyzed drug effects on heart rate and on systolic and diastolic blood pressures by separate two-way repeated-measures ANOVA with the factors DRUG (three levels) and TIME (two levels; we only considered the baseline measurement and the measurement immediately before fMRI). Drug effects on subjective mood (measured by visual analog scales) were likewise tested by subjecting the visual analog scales to two-way repeated-measures ANOVA with the factors DRUG (three levels) and ITEM (16 scale items).

Reinforcement learning task. We used an instrumental approach and avoidance learning task similar to that used by Pessiglione and colleagues (Fig. 1; Pessiglione et al., 2008; Fischer and Ullsperger, 2013). Each trial started with presentation of a central fixation cross for a mean duration of 1150 ms (jittered between 400, 900, 1400, and 1900 ms). Next, an abstract symbol was presented centrally. One second after symbol onset, participants were shown a response prompt that consisted of a check sign and an X to the left and right of the option, respectively (side of check sign and X balanced across participants). By pressing a left or right button, subjects could indicate whether they wanted to select (check sign) or reject (X) the option shown to them. There was a response window of 1.5 s within which subjects were required to respond. Each option had a fixed probability p of yielding a monetary reward when chosen and a fixed probability $1 - p$ of yielding a monetary loss when chosen. Once a response was made, the selected action (check sign for choose, X for avoid) was highlighted by a rectangular frame until the outcome was presented. Outcomes were revealed after a mean delay of 3.25 s (jittered between 2.5, 3.0, 3.5, and 4.0 s) and consisted of a smiling face for rewarded trials and a frowning face for punished trials. Outcomes were shown for 700 ms. If no choice was registered within the required response window, the words "please respond faster" appeared on the screen, followed directly by the intertrial interval. Thus, no outcomes were displayed on miss trials and hence no learning was possible following a missed response. Importantly, on trials where participants decided to avoid the option presented to them, they were nevertheless shown the outcome that would have happened if they had chosen the option, but this feedback was crossed out and shown in darker gray, indicating to them that this had no financial consequences for them. This manipulation allowed subjects to update their value estimates on each trial, regardless of their propensity to choose or avoid an option. For each rewarded approach response, participants earned €0.1; for each punished approach response, €0.1 was deducted. Outcome presentation was followed by an intertrial interval of 0.4–3.4 s, during which a black blank screen was presented. Participants completed three blocks of 90 trials each. In addition, 30 null events (black screen for 6 s) were randomly interspersed with the regular trials, resulting in a total experiment duration of ~40 min. In each block of 90 trials, there were three different symbols that were each shown 30 times. Depending on their associated probabilities to yield reward or punishment when chosen, they were designated as "good" (reward = 0.7, punishment = 0.3), "neutral" (0.5/0.5), or "bad" (0.3/0.7) options. Thus, subjects' goal was to learn to approach (select) the good options in each block and to avoid (reject) the bad option. For the neutral option, the ideal choice policy was to either always approach it or always avoid it. Subjects were not explicitly informed that there was always one good, one bad, and one neutral option. They were only told that across the entire experiment, there were options that lead to losses and some that lead to wins in the longer term. We calculated learning curves separately for the three options by averaging subjects' choices (avoid = 0, select = 1) over the three blocks. *Post hoc*, we observed that drug effects on learning became only apparent toward the end of each block when performance reached an asymptote. We therefore used the Bai–Perron multiple break point test (Bai and Perron, 1998) to identify the number and location of structural breaks in the learning curves. The test was implemented using the Matlab code provided online by the authors. Using a minimum window of k trials (here, we set k to five trials), the test can detect whether one or more breaks in the curve exist and, if so, and how many there are, based on the regression slope of each possible segment and the resulting model fits. The test may find any number of breaks between 0 and $n/k - 1$ trials, where n equals the number of trials. Thus, in our case, the test could reveal that between 0 and 5 breaks exist. The procedure was run using data from all subjects

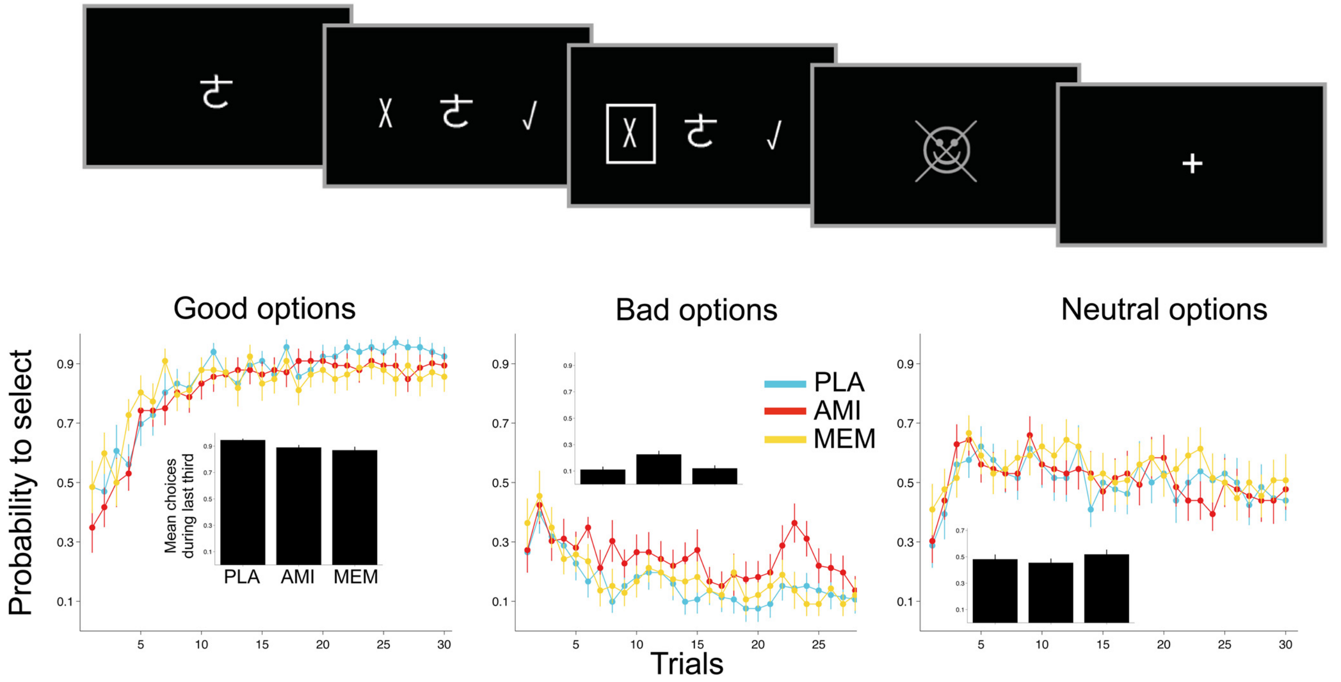

Figure 1. Top, Task schematic. Bottom, Learning curves for the rewarding (left), punishing (middle), and neutral (right) options under the three drug-treatment conditions. Values are mean ± SEM. Statistics were computed on the averaged scores over the final segment of the block (bar graph insets, group means + SEM). PLA, Placebo; AMI, amisulpride; MEM, memantine.

and all sessions, separately for the good and bad symbols. For both good and bad symbols, the test detected two breaks that cut the learning curves into three segments (see Results). We then calculated an index of final learning performance by averaging the choice probabilities over the last segment of the learning curve. This index was then subjected to two-way repeated-measures ANOVA with the factors DRUG (three levels) and VALENCE (two levels). Paired t tests were used as *post hoc* tests.

Forced-choice reaction-time task. This task was administered to test whether the drugs caused a general response slowing. 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 accurately as they could. Subjects' response times and number of correct responses were compared between drug treatments using one-way repeated-measures ANOVA with the factor DRUG (three levels).

Trail Making Test Part B. Trail Making Test Part B, a pencil–paper task, requires subjects to connect letters and numbers in ascending order, alternating between letters and numbers. This served as a test for attention and visuomotor speed. Time required for completion of the task was compared between drug treatments using one-way repeated-measures ANOVA with the factor DRUG (three levels).

Reinforcement learning model. To obtain trial-by-trial estimates of subjects' reward expectation and PEs, we fitted a simple Rescorla–Wagner model (Sutton and Barto, 1998; Jocham et al., 2011) to participants' sequence of choices (approach or avoid) in the learning task. For each of the three options in each block, the model learns a value by performing the following update on each trial: $V(k)_{t+1} = V(k)_t + \alpha\delta(k)_t$, where V is the estimated value of option k on trial t, δ_t is the PE on trial t, and α is the learning rate that determines the degree to which δ is used to update option values. δ_t Is the difference between r_t, the reward obtained in trial t, and $V(k)_t$, the reward expected from option k on trial t: $\delta(k)_t = r_t - V(k)_t$.

The model's probability of selecting option k on trial t is then given by a softmax choice rule as follows:

$$\frac{1}{1 + exp(-V(k)_t/\tau)}$$

where the softmax temperature τ is another free parameter that reflects the degree of stochasticity in subjects' choices. We used a Bayesian estimation procedure custom-implemented in Matlab [MathWorks, Research Resource Identifier (RRID):nlx_153890] routines to obtain the free model parameters α and τ that best describe subjects' behavior. The parameter space was set up as a two-dimensional grid in log space with 500 points in each dimension. The joint posterior distribution of the unknown model parameters is then given by the product of choice probabilities over trials under each possible parameter combination in the grid. The marginal posterior distributions on each parameter were then obtained by marginalizing (direct numerical integration) over the remaining dimension of the grid. The optimal parameters were then given by the means of the resulting marginal posterior distributions.

Relative value learning model. When regressing the model-derived PE against our fMRI data, we made an intriguing observation. PEs showed the usual positive correlation with the BOLD signal on approach trials (when subjects selected the option presented to them), but correlated negatively on avoid trials (when subjects rejected the option; see Fig. 3). By definition ($\delta_t = r_t - V_t$) the PE consists of separable contributions of the outcome term (higher rewards result in higher PE) and the expectation term (a higher expectation leads to lower PE). When decomposing the PE into the outcome and expectation terms, we found that both regressors had reverse effects on the BOLD signal on approach versus avoid trials: there was a positive correlation of the fMRI signal with the outcome on approach trials, but a negative correlation on avoid trials. Likewise, there was a negative correlation with expected value on approach trials, but a positive correlation on avoid trials. This is unexpected from standard reinforcement learning models. A learner that learns the value of the stimuli would always experience a positive PE on rewarded trials and a negative PE on punished trials, regardless of the action taken. A model that learns state–action value pairs like SARSA (state–action–reward–state–action) would behave the same way on approach trials, but generate no PE at all on avoid trials: subjects are clearly instructed that avoiding the option always results in an outcome of zero, so the expected value of option k given action A = avoid is always zero. If the outcome is set to be the really experienced outcome (zero), the PE is thus always zero; if the outcome is set to be the outcome that would have happened, the resulting PE would again be the same as in standard value learning (positive on rewarded trials, negative on punished trials). However, an alternative possibility is that subjects learn a relative action value, that is, how good action *A* is in comparison to the corresponding alternative action. For example, a reward following an approach response indicates to the subject that it has indeed been better to select this option rather than to avoid it. The same rewarding outcome following an avoid response, however, indicates to the subject that it would have been better to take the alternative approach action instead. We therefore use a simple modification of the above model, in which a relative value *RV* is learned by $RV(k)_{t+1} = RV(k)_t + \alpha\delta(k)_t$.

Here, the PE δ is computed as follows: $\delta(k)_t = ro_t - RV(k)_t$ where *ro* is the relative outcome, that is, the difference between the outcome obtained under the action taken by the subject and the outcome that would have happened under the alternative course of action, $o_{\text{chosen}} - o_{\text{unchosen}}$. Thus, *ro* simply corresponds to the experienced outcome on approach trials ($ro = o_{\text{chosen}} - 0$), but is the inverse on avoid trials [$ro = 0 - 1$ for rewarded trials and $ro = 0 - (-1) = 1$ for punished trials]. The net effect of this algorithm in the context of our experiment is therefore simply that the PE flips sign on avoid trials, which allows us to analyze approach and avoid trials together.

MRI data acquisition. MRI data were acquired on a 3T Siemens Magnetom Trio equipped with a standard birdcage head coil. We acquired 1170 volumes per subject, with 30 slices (voxel size, 3 mm isotropic; interslice gap, 0.3 mm) obtained parallel to the anterior commissure–posterior commissure line using a single-shot gradient echo-planar imaging (EPI) sequence [TR, 2000 ms; TE, 30 ms; bandwidth, 116 kHz; flip angle, 90°; 64 × 64 pixel matrix; field of view (FOV), 192 mm] sensitive to BOLD contrast. A high-resolution brain image (three-dimensional reference dataset) was recorded from each participant in a separate session using a 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.

MRI data analysis. Analysis of fMRI data was performed using tools from the Functional Magnetic Resonance Imaging of the Brain Software Library (Smith et al., 2004; RRID:birnlex_2067). Functional data were motion-corrected using rigid-body registration to the central volume (Jenkinson et al., 2002), corrected for geometric distortions using the field maps and an *n*-dimensional phase-unwrapping algorithm (Jenkinson, 2003), high-pass filtered using a Gaussian-weighted lines 0.01 Hz filter and 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 (MNI) space using affine registration (Jenkinson and Smith, 2001). A general linear model (GLM) was fitted into prewhitened data space to account for local autocorrelations (Woolrich et al., 2001). We constructed two separate GLMs, one designed to investigate correlates of PEs at the whole-brain level, the other to search for outcome-related activity. Note that all of our analyses focus on region-of-interest (ROI) analyses in the striatum. The results of GLM1 merely serve to display the whole-brain pattern of PE-related activity across all subjects, regardless of treatment condition. GLM2 provides the basis for defining orthogonal ROIs for subsequent analyses of drug-induced differences in the striatum. We focus our analyses on the striatum as the key region implicated in reinforcement learning under the control of reward PEs. GLM1 contained two regressors of interest: the relative value (modeled at stimulus onset), derived from the relative value learning model, and the relative outcome (modeled at outcome onset). These two regressors were entered into the design matrix along with regressors modeling the main effect of stimulus presentation, outcome presentation, and responses made by subjects. In addition, the six motion parameters from the motion correction were included in the model. Contrasts were calculated for the expected value (−1, because we expected negative correlations with expected value), relative outcome, and the contrast for PE was set up by contrasting the relative outcome regressor with the relative value regressor. GLM2 was identical to GLM1 except that the regressors coding for relative value and relative outcome were replaced by regressors coding for the actual outcomes on each trial and the foregone outcomes, respectively. The contrast for obtained outcomes was computed for this GLM. Contrast images from the first level were then taken to the group level using a random-effects analysis. Results are reported at a threshold of $p < 0.001$, uncorrected.

ROI analyses. To obtain independent ROIs that are not subject to any selection bias, we used the contrast of outcome-related activity from GLM2 (see above) across all subjects in all three treatment conditions. This yielded robust activations throughout the brain, particularly pronounced in the striatum (peak MNI coordinates: $x = -17$, $y = 7$, $z = -13$, *Z*-max = 7.28, and $x = 15$, $y = 9$, $z = -13$, *Z*-max = 7.27). The contrast image was thresholded at $z > 3.9$ and multiplied with anatomical masks, first for the entire striatum, and subsequently separately for the caudate nucleus, putamen, and ventral striatum from the Harvard–Oxford subcortical atlas. We merged the resulting masks for the two hemispheres, thus resulting in one ROI each for caudate, putamen, ventral striatum, and entire striatum. We extracted raw BOLD signal time courses from the above ROI. Using custom-written Matlab (RRID: nlx_153890) code, the following steps were then performed. The time series of each volunteer was cut into trials with a duration of 14.31 s, where options were presented at 0 s, the response was made at 1.42 s, and the outcome was presented at 4.31 s, which corresponds to the mean onsets of these events across subjects, trials, and sessions. Time series were resampled to a resolution of 200 ms. A GLM containing the parameters of interest was then fitted at each time point for each volunteer. This resulted in a time course of effect sizes for each regressor in the design matrix and for each volunteer. These time courses were then averaged across participants. The GLM was set up to decompose the PE into its constituent terms, expectation and outcome. For a signal to truly represent a PE, it has to comply with two key formal criteria: (1) it has to be sensitive to variations in both expectation and outcome, and (2) the effect of outcome on the BOLD signal has to be positive because, for any given expectation, a larger reward will generate a larger PE. In contrast, for any given outcome, a higher expectation decreases the PE; hence the effect of expectation on the BOLD signal (at the time of outcome) has to be negative (Caplin and Dean, 2008; Rutledge et al., 2010). The design matrix therefore contained the relative values and relative outcomes, with separate regressors for appetitive and aversive trials. Appetitive and aversive here refers to the outcome available for the symbol on each trial, e.g., trials on which a reward is observed (experienced or foregone) are considered “appetitive,” whereas trials where a punishment is observed are considered “aversive.” This classification is due to our hypothesis that amisulpride would be involved in both approach and avoidance learning by reporting both the increase and decrease in DA levels during positive and negative PEs, respectively, whereas we expected memantine to interfere selectively with approach learning by preventing the phasic increases in DA levels during positive PEs. For statistical testing, the time course of effect sizes was summed in the time window from 4–8 s after outcome presentation, which is where the peak effects are expected to occur. These values were then entered into a two-way repeated-measures ANOVA with the factors DRUG (three levels) and VALENCE (two levels). Paired *t* tests were used as *post hoc* tests.

Results

Behavior

We excluded two participants because of an extreme number of misses (>40) in ≥1 of the three sessions, which prevented reliable computation of learning curves. Response misses were rare among the remaining volunteers (median, 1, 2, and 2 misses for the placebo, amisulpride, and memantine conditions) and did not differ between drug conditions ($p > 0.08$). All further analyses were therefore performed on this final sample of $n = 22$ subjects.

Participants quickly learned to select rewarding options and to avoid punishing stimuli (Fig. 1). In contrast, for neutral options, the percentage of subjects selecting this option remained at ~50% throughout the course of learning. When considering the entire course of learning, no clear differences in approach learn-

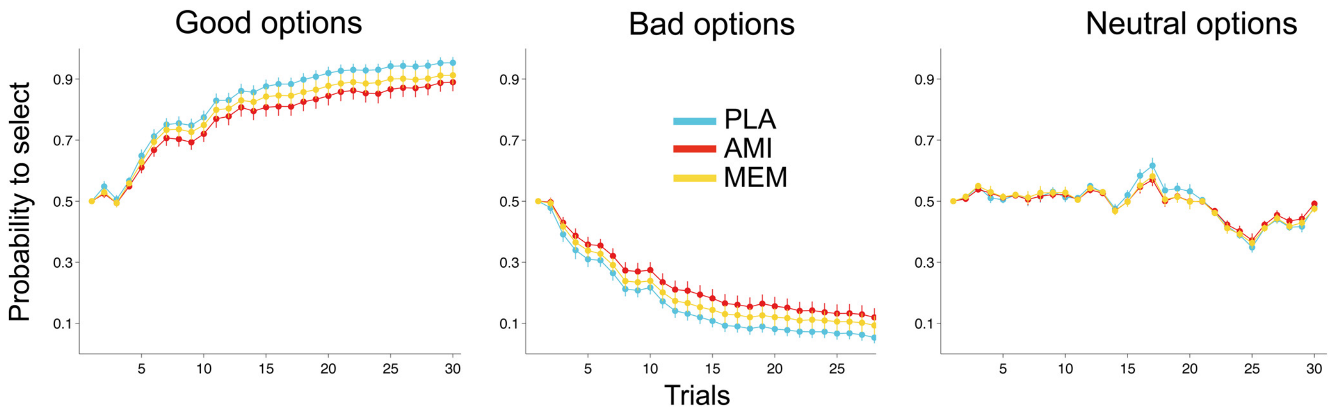

Figure 2. Learning curves derived from the reinforcement learning model. Model choice probabilities to select the option presented to it were computed using the learning rate and softmax temperature fitted for each subject and the sequence of options presented and outcomes observed by the model. Values are mean ± SEM.

ing between treatment conditions became evident. However, it appeared that learning curves diverged late in the blocks when learning had stabilized, with the plateau reached at the end of the block seeming to be higher under placebo than under either memantine or amisulpride. To find systematic breaks in the learning curves that allowed us to divide learning into early and final stages, we applied the Bai–Perron multiple break point test (see Materials and Methods). The test was applied to learning data across participants. Two breaks were identified for both approach and avoidance learning that split the learning curves into three stages. For approach learning, these breaks were detected at trials 10 and 19; for avoidance learning, breaks were detected at trials 6 and 15. In the final segment, approach learning was impaired by both amisulpride and memantine. In contrast, avoidance learning was impaired only by amisulpride, but not by memantine (Fig. 1). Two-way ANOVA on the averaged score during the final segment revealed an effect of VALENCE ($F_{(1,21)} = 18.78$, $p < 0.0001$), no effect of DRUG ($F_{(1,21)} = 2.237$, $p = 0.119$), but a marginally significant DRUG × VALENCE interaction ($F_{(2,42)} = 2.92$, $p = 0.065$), indicating that drugs affected learning and did so differently for approach versus avoidance learning. *Post hoc t* tests showed that final learning performance was reduced under amisulpride compared with placebo ($t_{(21)} = 2.1$, $p = 0.024$) and under memantine compared with placebo, although this effect fell just short of statistical significance ($t_{(21)} = 1.63$, $p = 0.059$). In contrast to approach learning, avoidance learning was impaired only by amisulpride ($t_{(21)} = 2.65$, $p = 0.0075$), but not by memantine ($t_{(21)} = 0.49$, $p > 0.3$). The effects of amisulpride on avoidance learning appeared more pronounced than those on approach learning because a significant learning impairment compared with placebo became evident even when considering the summed score over the entire learning curve ($t_{(21)} = 2.56$, $p = 0.009$).

The behavioral effects of these drugs occurred in the absence of any effects on response speed, measures of attention, subjective mood, heart rate, and blood pressure. Neither response time nor the number of incorrect responses on the forced-choice reaction-time task showed an effect of DRUG ($p > 0.44$). Likewise, time taken to complete the Trail Making Test Part B, a measure of visual attention, was not influenced by DRUG ($p > 0.36$), nor was there an effect of DRUG on subjective mood rating ($p > 0.89$). Finally, neither heart rate nor systolic or diastolic blood pressure was influenced by DRUG ($p > 0.37$).

Neither the learning rates [$\alpha = 0.0175 \pm 0.0042$, 0.0266 ± 0.0067, and 0.0221 ± 0.0061 (mean ± SEM) for placebo, amisulpride, and memantine, respectively] nor the temperatures ($\tau = 0.0533 \pm 0.0194$, 0.1328 ± 0.0446, and 0.1131 ± 0.0432) estimated for the reinforcement learning model differed as a function of DRUG ($p > 0.26$). However, there was a trend for an effect of DRUG on the model fits (negative log likelihoods, $p = 0.094$). When we followed this up by *post hoc t* tests, we found that the model fit was worse under amisulpride compared with placebo ($p = 0.0052$), but not under memantine ($p > 0.5$) compared with placebo [negative log likelihoods: 126.23 ± 5.73, 140.41 ± 6.5, and 130.58 ± 6.6 (mean ± SEM), for placebo, amisulpride, and memantine, respectively]. The model's probability for generating the choices made by participants was significantly above chance for good [0.75 ± 0.015 (mean ± SEM) model choice probability], bad (0.75 ± 0.017), and neutral symbols (0.52 ± 0.0037, all $t_{(65)} > 43$, $p < 10 * 10^{-37}$). The model was able to accurately reproduce subjects' choice behavior (Fig. 2). Again, two-way ANOVA on the averaged score during the final segment revealed an effect of VALENCE ($F_{(1,21)} = 403.58$, $p < 0.0001$), no effect of DRUG ($F_{(1,21)} = 0.252$, $p = 0.778$), but a marginally significant DRUG × VALENCE interaction ($F_{(2,42)} = 3.19$, $p = 0.051$). Likewise, *post hoc t* tests showed that the modeled final learning performance was reduced under amisulpride compared with placebo ($t_{(21)} = 3.17$, $p = 0.0023$) and under memantine compared with placebo, although this effect fell just short of statistical significance ($t_{(21)} = 1.66$, $p = 0.056$). In contrast, the model's avoidance learning was impaired only by amisulpride ($t_{(21)} = 3.16$, $p = 0.0024$), but not by memantine ($t_{(21)} = 1.53$, $p = 0.07$).

BOLD correlates of reward PEs

When we first performed our fMRI analyses using a simple stimulus value learning method, we made the intriguing observation that the correlation of the fMRI signal with the model-derived PEs flipped signs between approach and avoid trials (see Materials and Methods). We show this in Figure 3 using data from our ROI in the putamen, but the same results are obtained when using the other two striatal ROIs. This figure shows the PEs across all subjects in all three drug treatments, split up into trials where the subjects selected the option or where they avoided it (Fig. 3, left). The positive PE correlate is a result of both a negative correlation with the expected value and a positive correlation with the outcome. When decomposing the PEs into these constituent terms, we find that both effects flip signs. On approach trials, the usual pattern is observed (positive correlation with outcome and negative correlation with expected value; Fig. 3, middle). On

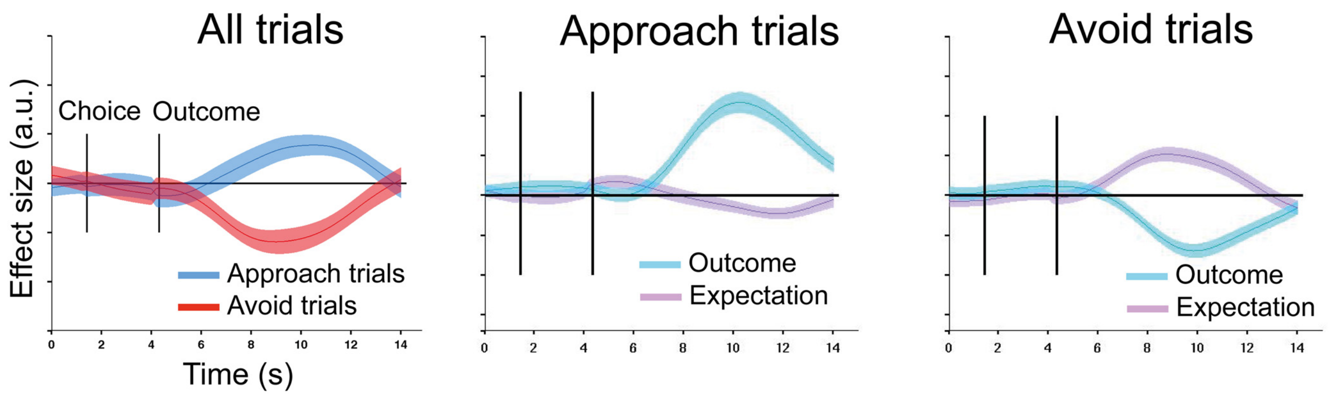

Figure 3. Time course over the course of a trial of the correlation of the BOLD signal from an ROI in the putamen with the PE derived from a standard Rescorla–Wagner model. Left, PE correlates positively with putaminal BOLD signal on approach trials. On avoid trials, however, the effect reverses to a negative correlation between PE and BOLD signal. When further decomposing the PE into its component terms, expected value and outcome, separately for approach (middle) and avoid (right) trials, it shows that both the effects of outcome and expectation reverse signs on the avoid trials. Values represent regression coefficients obtained from multiple linear regression; solid lines are means of all subjects collapsed across the three conditions. Shaded areas represent the SEM.

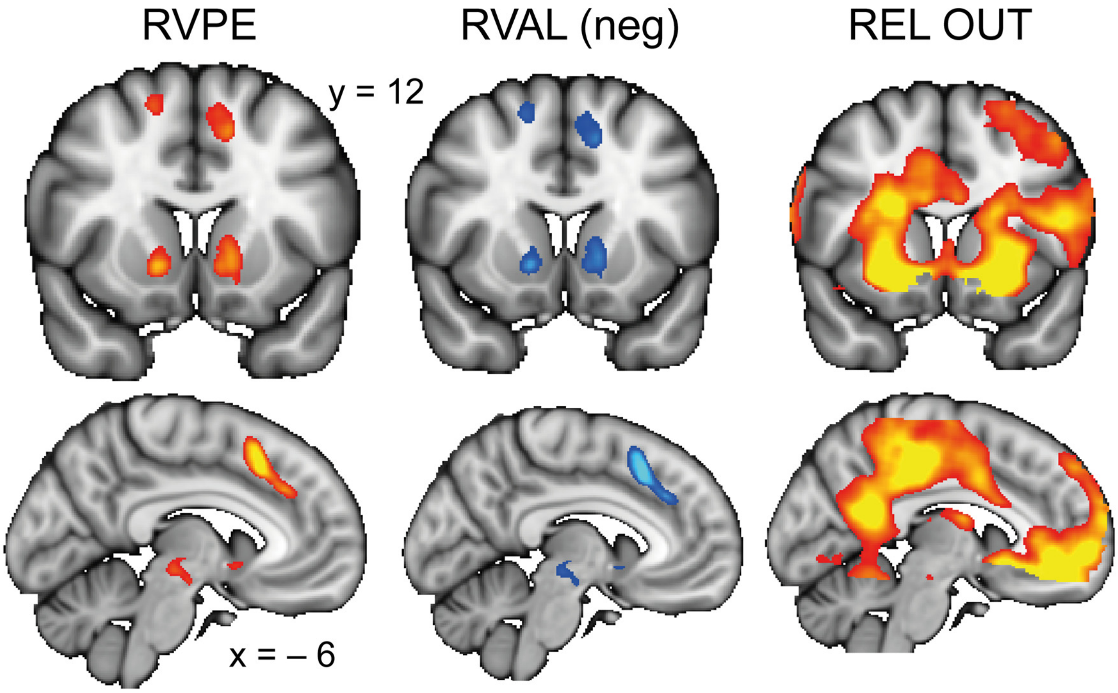

Figure 4. Whole-brain effects of the relative value PE at outcome time. Left, Effect of the contrast relative value PE (relative outcome minus relative value). Middle, Right, Effects of the component terms relative value (middle) and relative outcome (right) show that the positive effect of relative value PE in the striatum (left) is a result of both a negative effect of relative value and a pronounced positive effect of relative outcome (right). In the aMCC, the relative value PE effect seems to arise primarily from a negative effect of relative value, without a clear positive effect of relative outcome in the same region. Note, however, the strong positive relative outcome effects in adjacent mesial cortical areas. Images are thresholded at $Z > 3.1$.

avoid trials, however, both effects reverse in the opposite direction (Fig. 3, right). This might be suggestive of a PE that is used to update the relative, rather than absolute value of actions. We implemented a relative value learning algorithm that learns relative values simply by substituting the relative outcome (outcome of the chosen action minus outcome of the alternative action) for the experienced reward. This has interesting implications for studies with two (or more) separate options with uncorrelated outcomes and new choice contexts. For the purpose of our current study, this procedure simply serves to bring PEs from approach and avoid trials into the same reference frame so they can be analyzed together. The resulting model choice probabilities of the two algorithms are (by design) identical. Our following results therefore report the effects of relative value PEs, or their component terms, relative values, and relative outcomes.

Whole-brain correlates of relative value PEs

We found activity covering ventral parts of the striatum (ventral caudate, ventral putamen and ventral striatum, peaks at MNI $x = -14$, $y = 7$, $z = -4$, Z-max $= 4.77$, and $x = 15$, $y = 15$, $z = -3$, Z-max $= 4.7$) and in the anterior midcingulate cortex (aMCC; MNI: $x = -7$, $y = 20$, $z = 43$, Z-max $= 5.47$, and $x = 12$, $y = 27$, $z = 39$, Z-max $= 5.2$) to correlate with the relative value PE (Fig. 4, left). When looking at the component terms (relative value and relative outcome), we found both a strong negative effect of relative value (MNI: $x = -14$, $y = 7$, $z = -4$, Z-max $= -4.65$, and $x = 15$, $y = 15$, $z = -3$, Z-max $= -4.5$) and a strong positive effect of relative outcome (MNI: $x = -19$, $y = 9$, $z = -11$, Z-max $= 8.84$, and $x = 20$, $y = 9$, $z = -10$, Z-max $= 8.6$) in the striatum (Fig. 4, middle, right). Interestingly, in the aMCC we only found a strong negative effect of relative value (MNI: $x =$

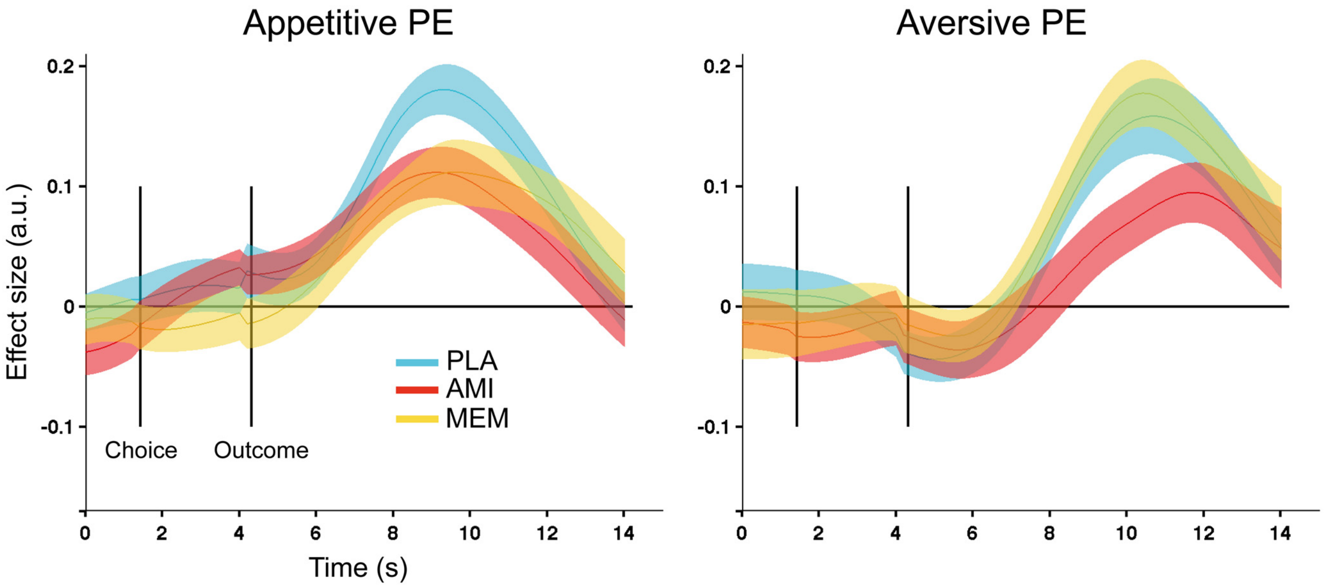

Figure 5. Drug effects on striatal PE representations. Here, the striatal ROIs (caudate, putamen, and ventral striatum) were merged into one conglomerate striatal ROI. Time courses represent the correlation of the effect size differences of relative outcome minus relative value on the BOLD signal from this ROI. Amisulpride diminished both appetitive (left) and aversive (right) PEs, while memantine only attenuated appetitive, but not aversive PEs. Values represent regression coefficients obtained from multiple linear regression; solid lines are group means for each drug condition. Shaded areas represent the SEM. PLA, Placebo; AMI, amisulpride; MEM, memantine.

-7, $y = 20$, $z = 43$, Z-max $= -5.47$, and $x = 12$, $y = 27$, $z = 39$, Z-max $= -5.2$), but no positive effect of relative outcome.

Pharmacological effects on striatal PE coding

Using the procedure above, we were able to obtain data from three ROIs that are not subject to any selection bias. Those ROIs lie in the caudate, putamen, and ventral striatum. In addition, an overall striatal ROI was created by merging the three subterritories. We divided PEs up into appetitive and aversive trials, because the availability of a reward (appetitive trial), experienced or observed, will promote approach behavior, whereas the availability of a punishment (aversive trial), experienced or observed, will promote avoidance behavior. We find that both amisulpride and memantine diminished appetitive PEs. In contrast, aversive PEs were only reduced under amisulpride, while memantine had no effect. Two-way ANOVA on the whole-striatal PEs yielded an effect of DRUG ($F_{(2,42)} = 3.8$, $p = 0.03$), but no effect of VALENCE ($F_{(1,21)} = 0.016$, $p = 09$), and no VALENCE × DRUG interaction ($F_{(2,42)} = 1.96$, $p = 0.15$). Again, *post hoc* tests showed that amisulpride diminished both the appetitive ($t_{(21)} = 2.23$, $p = 0.018$) and aversive PEs ($t_{(21)} = 1.852$, $p = 0.039$). In contrast, memantine only reduced the appetitive PEs ($t_{(21)} = 2.38$, $p = 0.014$), but had no influence on the aversive PEs ($t_{(21)} = 0.44$, $p = 0.33$; Fig. 5). To investigate this effect in more detail, we used a design matrix (see Materials and Methods) that decomposed the PE into its component terms, relative value, and relative outcome, separately for approach and avoid trials and applied this analysis separately to each of the three striatal subterritories. We find the expected negative effect of relative value following presentation of the outcome. However, this representation was not affected by either of the drugs. Two-way ANOVA revealed neither an effect of DRUG nor a DRUG × VALENCE interaction in any of the three striatal ROIs (Fig. 6; all $p > 0.2$). In contrast to this, for relative outcome (Fig. 7), two-way ANOVA in the caudate yielded an effect of DRUG ($F_{(2,42)} = 4.34$, $p = 0.0194$), VALENCE ($F_{(1,21)} = 11.92$, $p = 0.0031$), but no VALENCE × DRUG interaction ($F_{(2,42)} = 1.04$, $p = 0.36$). *Post hoc* tests showed that under amisulpride, both the appetitive ($t_{(21)} = 2.32$, $p = 0.015$) and aversive ($t_{(21)} = 2.02$, $p = 0.0277$) outcome signal was reduced compared with placebo. Under memantine, only the relative outcome signal on appetitive trials ($t_{(21)} = 2.2$, $p = 0.0195$), but not on aversive trials ($p > 0.46$), was attenuated compared with placebo. In the putamen, two-way ANOVA yielded an effect of DRUG ($F_{(2,42)} = 3.586$, $p = 0.0365$), but no effect of VALENCE ($F_{(1,21)} = 2.138$, $p = 0.1585$), and no VALENCE × DRUG interaction ($F_{(2,42)} = 1.608$, $p = 0.2123$). As in the caudate, *post hoc* tests again showed that under amisulpride, the relative outcome effect in both appetitive ($t_{(21)} = 1.99$, $p = 0.03$) and aversive trials ($t_{(21)} = 1.87$, $p = 0.0375$) was reduced compared with placebo. Likewise, under memantine, again only the appetitive ($t_{(21)} = 2.14$, $p = 0.022$) but not the aversive ($p > 0.4$) outcome signal was attenuated compared with placebo. In the ventral striatum, two-way ANOVA yielded a trend for an effect of DRUG ($F_{(2,42)} = 2.825$, $p = 0.07$), an effect of VALENCE ($F_{(1,21)} = 5.489$, $p = 0.0291$), and a trend for a VALENCE × DRUG interaction ($F_{(2,42)} = 2.73$, $p = 0.077$). Despite the effect of drug being only present as a trend, we followed this up with our preplanned comparisons because of our strong a priori hypothesis. We found that both amisulpride ($t_{(21)} = 3.17$, $p = 0.0023$) and memantine ($t_{(21)} = 2.7$, $p = 0.0067$) treatment decreased the outcome signal on appetitive trials, thus mirroring the pattern in the other two striatal ROIs. In contrast to this, the outcome signal on aversive trials in the ventral striatum remained unaffected by either drug (both $p > 0.16$). Together, by showing that the striatal signal shows both the positive correlation with relative outcome and the negative correlation with relative value at the time of outcome, we demonstrate that the striatal signal fulfils the required formal criteria for a PE signal. We further demonstrated that both drugs modulated the effects of the (relative) outcome term and did not affect representations of expected (relative) value. Finally, to test whether the drugs differentially affected relative outcome representations in the three striatal regions, we performed a three-way ANOVA with the factors ROI (caudate, putamen, ventral striatum), DRUG (placebo, amisul-

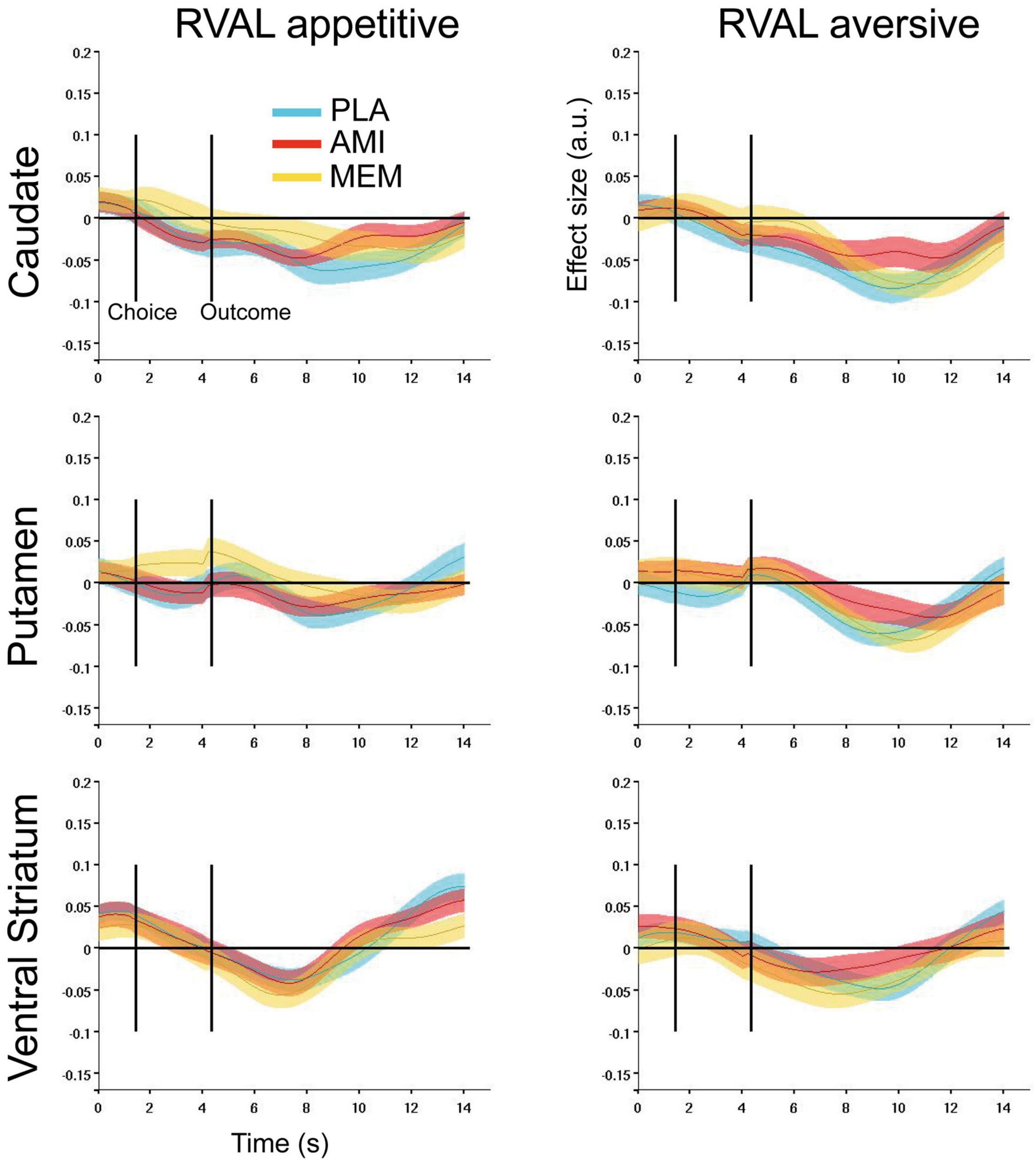

Figure 6. Time course over the course of a trial of the correlation of relative value (RVAL) with the BOLD signal in three independently defined striatal ROIs in the caudate (top), putamen (middle), and ventral striatum (bottom), separately for appetive and aversive trials (trials in which reward or punishment, respectively, were available, regardless of whether or not this outcome had been experienced or only observed). Values represent regression coefficients obtained from multiple linear regression; solid lines are group means for each drug condition. Shaded areas represent the SEM. PLA, Placebo; AMI, amisulpride; MEM, memantine.

pride, memantine), and VALENCE (appetitive, aversive). We found main effects of DRUG ($F_{(2,42)} = 3.97$, $p = 0.026$), VALENCE ($F_{(1,21)} = 7.647$, $p = 0.012$), and ROI ($F_{(2,42)} = 16.131$, $p < 0.001$), but no DRUG × ROI nor a DRUG × ROI × VALENCE interaction ($p > 0.45$), suggesting that drug effects did not differ between the three ROIs. When looking at the pattern of the effects (Fig. 7, bottom right), we note that the aversive outcome correlate in the ventral striatum seems to have temporal characteristics that slightly deviate from the other signals. In particular, we observe an initial negative dip following the outcome, followed by a later positive peak. We do not know the reason for this, but we observe that under placebo, the negative dip seems slightly more pronounced than under drug. Thus, it might be possible that when considering the amplitude of this effect from base to peak, there may still be a significant difference between conditions.

Discussion

In this study we combined fMRI with pharmacological antagonism of D2 and NMDA receptors to test their differential involvement in approach and avoidance learning and in the expression of appetitive and aversive PEs. We found that D2 antagonism with amisulpride impaired both approach and avoidance learning concomitant with blunting of both appetitive and aversive PE signals in subregions of the striatum. In contrast, memantine mildly attenuated approach learning without affecting avoidance learning, which was paralleled by a reduction of appetitive, but not aversive PE signals in striatal subregions.

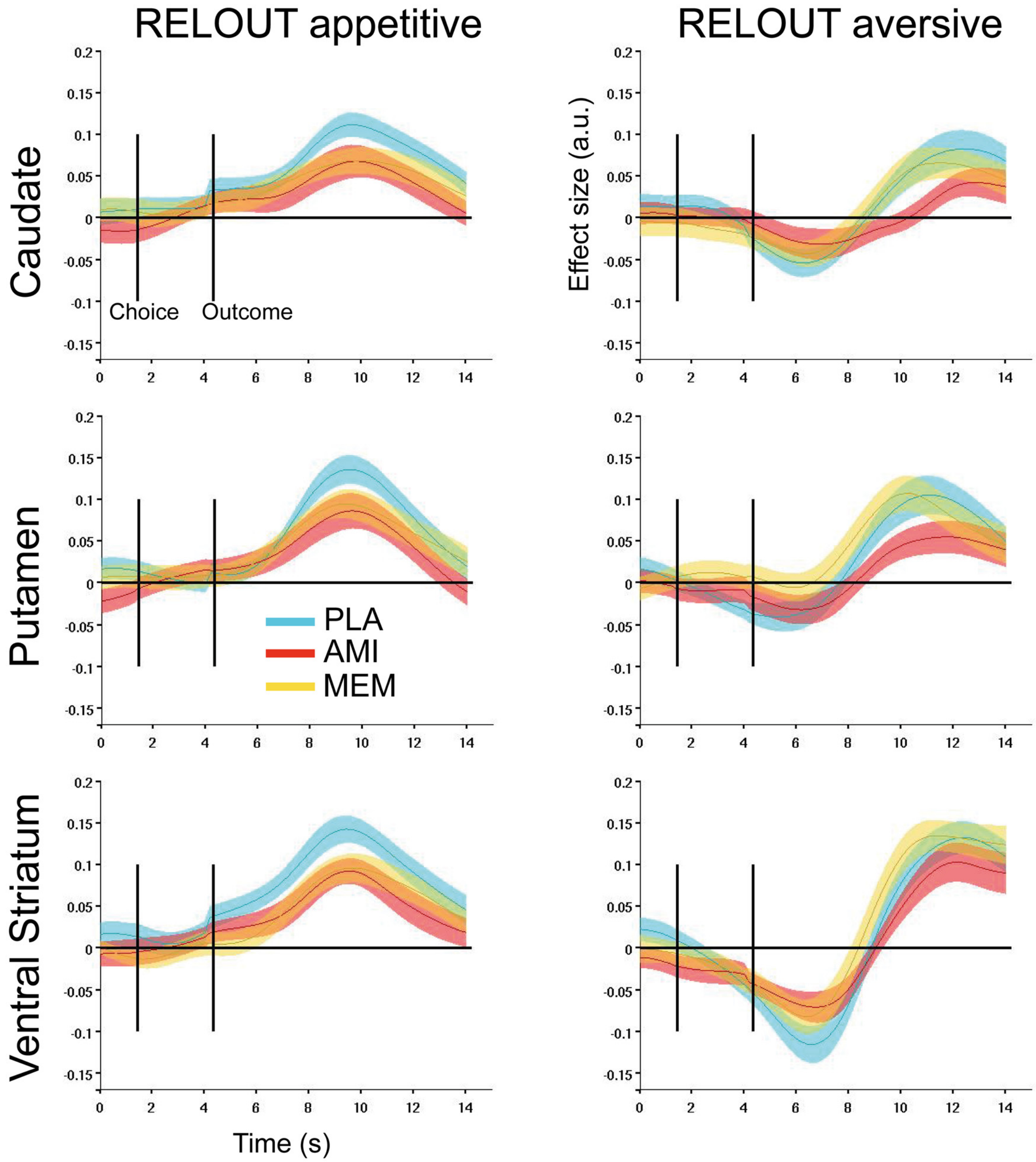

Figure 7. Same as in Figure 6, but here the effect of relative outcome is shown (obtained from the same GLM as the relative value effects in Fig. 6). PLA, Placebo; AMI, amisulpride; MEM, memantine. The representation of outcomes was attenuated by both amisulpride and memantine on appetitive trials, but only by amisulpride on aversive trials.

There are ≥3 candidate mechanisms that could explain the attenuation of appetitive learning after NMDA receptor antagonism. Our working hypothesis was built on the evidence that phasic activation of midbrain DA neurons is driven by glutamatergic afferents acting via NMDA receptors localized on DA neurons in the midbrain (Overton and Clark, 1992; Wang et al., 2011). However, DA release in the striatum is also enhanced by local NMDA receptors, possibly localized directly on the dopaminergic terminals (Krebs et al., 1991; Ohta et al., 1994; Iravani and Kruk, 1996; Chéramy et al., 1998). This mechanism could also account for the diminished appetitive PEs we observed in the striatum. In addition, direct infusion of NMDA receptor antagonists into the ventral striatum impairs appetitive instrumental learning (Kelley et al., 1997). This effect might operate at the postsynaptic level, independently of modulation of DA release. The effects of NMDA receptors and DA converge on the same signaling pathways (Konradi et al., 1996) and combined administration of NMDA and DA D1 receptor antagonists at doses that are ineffective when given alone impair appetitive instrumental conditioning (Smith-Roe and Kelley, 2000). Finally, it has to be noted that NMDA receptors are widely expressed throughout the brain, in particular in neocortex. Thus, given that we used systemic drug administration, an action at extrastriatal sites cannot be ruled out.

In contrast to NMDA receptors, D2 receptors are required for detecting both phasic increases and decreases in striatal DA levels. Our finding that amisulpride interfered with both approach and avoidance learning is in agreement with rodent studies showing impairments in appetitively and aversively motivated instrumental learning following blockade of striatal D2 receptors (Salamone, 2002; Wise, 2004; Salamone et al., 2005). The evidence in humans is more mixed (for review, see Ullsperger et al., 2014).

Studies using the paradigm developed by Frank and colleagues (Frank et al., 2004) found that Parkinson's patients (who suffer from DA depletion in the dorsal striatum) are biased to better learning from negative outcomes, which is reversed under treatment with L-DOPA (Frank et al., 2004, 2007). Using the same task in healthy volunteers, the D2 receptor antagonists haloperidol or amisulpride (Frank and O'Reilly, 2006; Jocham et al., 2011) biased learning toward better learning from positive outcomes, whereas the reverse bias was found under the influence of the D2 agonist cabergoline (Frank and O'Reilly, 2006). These effects are remarkably consistent with a mechanistic model of basal ganglia DA function that relates approach and avoidance learning to differential DA effects on synaptic plasticity in the direct and indirect pathway of the striatum (Frank, 2005). It has to be borne in mind, however, that the task used in these studies consists of an initial learning stage and a later transfer phase that requires participants to make choices on the basis of the previously learned values. None of the above studies report any effects on the initial learning. The effects are found when choices between novel pairings are made, in the absence of any feedback. There is evidence that DA effects in this task may, at least in part, be mediated by affecting instrumental performance during the transfer phase, rather than, or in addition to, an effect on learning (Jocham et al., 2011; Shiner et al., 2012). In a task using an asymmetric reward schedule, pramipexole reduced reward learning (Pizzagalli et al., 2008). This task, however, did not capture effects on punishment learning, and there was evidence for nonspecific effects (sedation) of pramipexole. Using a subliminal instrumental learning task, Palminteri and colleagues found that unmedicated Tourette's patients (likely suffering from a hyperdopaminergic state) were better at appetitively motivated learning than at aversively motivated learning. This bias was reversed under treatment with neuroleptic drugs (risperidone, pimozide), agents that primarily target D2 receptors. The reverse pattern was found in Parkinson's patients, who showed a bias to better punishment learning in the nonmedicated state, which was reversed in the medicated state (Palminteri et al., 2009). The same group also found that haloperidol attenuated approach learning and appetitive striatal PEs relative to L-DOPA, without affecting avoidance learning and aversive striatal PEs in healthy humans (Pessiglione et al., 2006). We have previously found improved learning from rewarding outcomes and enhanced striatal PE coding under a lower dose of amisulpride (200 mg; Jocham et al., 2011). This dose was chosen to primarily achieve a blockade of somatodendritic autoreceptors, thereby possibly facilitating phasic DA neuron firing and subsequent DA release during positive PEs. In contrast, the dose in the current study (400 mg) was chosen on the basis of studies showing that this dose is sufficient to achieve occupation of ∼50–80% of postsynaptic striatal D2 receptors (Martinot et al., 1996; Bressan et al., 2004; la Fougère et al., 2005; Meisenzahl et al., 2008), thereby preventing the detection of either increases or decreases in DA levels by striatal D2 receptors.

The drug effects on reinforcement learning we observed were rather modest. One possible reason for this is that learning in our task was not exclusively reliant on striatal mechanisms. It is possible that our subjects deployed additional, possibly more declarative strategies to solve the task. Our task was similar to the paradigm used by Pessiglione and colleagues (Pessiglione et al., 2008; Palminteri et al., 2009). In their studies, due to the subliminal presentation of stimuli, performance never reached the same plateau as in our present findings. It is possible that such learning, which possibly relies predominantly on the striatum, would have been easier to perturb with pharmacological challenges. Regarding the choice of our NMDA antagonist, there are drugs that are more potent at blocking NMDA receptors, e.g., ketamine. It is possible that NMDA receptor antagonism with ketamine would have revealed more pronounced effects. However, due to the general role of NMDA receptors in cognition (Robbins and Murphy, 2006), one might easily obtain diffuse cognitive impairments under highly potent NMDA receptor antagonism. A third possibility is that the high-amplitude PEs that occur early in the course of learning are still able to override the pharmacological effects (at least in the case of phasic DA activation), whereas the smaller PEs during the final stages as learning stabilizes may be more vulnerable to our drug challenge. Finally, an alternative explanation of our findings is that drugs did not affect reinforcement learning at all, but instead reflected effects on instrumental performance. At least in the case of DA, there is clear evidence for drugs also affecting instrumental performance (Jocham et al., 2011; Shiner et al., 2012; Eisenegger et al., 2014). Future studies might employ protocols that run either a learning phase in the drugged state or a later probe trial phase in the drugged state (with learning in the drug-free state). However, while we cannot exclude a performance effect, we also note that the behavioral effects are precisely mirrored by the drug effects on striatal representations of reward PEs.

The fit of our reinforcement learning model to subjects' behavior was poorer under amisulpride compared with placebo. One might argue that our finding of reduced appetitive PE coding in the striatum under amisulpride might be a reflection a less accurate estimation of subjects' PEs in this drug treatment. While this is possible, we have decomposed the PEs into its component terms, relative values, and relative outcomes. We find that the drug effects on striatal PEs arise from a modulation of the effect of relative outcomes, not relative values. While relative value is a parameter estimated by the model, relative outcome is simply the difference between the experienced outcome and the outcome that would have resulted from the alternative course of action. Therefore, our results are immune to differences in model fits between drug treatments.

We made the entirely unpredicted observation that our PEs derived from a standard Rescorla–Wagner model showed the expected positive correlation with the striatal fMRI signal on approach trials, but reversed to a negative correlation on avoid trials. This pattern is not expected from standard reinforcement learning models and may provide a hint that this signal instead codes a PE that is used to learning a relative value, that is, how good an action is relative to its alternatives. We are currently exploring this possibility further. Here, we can't provide further evidence in support of this idea because we do not have uncorrelated alternative outcomes. Other fMRI studies have already found evidence for so-called fictive PEs (difference between reward obtained and maximum possible reward) in the human striatum (Lohrenz et al., 2007; Chiu et al., 2008). Taking advantage of the millisecond resolution of EEG, Fischer and Ullsperger (2013) used a task almost identical to the one used here and showed that the spatiotemporal dynamics of real and fictive outcome processing differed during the initial 400 ms after outcome presentation, but then converged onto a common final pathway.

Our findings show that D2 receptor antagonism interfered with both approach and avoidance learning, concomitant with reduced expression of both appetitive and aversive PEs in the striatum. NMDA receptor antagonism attenuated approach learning and appetitive PE coding in the striatum, without affecting avoidance learning and aversive PEs in the striatum.

References

Bai JS, Perron P (1998) Estimating and testing linear models with multiple structural changes. Econometrica 66:47–78. CrossRef

Bayer HM, Glimcher PW (2005) Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron 47:129–141. CrossRef Medline

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. CrossRef Medline

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

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

Bressan RA, Erlandsson K, Spencer EP, Ell PJ, Pilowsky LS (2004) Prolactinemia is uncoupled from central D2/D3 dopamine receptor occupancy in amisulpride treated patients. Psychopharmacology 175:367–373. CrossRef Medline

Bromberg-Martin ES, Matsumoto M, Hong S, Hikosaka O (2010) A pallidus-habenula-dopamine pathway signals inferred stimulus values. J Neurophysiol 104:1068–1076. CrossRef Medline

Caplin A, Dean M (2008) Axiomatic methods, dopamine and reward prediction error. Curr Opin Neurobiol 18:197–202. CrossRef Medline

Chéramy A, L'hirondel M, Godeheu G, Artaud F, Glowinski J (1998) Direct and indirect presynaptic control of dopamine release by excitatory amino acids. Amino Acids 14:63–68. CrossRef Medline

Chiu PH, Lohrenz TM, Montague PR (2008) Smokers' brains compute, but ignore, a fictive error signal in a sequential investment task. Nat Neurosci 11:514–520. CrossRef Medline

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. CrossRef Medline

D'Ardenne K, McClure SM, Nystrom LE, Cohen JD (2008) BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. Science 319:1264–1267. CrossRef Medline

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. CrossRef Medline

Eisenegger C, Naef M, Linssen A, Clark L, Gandamaneni PK, Müller U, Robbins TW (2014) Role of dopamine D2 receptors in human reinforcement learning. Neuropsychopharmacology 39:2366–2375. CrossRef Medline

Fischer AG, Ullsperger M (2013) Real and fictive outcomes are processed differently but converge on a common adaptive mechanism. Neuron 79:1243–1255. CrossRef Medline

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. CrossRef Medline

Frank MJ, O'Reilly RC (2006) A mechanistic account of striatal dopamine function in human cognition: psychopharmacological studies with cabergoline and haloperidol. Behav Neurosci 120:497–517. CrossRef Medline

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

Frank MJ, Samanta J, Moustafa AA, Sherman SJ (2007) Hold your horses: impulsivity, deep brain stimulation, and medication in parkinsonism. Science 318:1309–1312. CrossRef Medline

Hollerman JR, Schultz W (1998) Dopamine neurons report an error in the temporal prediction of reward during learning. Nat Neurosci 1:304–309. CrossRef Medline

Iravani MM, Kruk ZL (1996) Real-time effects of N-methyl-D-aspartic acid on dopamine release in slices of rat caudate putamen: a study using fast cyclic voltammetry. J Neurochem 66:1076–1085. Medline

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

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

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. CrossRef Medline

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. CrossRef Medline

Jocham G, Klein TA, Ullsperger M (2011) Dopamine-mediated reinforcement learning signals in the striatum and ventromedial prefrontal cortex underlie value-based choices. J Neurosci 31:1606–1613. CrossRef Medline

Kelley AE, Smith-Roe SL, Holahan MR (1997) Response-reinforcement learning is dependent on N-methyl-D-aspartate receptor activation in the nucleus accumbens core. Proc Natl Acad Sci U S A 94:12174–12179. CrossRef Medline

Klein-Flügge MC, Hunt LT, Bach DR, Dolan RJ, Behrens TE (2011) Dissociable reward and timing signals in human midbrain and ventral striatum. Neuron 72:654–664. CrossRef Medline

Konradi C, Leveque JC, Hyman SE (1996) Amphetamine and dopamine-induced immediate early gene expression in striatal neurons depends on postsynaptic NMDA receptors and calcium. J Neurosci 16:4231–4239. Medline

Krebs MO, Desce JM, Kemel ML, Gauchy C, Godeheu G, Cheramy A, Glowinski J (1991) Glutamatergic control of dopamine release in the rat striatum: evidence for presynaptic N-methyl-D-aspartate receptors on dopaminergic nerve terminals. J Neurochem 56:81–85. CrossRef Medline

la Fougère C, Meisenzahl E, Schmitt G, Stauss J, Frodl T, Tatsch K, Hahn K, Moller HJ, Dresel S (2005) D2 receptor occupancy during high- and low-dose therapy with the atypical antipsychotic amisulpride: a 123I-iodobenzamide SPECT study. J Nucl Med 46:1028–1033. Medline

Lohrenz T, McCabe K, Camerer CF, Montague PR (2007) Neural signature of fictive learning signals in a sequential investment task. Proc Natl Acad Sci U S A 104:9493–9498. CrossRef Medline

Martinot JL, Paillère-Martinot ML, Poirier MF, Dao-Castellana MH, Loc'h C, Mazière B (1996) In vivo characteristics of dopamine D2 receptor occupancy by amisulpride in schizophrenia. Psychopharmacology 124:154–158. CrossRef Medline

Matsumoto M, Hikosaka O (2007) Lateral habenula as a source of negative reward signals in dopamine neurons. Nature 447:1111–1115. CrossRef Medline

Matsumoto M, Hikosaka O (2009) Representation of negative motivational value in the primate lateral habenula. Nat Neurosci 12:77–84. CrossRef Medline

Meisenzahl EM, Schmitt G, Gründer G, Dresel S, Frodl T, la Fougère C, Scheuerecker J, Schwarz M, Boerner R, Stauss J, Hahn K, Möller HJ (2008) Striatal D2/D3 receptor occupancy, clinical response and side effects with amisulpride: an iodine-123-iodobenzamide SPET study. Pharmacopsychiatry 41:169–175. CrossRef Medline

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. CrossRef Medline

Ohta K, Araki N, Shibata M, Komatsumoto S, Shimazu K, Fukuuchi Y (1994) Presynaptic ionotropic glutamate receptors modulate in vivo release and metabolism of striatal dopamine, noradrenaline, and 5-hydroxytryptamine: involvement of both NMDA and AMPA/kainate subtypes. Neurosci Res 21:83–89. CrossRef Medline

Overton P, Clark D (1992) Iontophoretically administered drugs acting at the N-methyl-D-aspartate receptor modulate burst firing in A9 dopamine neurons in the rat. Synapse 10:131–140. CrossRef Medline

Palminteri S, Lebreton M, Worbe Y, Grabli D, Hartmann A, Pessiglione M (2009) Pharmacological modulation of subliminal learning in Parkinson's and Tourette's syndromes. Proc Natl Acad Sci U S A 106:19179–19184. CrossRef Medline

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. CrossRef Medline

Pessiglione M, Petrovic P, Daunizeau J, Palminteri S, Dolan RJ, Frith CD (2008) Subliminal instrumental conditioning demonstrated in the human brain. Neuron 59:561–567. CrossRef Medline

Pizzagalli DA, Evins AE, Schetter EC, Frank MJ, Pajtas PE, Santesso DL, Culhane M (2008) Single dose of a dopamine agonist impairs reinforcement learning in humans: behavioral evidence from a laboratory-based measure of reward responsiveness. Psychopharmacology 196:221–232. CrossRef Medline

Rescorla RA, Wagner AR (1972) A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In:

Classical conditioning II: current research and theory (Black AH, Prokasy WF, eds), pp 64–99. New York: Appleton Century Crofts.

Robbins TW, Murphy ER (2006) Behavioural pharmacology: 40+ years of progress, with a focus on glutamate receptors and cognition. Trends Pharmacol Sci 27:141–148. CrossRef Medline

Roesch MR, Calu DJ, Schoenbaum G (2007) Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. Nat Neurosci 10:1615–1624. CrossRef Medline

Rutledge RB, Dean M, Caplin A, Glimcher PW (2010) Testing the reward prediction error hypothesis with an axiomatic model. J Neurosci 30:13525–13536. CrossRef Medline

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

Salamone JD, Correa M, Mingote SM, Weber SM (2005) Beyond the reward hypothesis: alternative functions of nucleus accumbens dopamine. Curr Opin Pharmacol 5:34–41. CrossRef Medline

Satoh T, Nakai S, Sato T, Kimura M (2003) Correlated coding of motivation and outcome of decision by dopamine neurons. J Neurosci 23:9913–9923. Medline

Schönberg T, Daw ND, Joel D, O'Doherty JP (2007) Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. J Neurosci 27:12860–12867. CrossRef Medline

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

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

Shiner T, Seymour B, Wunderlich K, Hill C, Bhatia KP, Dayan P, Dolan RJ (2012) Dopamine and performance in a reinforcement learning task: evidence from Parkinson's disease. Brain 135:1871–1883. CrossRef Medline

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. Medline

Smith-Roe SL, Kelley AE (2000) Coincident activation of NMDA and dopamine D1 receptors within the nucleus accumbens core is required for appetitive instrumental learning. J Neurosci 20:7737–7742. Medline

Steinberg EE, Keiflin R, Boivin JR, Witten IB, Deisseroth K, Janak PH (2013) A causal link between prediction errors, dopamine neurons and learning. Nat Neurosci 16:966–973. CrossRef Medline

Sutton RS, Barto AG (1998) Reinforcement learning: an introduction. Cambridge, MA: MIT.

Takahashi YK, Roesch MR, Stalnaker TA, Haney RZ, Calu DJ, Taylor AR, Burke KA, Schoenbaum G (2009) The orbitofrontal cortex and ventral tegmental area are necessary for learning from unexpected outcomes. Neuron 62:269–280. CrossRef Medline

Ullsperger M, von Cramon DY (2003) Error monitoring using external feedback: specific roles of the habenular complex, the reward system, and the cingulate motor area revealed by functional magnetic resonance imaging. J Neurosci 23:4308–4314. Medline

Ullsperger M, Danielmeier C, Jocham G (2014) Neurophysiology of performance monitoring and adaptive behavior. Physiol Rev 94:35–79. CrossRef Medline

Valentin VV, O'Doherty JP (2009) Overlapping prediction errors in dorsal striatum during instrumental learning with juice and money reward in the human brain. J Neurophysiol 102:3384–3391. CrossRef Medline

Wang LP, Li F, Wang D, Xie K, Wang D, Shen X, Tsien JZ (2011) NMDA receptors in dopaminergic neurons are crucial for habit learning. Neuron 72:1055–1066. CrossRef Medline

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

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

Zaghloul KA, Blanco JA, Weidemann CT, McGill K, Jaggi JL, Baltuch GH, Kahana MJ (2009) Human substantia nigra neurons encode unexpected financial rewards. Science 323:1496–1499. CrossRef Medline