NeuroImage 186 (2019) 113–125

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

L-DOPA reduces model-free control of behavior by attenuating the transfer of value to action

Nils B. Kroemer [a,b], Ying Lee [a], Shakoor Pooseh [a,c], Ben Eppinger [d,e], Thomas Goschke [d], Michael N. Smolka [a,*]

[a] Department of Psychiatry and Neuroimaging Center, Technische Universität Dresden, Dresden, Germany
[b] Department of Psychiatry and Psychotherapy, University of Tübingen, Tübingen, Germany
[c] Freiburg Center for Data Analysis and Modelling, Freiburg, Germany
[d] Department of Psychology, Technische Universität Dresden, Dresden, Germany
[e] Department of Psychology, Concordia University, Montreal, Canada

ARTICLE INFO

Keywords:
Reward
Dopamine
fMRI
Goal-directed behavior
Pharmacology
Computational modeling

ABSTRACT

Dopamine is a key neurotransmitter in action control. However, influential theories of dopamine function make conflicting predictions about the effect of boosting dopamine neurotransmission. Here, we tested if increases in dopamine tone by administration of L-DOPA upregulate reward learning as predicted by reinforcement learning theories, and if increases are specific for deliberative "model-based" control or reflexive "model-free" control. Alternatively, L-DOPA may impair learning as suggested by "value" or "thrift" theories of dopamine. To this end, we employed a two-stage Markov decision-task to investigate the effect of L-DOPA (randomized cross-over) on behavioral control while brain activation was measured using fMRI. L-DOPA led to attenuated model-free control of behavior as indicated by the reduced impact of reward on choice. Increased model-based control was only observed in participants with high working memory capacity. Furthermore, L-DOPA facilitated exploratory behavior, particularly after a stream of wins in the task. Correspondingly, in the brain, L-DOPA decreased the effect of reward at the outcome stage and when the next decision had to be made. Critically, reward-learning rates and prediction error signals were unaffected by L-DOPA, indicating that differences in behavior and brain response to reward were not driven by differences in learning. Taken together, our results suggest that L-DOPA reduces model-free control of behavior by attenuating the transfer of value to action. These findings provide support for the value and thrift accounts of dopamine and call for a refined integration of valuation and action signals in reinforcement learning models.

1. Introduction

Goal-directed actions require considerations along several dimensions. How valuable are the current options given the implied costs of action? Is it still worthwhile to pursue them if the rate of rewards decreases over time? Accordingly, dual-action choice models of reinforcement learning distinguish two basic modes of action control (Daw et al., 2005). On the one hand, "model-free" (MF) control appears as reflexive because it is based on direct reinforcement of successful actions via reward prediction errors (RPE), putatively encoded in mesocorticolimbic dopamine neurons (O'Doherty et al., 2007; Schultz et al., 1997). On the other hand, "model-based" (MB) control appears as deliberative because it is based on a learned model of the task structure, which allows agents to plan actions ahead of time by simulating their potential outcome (Daw et al., 2011). Similar to rewards, model states are learned via state prediction errors (Gläscher et al., 2010) and the learned structure of the task can be used to adjust reward expectations via MB control. Since MB control operates on "goal values", it has been associated with goal-directed behavior. In contrast, MF control operates on "habit values", which are divorced from the goal of reinforcement (Balleine et al., 2008). Whereas the corresponding computations recruit partly divergent brain networks (Smittenaar et al., 2013), the information is integrated in a common region: the nucleus accumbens (NAcc) (Daw et al., 2011). The NAcc acts as a limbic–motor interface and has been referred to as "nexus of goals" (Mannella et al., 2013). While there is mounting evidence each supporting the role of mesocorticolimbic

* Corresponding author. Department of Psychiatry and Psychotherapy, Technische Universität Dresden, Würzburger Str. 35, 01187, Dresden, Germany.
E-mail address: michael.smolka@tu-dresden.de (M.N. Smolka).

https://doi.org/10.1016/j.neuroimage.2018.10.075
Received 9 February 2018; Received in revised form 25 October 2018; Accepted 26 October 2018
Available online 28 October 2018

dopamine in the control of vigor (Beeler et al., 2012, 2016; Niv et al., 2007) and reinforcement learning (Schultz, 2013, 2015), little is known about how these two aspects go hand in hand to support adaptive control of behavior.

Recent work in animals has demonstrated that signals encoding demands for acting and learning are mixed in dopamine transmission (Hamid et al., 2016; Syed et al., 2016). Two distinct, yet interdependent aspects of signaling have been commonly differentiated: phasic dopamine and dopamine tone. Phasic dopamine appears to underlie acting and learning at different time points of a trial: it invigorates actions at the beginning whereas it reinforces actions at the end when the outcome is being presented (Hamid et al., 2016; Steinberg et al., 2013). The amplitude of phasic RPE signals is dependent on dopamine as increasing (via L-DOPA) versus reducing (via haloperidol) dopamine transmission leads to corresponding changes in RPE BOLD signals in humans (Pessiglione et al., 2006). In addition to phasic signaling, dopamine tone has been hypothesized to track the average reward rate (Niv et al., 2007) and plays an important role in facilitating response vigor (Beierholm et al., 2013; Rigoli et al., 2016a). Collectively, these observations raise a fundamental question of how different modes of action control and dopamine signaling are functionally integrated and balanced in a task-adaptive manner.

To address this question, we tested the effect of an exogenous increase in dopamine on the performance of a two-stage Markov task that has been designed to distinguish MF and MB control, including their RPE correlates (Fig. 1; Daw et al., 2011). As exogenous modulation of dopamine, we used the synthetic precursor L-DOPA, which increases presynaptic levels of dopamine and leads to an increase in tone (Harun et al., 2015). We focused a priori on the NAcc where both MF and MB RPE signals are commonly observed and explored potential effects of L-DOPA at a whole-brain level. Based on Wunderlich et al. (2012), we would expect an upregulation of deliberative MB control without changes in MF control.

With regard to actions, we would expect that boosting dopamine increases vigor (Beierholm et al., 2013; Rigoli et al., 2016a) and decreases thrift as organisms are more willing to expend energy (Beeler et al., 2012). One manifestation of increased vigor and/or decreased thrift is an increase in exploratory behavior. According to the thrift theory (Beeler, 2012), increased dopamine tone not only increases the willingness to expend energy; it also reduces the degree to which learned reward values affect behavioral choice. In other words, actions become more independent of expected values because a person no longer needs to exploit the best option. Likewise, the value theory (Hamid et al., 2016) proposes that heightened dopamine tone would reduce the local change in dopamine, thereby reducing the impact of phasic RPEs on actions (Collins and Frank, 2016; Grace, 1991). Whereas both of these theories predict that heighted dopamine increases exploration when expected values are similar, the thrift theory predicts that exploration should be facilitated even when there are larger differences in expected values because exploration is seen as less costly.

However, with regard to learning, the expected results are less straightforward as different theories make conflicting predictions (Collins and Frank, 2016). According to reinforcement learning theories, we would predict that boosting dopamine increases reinforcement learning (Pessiglione et al., 2006), at least for positive outcomes because high dopamine tone may impair negative outcome learning via D2 receptors (Collins and Frank, 2014; Frank et al., 2004). Notably, Wunderlich et al. (2012) did not observe changes in MF control, unlike what we would expect based on Pessiglione et al. (2006). Furthermore, given that both "thrift" (Beeler, 2012) and "value" theories (Hamid et al., 2016)

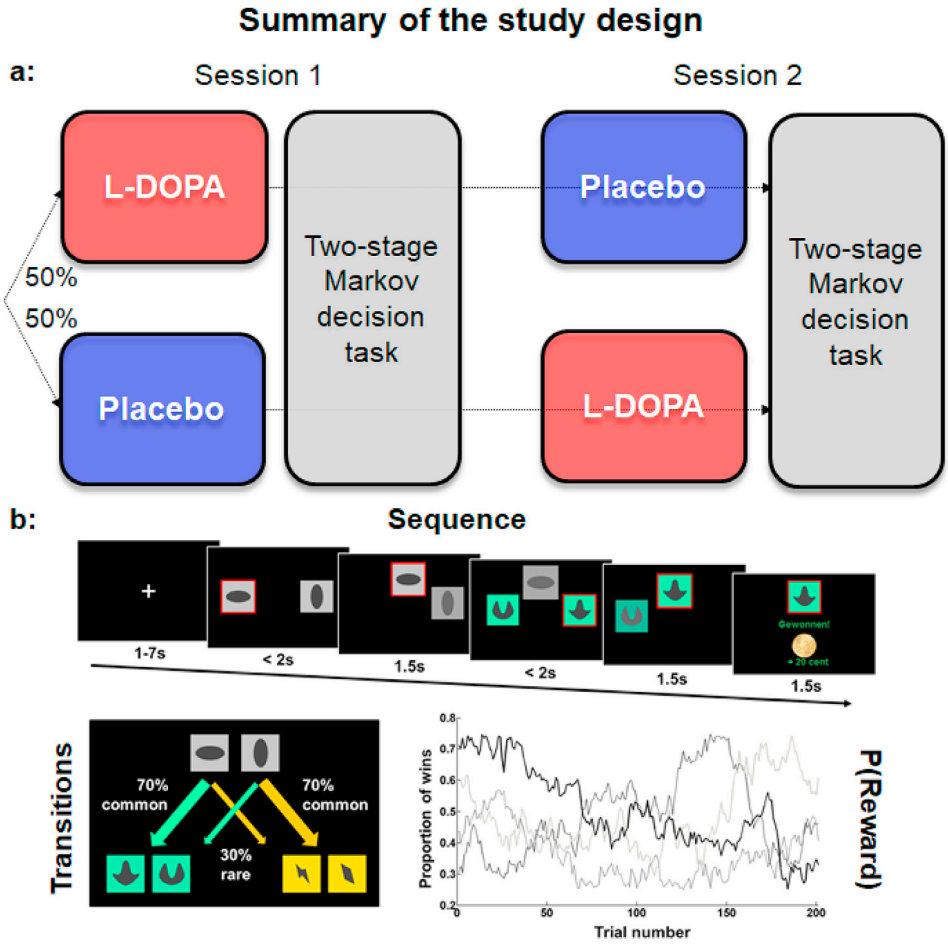

Fig. 1. Schematic summary of the study and task design. A: Participants completed a double-blind randomized cross-over design. Accordingly, they were randomly assigned to L-DOPA or placebo conditions during the first session and then completed the two-stage decision task. They returned for a second session where they completed the other condition before they were scanned a second time. B: Schematic of the task sequence in one trial, transition contingencies, and random walks of the probabilities to win for the four options (green/yellow boxes at stage 2).

suggest that boosting dopamine decreases the impact of learning on actions, one might even expect that learning becomes impaired with L-DOPA. Our results demonstrate that increased dopamine tone reduces reflexive MF control of behavior and increases exploration without affecting learning rates thus providing support for "thrift" (Beeler, 2012) and "value" (Hamid et al., 2016) theories of dopamine.

2. Methods

2.1. Participants

This dataset is part of an ongoing study investigating dopaminergic modulation of reinforcement learning. In order to ensure that the participants of our study were representative of the general population, we requested for postal addresses of individuals randomly selected by the residents' registration office of Dresden, Germany (N = 15,778) and invited them to our study. To limit a confounding effect of age (Eppinger et al., 2013), we recruited participants within a restricted age range of 30–40 years. As part of the general study protocol, participants were invited to a total of four visits, which comprised of a pre-screening visit and two fMRI visits at the Neuroimaging Center, Technische Universität Dresden, followed by a positron emission tomography (PET) visit at the PET center (Lee et al., 2018). For the current analysis, we included 65 healthy participants (49 male; $M_{age} = 37.0$ years, $SD_{age} \pm 3.56$, range [30–42]) who completed both drug and placebo sessions without severe side effects ($N = 10$ had only one visit; $N = 5$ reported severe side effects in the second session) and passed extensive quality control ($N = 4$ were excluded because of low fMRI data quality, $N = 2$ because of >20% missing trials in one session; Fig. 1a). Furthermore, four participants had to be excluded due to 1) brain atrophy, 2) a positive drug test for THC, 3) data loss, 4) erroneous drug manipulation. This sample size provides sufficient power ($1 - \beta = 0.80$) to detect small-to medium-sized effects of repeated measures at the behavioral and ROI level ($\alpha = 0.05$, dz = 0.36) and medium-sized effects at the voxel level ($\alpha = 0.001$, dz = 0.54).

Inclusion criteria for the study were as follows: (1) at least 30 years old at the date of the PET scan, (2) no history of neurological or mental disorders according to the Screening Version of the Structured Clinical Interview for DSM-IV (Wittchen et al., 1997) except for nicotine dependence, (3) no MRI, PET nor L-DOPA contraindications, (4) normal or corrected-to-normal vision, (5) no recent use of illicit drugs (urine test on first fMRI visit; Kombi/DOA10-Schnelltest, MAHSAN Diagnostika GmbH, Reinbek, Germany) nor alcohol consumption (breath-alcohol analysis on both fMRI visits; Alcotest 6510, Drägerwerk AG & Co. KGaA, Lübeck, Germany). Since the main goal of recruitment was to maximize generalization to the population, we deliberately included smokers (approximately every fourth adult in Germany; (Pötschke-Langer et al., 2016)). Smokers were allowed to smoke cigarettes prior to the fMRI visit and breath carbon monoxide levels were measured to assess recent use of cigarettes. We present the results for non-smokers and smokers combined because there were no significant differences between the groups in the main outcomes of the drug intervention. The institutional review boards of Technische Universität Dresden approved the study and we obtained informed consent was obtained from all participants prior to taking part in the experiment.

2.2. Paradigm

Participants performed an adapted version (Sebold et al., 2014) of the two-stage Markov decision task developed by Daw et al. (2011) while undergoing fMRI. For the current study, (a) the instructions were translated into German, (b) visual stimuli were adapted to present different sets of stimuli across visits (pseudo-randomized across participants), and (c) outcome presentation times at both stages were decreased by a factor of 2 to reduce trial duration. The task consisted of a total of 201 trials, separated by inter-trial intervals sampled from an exponential distribution (M = 2 s; range: 1–7 s). For each trial, there were two stages (Fig. 1b). At the first stage, participants had to choose between two grey boxes. After the first-stage choice, they were led to a second stage where they had to make a choice between two colored (green/yellow) boxes. After they have made the second-stage choice, the monetary outcome (win 20 cents or 0 cents) for the trial was presented. In order to optimize their performance, participants had to learn two aspects of the task. First, they had to learn the transition structure, that is, which grey stimulus led to the yellow pair of stimuli in 70% of the trials ("common" trials) and to the green pair in 30% of the trials ("rare" trials; and vice versa for the other grey stimulus). Second, they had to infer the reward probabilities associated with each second-stage stimulus, which followed random Gaussian walks that changed slowly and independently of each other with reflecting boundaries at 0.25 and 0.75. Both aspects were emphasized in the instructions and participants completed 50 practice trials using an independent set of stimuli. After participants completed the practice, they were queried to ensure that they understood a) the transition structure and the difference between common and rare transitions and b) that the best option changes over time due to the random walks. If participants did not answer the query questions correctly, the experimenter repeated the instructions.

This task has been employed previously in multiple studies for characterizing weighted contributions of MF and MB systems in individuals during adaptive learning (e.g., Daw et al., 2011; Deserno et al., 2015a; Deserno et al., 2015b; Sebold et al., 2014; Wunderlich et al., 2012). The key feature differentiating between MF and MB strategies is how first-stage choices are influenced by the "model", that is, the transition structure between the two stages of the task. For example, suppose an individual was rewarded for a second-stage choice during a rare transition trial. In order to be rewarded again for the same second-stage option, MF individuals would repeat their first-stage choice simply because it was rewarded. On the other hand, MB individuals would consider the transition probabilities of both first-stage stimuli. Hence, they would switch to the *other* first-stage stimulus. The reason is that according to the transition structure, switching to the other first-stage option would give them a higher chance to select the same second-stage stimulus, which increases the chances of being rewarded again (Daw et al., 2011).

2.3. Procedure

During the pre-screening visit, we measured height and weight, drew blood samples, and participants completed tasks and questionnaires at a computer. Participants then returned to the scanning facility for their fMRI visits. In order to minimize the influence of medication on BOLD signal, participants were asked to abstain from medication for at least 24 h prior to their visit. As presence of food in the bowels influences the rate of levodopa absorption (Crevoisier et al., 2003), we wanted to control for the amount of food present by asking participants to fast overnight before arriving at the scanning facility. Participants were then given a small standardized breakfast (about 25 g butter biscuits, ~120 kcal) upon arrival (between 05:30–08:45 a.m.) and dextrose tablets throughout the session (about 4 g, ~17.4 kcal/h) to reduce side effect of the drug administration.

Next, participants were instructed and trained on the task. In line with previous studies, we explained that (a) transition contingencies (mapping between first and second stage) would remain fixed throughout the experiment, (b) reward probabilities of each option ("states") at second stage would vary slowly over time independently of each other, and (c) they should try to maximize their monetary outcome throughout the experiment. To familiarize the participants with the task, they were given a computerized practice that consisted of 50 trials prior to the scanning session with the same transition contingencies (70%/30%). To minimize transfer of expectations to the experiment, the practice task had a different set of reward probabilities and stimuli from that used inside the scanner.

Furthermore, participants were asked to complete a series of

computerized and pen-and-paper questionnaires on psychological functioning (e.g., mood). On the second fMRI visit, they completed working memory tests instead of being trained on the two-stage Markov task. After approximately 80 min, we administered 150mg/37.5 mg L-DOPA/benserazide (Madopar; Levodopa and Berazidhydrochlorid; Roche) orally following a double-blind, placebo-controlled (P-Tabletten, Lichtenstein; Winthrop Arzneimittel GmbH) randomized cross-over design. The order of drug condition during the fMRI visits was pseudo-randomized across participants prior to the study. Participants then entered the scanner for structural scans and a second blood sampling (T1) before they proceeded with the two-stage Markov task, which took about 36 min.

After participants completed the two-stage Markov task, a 6-min resting-state scan was collected. The resting-state scan was followed by a booster dose of L-DOPA, a second task inside the scanner, and additional behavioral testing outside the scanner. Participants took approximately 5 h to complete one session. At least seven days after the first fMRI visit (M = 12.2 d, ±9.6 d), participants returned and completed a second visit following the same procedure (except for exchanging the task training for a set of working memory tests), but receiving the complementing drug condition instead.

2.4. fMRI data acquisition and preprocessing

MRI images were acquired on a 3 T Magnetom Trio Tim system (Siemens, Erlangen, Germany) equipped with a 32-channel head coil. During the in-scanner task, stimuli were presented on an MR compatible screen and rearview mirror system. Participants responded by pressing their index fingers on two separate button boxes, one held in each hand. Psychophysics Toolbox Version 3 (Brainard, 1997; Kleiner et al., 2007) implemented within MATLAB R2010a software (The Mathworks, Inc., MA, USA) was used to present the stimuli and collect behavioral data. Functional images were acquired using a gradient echo-planar imaging (EPI) sequence, repetition time TR = 2.41 s; echo time TE = 25 ms; flip angle: 80°; field of view: $192 \times 192\,mm^2$; matrix size: 64×64; voxel size: $3 \times 3 \times 2\,mm^3$ (slice thickness: 2 mm; gap: 1 mm). Every volume consisted of 42 transverse slices acquired descending from the top, manually adjusted ~25° clockwise from the anterior commissure-posterior commissure plane (total ~900 vol for each participant, total scan time: ~36min). A corresponding field map was also recorded for distortion correction of the EPI images. Structural images were acquired using a T1-weighted magnetization prepared rapid acquisition with gradient echo (MPRAGE) sequence for normalization, anatomical localization as well as screening for structural abnormalities by a neuro-radiologist (TR: 1.90 s; TE: 2.52 ms; flip angle: 9°; field of view: $256 \times 256\,mm^2$; number of volumes: 192; voxel size: $1 \times 1 \times 1\,mm^3$).

Functional brain data were preprocessed using SPM8 (Wellcome Trust Center for Neuroimaging, London, UK) implemented within Nipype Version 0.9.2 (Gorgolewski et al., 2011). The first 4 vol of the EPI images were discarded to allow for magnetic saturation. The remaining 896 vol were subjected to slice-time correction (reference: middle slice), followed by realignment to the first volume of the run to correct for motion. Distortion correction based on the field map was then applied to the realigned EPI images. Each individual anatomical T1 image was first co-registered to the individual mean EPI image before segmentation and normalization to MNI space. The resulting transformation parameters were then applied to the distortion-corrected EPI images to spatially normalize them to MNI space (non-linear; resampled to $2 \times 2 \times 2\,mm^3$). Finally, normalized EPI images were spatially smoothed with an isotropic Gaussian kernel (full width at half maximum = 8 mm). During first-level analyses, the data was high-pass filtered at 128 s. As mentioned previously, four participants were excluded from the final analysis because of low fMRI data quality due to excessive in-scanner motion ($N = 2$, >3 mm translation or 3° rotation volume-to-volume), anatomical abnormality ($N = 1$), and failure in image segmentation ($N = 1$).

2.5. Data analysis

2.5.1. Factorial analysis of model-free versus model-based behavior and response time

We investigated the effects of increases in tonic dopamine on stay/switch behavior using the task conditions (reward, transition) in a factorial analysis of the two-stage Markov decision-task (Fig. 1b) (Daw et al., 2011; Deserno et al., 2015a; Wunderlich et al., 2012). We had two behavioral measures of interest, namely the tendency to select the same choice as in the previous trial ("stay") and response time (RT). For this set of behavioral analyses, we estimated the main effects of reward, transition, and the Reward × Transition (RxT) interaction on the repetition of the same choice at the first stage of the next trial using full mixed-effect logistic regression analysis of placebo and drug sessions as implemented in hierarchical generalized linear modeling (HGLM, outcome distribution Bernoulli). The two main effects and the interaction term were treated as random effects, that is, we computed the deviation of each individual from the group effect for drug and placebo sessions to freely estimate individual values of the effects. To test for significant differences, we used empirical Bayes (EB) estimates that take group priors into account and ordinary least squares (OLS) estimates, which are not affected by group priors. We used both estimates because OLS values are "unbiased" and provide a lower-bound estimate of the group effect, but they come at the cost of power. This is because the estimates are not shrunk based on the likelihood of observed values across the group, unlike empirical Bayes estimates, leading to deviations from the normal distribution. For the second set of behavioral analyses, we set up two separate models for RT at first stage and RT at second stage. To estimate factorial effects on RT, we log transformed RT to normalize the distribution and used hierarchical linear modeling (HLM). The RT model at first stage was set up analogously to the choice model. The RT model at second stage included first-stage RT and the current transition as additional predictors. Both RT models included trial number as a random effect to account for potential reductions in RT across the run. All these analyses were conducted with HLM 7 (Raudenbush and Bryk, 2002; Raudenbush et al., 2011).

2.5.2. Computational model

As detailed by Daw et al. (2011), we assumed that agents learn by updating state-action values at each trial, t, through a weighted combination of MB and MF components. The MF component learns the system using a temporal difference algorithm, whereas the MB component does so by maximizing the expected value by taking the contingencies into account.

The second stage, $S^{'}$, has only an MF component driven by the final reward, r_t,

$$Q(S^{'}_{t}, a) = Q(S^{'}_{t-1}, a) + \alpha[r_t - Q(S^{'}_{t-1}, a)],$$

where $0 \leq \alpha \leq 1$ is the learning rate. The MF component of the state-action values at the first stage, S, is updated similarly with the addition of an eligibility trace parameter, $0 \leq \lambda \leq 1$. However, the presence of different transition probabilities (the task "model") leads to a potential MB learning component as

$$Q_{MB}(S_t, a) = P(S^{'}_{1}|S, a)\max_a Q_{MF}(S^{'}_{1,t-1}, a) + P(S^{'}_{2}|S, a)\max_a Q_{MF}(S^{'}_{2,t-1}, a).$$

The overall state-action values at the first stage are computed as a linear combination of the MF and MB components,

$$Q(S_t, a) = \omega Q_{MB}(S_t, a) + (1 - \omega)Q_{MF}(S_t, a) + \pi \cdot rep(a),$$

in which $0 \leq \omega \leq 1$. The tendency to stick to the first-stage action taken at the previous trial is captured by a perseveration parameter π. In order to choose an action, a softmax function maps the state-action values to

choice probabilities at every stage as

$$P(\text{action}|\text{state}) = \frac{\exp(\beta Q(\text{state}, \text{action}))}{\sum_{\text{all actions}} \exp(\beta Q(\text{state}, \text{action}))},$$

in which $\beta \geq 0$ is the so called inverse temperature. Thereby, it represents choice stochasticity, that is, how strongly choices made are related to expected values. The strength of choice consistency is related to the exploration/exploitation trade-off (Sutton and Barto, 1998) and higher values correspond to stronger exploitation of learned values. Considering different learning rates and stochasticity parameters for the two stages, the model comprises seven parameters.

Other variants of reinforcement learning models have been studied for this task before and further details are provided in the SI. Wunderlich et al. (2012) investigated the effects of simplifying the original model to have four parameters $(\alpha, \beta, \omega, \pi)$ and, alternatively, considering different learning rates for positive and negative prediction errors $(\alpha_+, \alpha_-, \beta, \omega, \pi)$. Recently, new formulations have been presented, which decouple MB and MF components and relax the assumption of a linear weighting (Doll et al., 2016; Otto et al., 2013). Furthermore, an alternative is provided by integrated reinforcement learning architectures such as DYNA, which assume that behavior is completely controlled by the model-free system. In this model, the model-based system only has an indirect influence by training the model-free system offline using simulations of the state space of the task (Gershman et al., 2014). However, due to our a priori hypotheses, which were focused on the original formulation of the model, we did not evaluate these alternatives in detail in the current study.

Finally, for any individual, we sought to find a set of parameters which yields the highest likelihood of the data for a given model. Here, we used maximum a posteriori estimation (*MAP*) to improve the estimation of parameters for parametric statistics (Deserno et al., 2015a).

2.5.3. Statistical modeling of fMRI data at first level

First and second-level analyses of fMRI data were carried out using SPM8. For this study, we calculated two sets of first-level statistics that correspond with each of the behavioral analyses. The first set of analyses was conducted in accordance to Daw et al. (2011). The regressors are based on the computational model of behavior (e.g., Deserno et al., 2015a; Deserno et al., 2015b) and use the mean parameter values of the computational model across both sessions. Briefly, it incorporated the onset of stage 1 as event with the parametric regressors first-stage action values, $P(a_{1,t}|s_A)$, and its partial derivative with respect to ω, a combined event regressor for the onset of stage 2 and the onset of outcome presentation with the parametric regressors MF RPEs and MB RPEs. The two events were merged to enable the conjoint estimation of MF and MB RPEs on one onset regressor. To capture potential differences in BOLD response between the two events, the onset of the outcome event was modeled separately in addition. Parametric predictions of MF and MB RPE were derived for both event onsets given the assumption of absolute MF ($\omega = 0$) or MB ($\omega = 1$) control, respectively. Notably, the MB regressor was set up to capture only variance that is not accounted for by MF RPEs since MB RPEs are assumed to act on top of MF control (Daw et al., 2011). As there is no transition involved at the second stage, reinforcement learning is primarily under MF control when the outcome is being presented and MB RPEs are therefore set to 0.

For the second set of first-level statistics, we employed a factorial approach that mirrors the behavioral analysis of stay/switch behavior. Here, we used a regressor for the onset of stage 1 with the parametric regressors previous reward (coded as −0.5/0.5), previous transition (coded as −0.5/0.5), and their interaction, a regressor for the onset of a common transition trial at stage 2, a regressor for the onset of a rare transition trial at stage 2, a regressor for a reward at outcome onset (i.e., the result of the choice at stage 2), and a regressor for reward omission at outcome onset. Specifically, we used the contrast rewarded – unrewarded trials at outcome onset and the parametric effect of preceding reward at first-stage onset to evaluate the simple main effects of L-DOPA on processing of reward, which were tested based on the observed behavioral differences between drug conditions.

For group inferences, we computed second-level group statistics with placebo vs. L-DOPA as repeated measures factor for the parametric contrast images (MF RPE, MB RPE, first-stage action values, rewarded vs. unrewarded trials). The second-level statistics based on the computational model included order as a nuisance variable. Since the effects of order were negligible, we included only the estimated effect of the reflexive effect of reward on stay behavior when we evaluated brain-behavior interactions.

ROI analyses were focused a priori on the NAcc and vmPFC where both types of signal correlates are evident (Daw et al., 2011; Deserno et al., 2015a). Valence effects of reward and MF RPE signals were also expected a priori to occur in the VTA/SN (e.g., see the term "reward" at www.neurosynth.org; (Yarkoni et al., 2011)). For brain–behavior coupling, we furthermore included the ACC since it is critically involved in the allocation of effort according to learned action policies (Silvetti et al., 2013; Verguts et al., 2015).

2.5.4. Full mixed-effects modeling of time courses

To maximize the sensitivity in detecting potential drug effects on RPE signals, we complemented the common so-called summary statistic random effects analysis by implementing a full mixed-effects design (i.e., two-level HLM). Here, we estimated drug effects by incorporating group priors based on extracted ROI time series as described by Kroemer et al. (2014). To this end, we extracted the first eigenvariates from anatomical masks of the NAcc and vmPFC where both MF and MB RPE signals can be observed. This method improves sensitivity of parametric statistics similar to MAP estimation for the behavioral data. To estimate drug effects on variables of interest, we allowed for interactions of L-DOPA with MF RPEs, MB RPEs, and first-stage action values. All events, parametric regressors, and interactions of L-DOPA with parametric regressors were modeled as random effects at the participant level and, thereby, led to parameter estimates for each individual.

2.5.5. Full mixed-effects modeling of exploratory behavior and response time (as depicted in Fig. 7)

We evaluated choice probabilities and RT in more detail using the fitted predictions of the computational model (M_{7P}) obtained after the MAP estimation for each individual participant. We concatenated data from placebo and drug sessions and predicted choice probabilities at the first (Fig. 7a–b) and second stage (Fig. 7a–e). For Fig. 7f, we predicted second stage RT as the outcome (ln transformed to meet distributional assumptions). The setup and output of the full-mixed effects models is detailed in the SI (Table S4). The average reward rate was computed by using a Gaussian moving window incorporating the past 5 wins according to a recency-weighting scheme and varied between 0 (5 omissions in a row) and 1 (5 wins in a row). For statistical analyses, we centered the reward rate for each individual session. Thus, the regressor captures residual variance throughout a session that is accounted for by the recent stream of success.

2.5.6. Statistical threshold and software

For behavioral and ROI analyses, we used $\alpha = 0.05$ (two-tailed) as significance threshold and performed correction for multiple comparisons based on the hypothesis test as detailed in the results section. For fMRI whole-brain analyses, we used one-sided contrast maps (mass-univariate *t*-tests) thresholded at a voxelwise $p < .001$ to assess cluster size. To facilitate the visualization of the involved brain structures, we show the images at a slightly lower threshold as indicated in the figures. To check for robustness, key parametric results were also assessed with non-parametric equivalents and corresponding *p*-values were derived by Monte Carlo simulations. To analyze and plot data, we used SPSS v21–23, R v3.2.2 (R Core Team, 2015), R Deducer (Fellows, 2012), HLM v7 (Raudenbush and Bryk, 2002; Raudenbush et al., 2011), Mango v3.6–3.8, and MATLAB v2012-2015.

3. Results

3.1. L-DOPA reduces the effect of reward on stay/switch behavior

In line with previous studies (Daw et al., 2011; Wunderlich et al., 2012), we estimated the main effects of reward, transition, and the Reward × Transition (RxT) interaction on the repetition of the same choice at the first stage of the next trial ("stay") using full-mixed effects modeling of placebo and drug sessions (see methods). The obtained individual estimates of the effects were then compared between conditions and used for further correlational analyses. Within this hierarchical logistic regression analysis, the main effect of reward corresponds to MF control (i.e., the reflexive effect of reward regardless of transition) whereas the RxT interaction term captures the use of the transition structure in the goal-directed pursuit of reward, corresponding to MB control.

Administration of L-DOPA attenuated the effect of preceding reward on stay probabilities of the next trial, $t(64) = -2.50$; $p = .015$; Fig. 2b–d, thereby reducing the reflexive MF facilitation of response after reward. An equivalent yet less powerful non-parametric test (i.e., Wilcoxon signed rank test on OLS estimates disregarding group priors) yielded a slightly attenuated estimate of the drug effect on MF control, which was only significant at trend level ($Z = -1.74$; Monte Carlo $p = .084$ 95% CI [0.079-0.089]). Descriptively, receiving a reward increased the odds of repeating the same choice at stage 1 of the next trial by 116.8% during placebo sessions ($b_{P.REW} = 0.774$) whereas it only increased the odds by 91.8% during L-DOPA sessions ($b_{D.REW} = 0.651$). In contrast, L-DOPA had no significant effect on MB control as indicated by RxT ($p = .27$), nor the intercept, which reflects a general bias to repeat the previous choice (at centered reward and transition terms, $p = .75$).

3.2. L-DOPA reduces model-free control at second stage in the computational model

To provide a more nuanced assessment of drug effects on reinforcement, we set up the seven parameter computational model (M_{7P}) proposed by Daw et al. (2011) and used maximum a posteriori (MAP) as fitting algorithm to better approximate the normal distribution of model parameters for parametric statistics (Deserno et al., 2015a). L-DOPA increased stochasticity of choices at the second stage (i.e., reduced $\beta 2$), which is primarily under MF control because there is no transition involved at the second stage ($p = .008$). This significant difference was also seen in the non-parametric Wilcoxon signed rank test ($Z = -2.80$; Monte Carlo $p = .004$ 95% CI [0.003-0.005]). Lower stochasticity at the second stage was also associated with a greater reflexive effect of reward at the first stage, $r_{P+D}(128) = 0.23$, $p = .010$, corroborating a contribution of MF control on the effects of $\beta 2$. Notably, L-DOPA appeared to have opposite effects on stochasticity, leading to a positive Stage × Drug interaction ($p < .001$; Fig. 3) despite the positive correlation between βs at both stages, $r_D(63) = 0.43$, $p < .001$; $r_P(63) = 0.36$, $p = .004$.

However, no main or interaction effect of drug was seen for learning rates at both stages and all other parameters were not significantly different. In line with previous research (Cools et al., 2001; Wunderlich et al., 2012), L-DOPA tended to increase switching as captured by the repetition bias parameter π ($p = .067$; Fig. 3), but this effect was also not significant. Lastly, there were no significant effects of gender nor order in factorial or computational analyses of behavior.

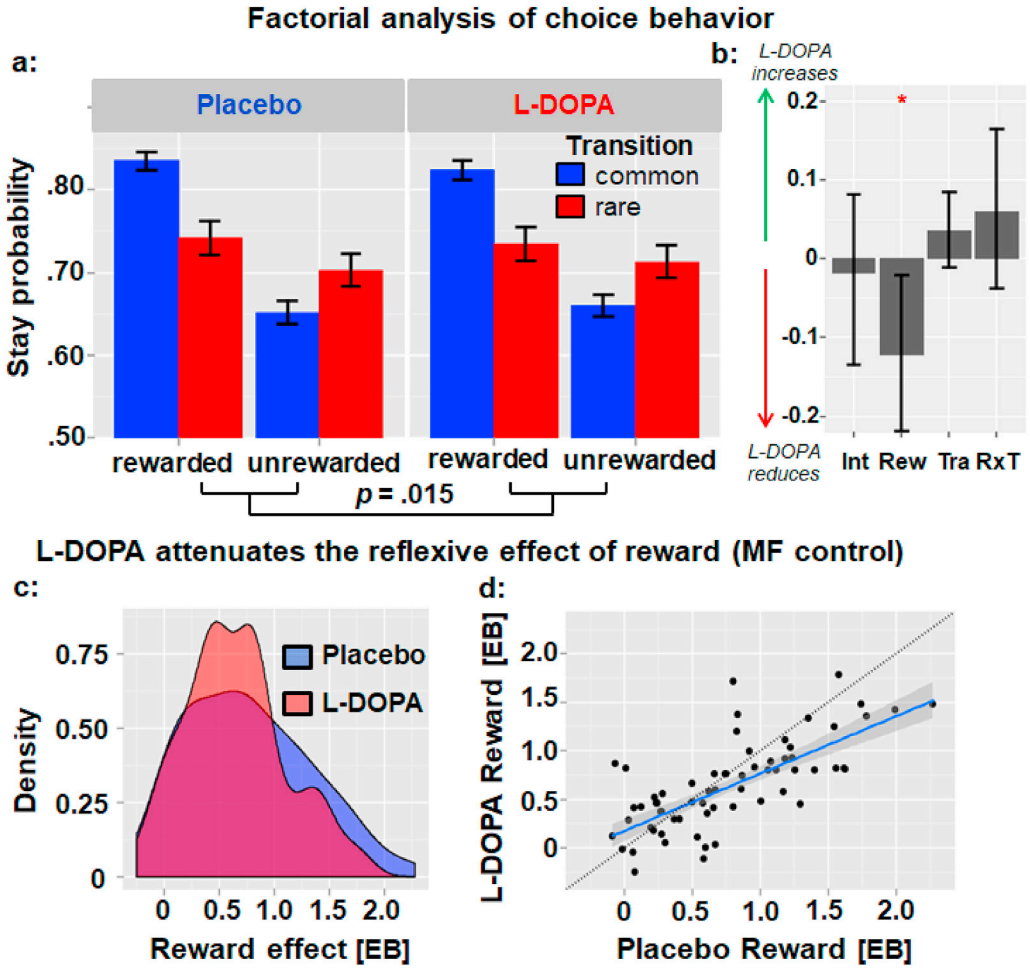

Fig. 2. L-DOPA decreases the reflexive effect of reward on stay/switch behavior. A: The factorial analysis of stay/switch behavior indicated that L-DOPA attenuated the effect of preceding reward on stay probability of the next trial, which is a signature of model-free (MF) control of behavior. Error bars depict 95% confidence intervals at the trial level. B: Differences in parameter estimates extracted from the full mixed-effects model of stay/switch behavior. L-DOPA led to a significant decrease in the regression weight of the previous reward. Here, error bars depict 95% confidence intervals at the participant level and exclusion of 0 corresponds to a significant difference in a random-effects analysis. (Int: intercept, Rew: Reward, Tra: Transition, RxT: Reward x Transition). C: Density plots of the distribution of empirical Bayes (EB) parameter estimates of the preceding reward effect on stay/switch behavior. Inspection of the density and scatter plots (d) indicates that L-DOPA "shrinks" strong model-free control of behavior during placebo towards zero, leading to an overall decrease in the impact of reward on stay probabilities. Blue line = robust linear fit, black dotted line = identity.

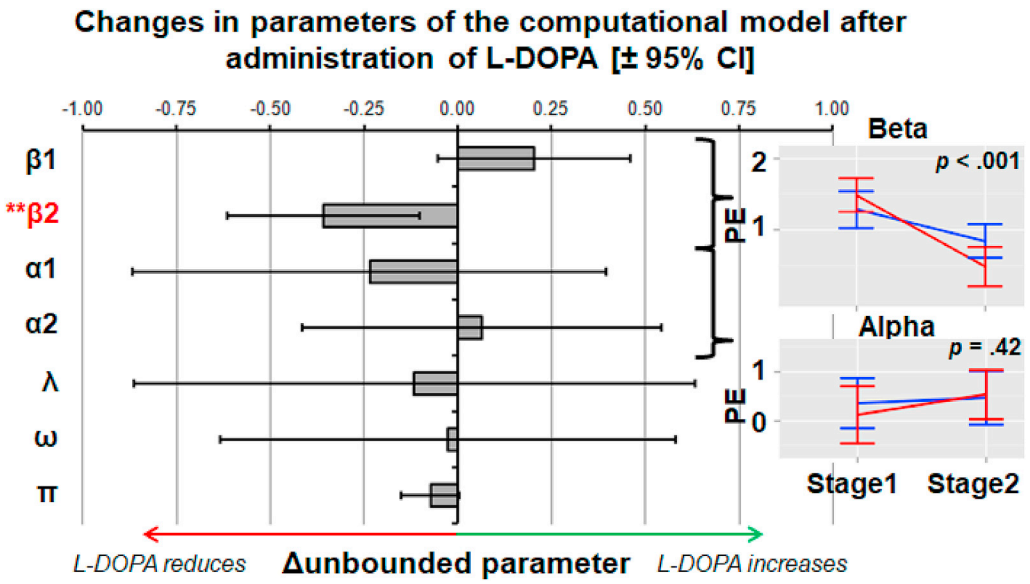

Fig. 3. Administration of L-DOPA increased stochasticity of choices at the second (i.e., primarily model-free) stage. L-DOPA reduces $\beta 2$, which reflects the consistency of choices with values at stage 2, $t(64) = -2.737$; $p = .008$. Interactions of L-DOPA with task stage on α and β parameters are shown in the inset graph. EB = empirical Bayes, PE = parameter estimate.

Due to the fact that we could not replicate the increase of MB control after administration of L-DOPA reported in Wunderlich et al. (2012), we tested if drug effects were dependent on overall model fit. First, L-DOPA did not significantly alter the log likelihood (LL) of the computational model, $t(64) = 1.50$, $p = .15$. Second, we added $LL_{M7.D}$ to repeated measures ANOVAs where estimates of MB control (RxT or ω) were the dependent variables, respectively. Administration of L-DOPA increased the RxT interaction term when the model fit was high, $F(1,63) = 10.92$, $p = .002$. A similar trend was observed for ω, $F(1,63) = 2.84$, $p = .097$. In other words, increases in MB control elicited by L-DOPA might be dependent on how well the model captures behavior. To test if the differences in drug effects might be driven by differences in the characteristics of the selected samples (i.e., undergraduates vs. a representative of adults), we used a working memory score computed from the operation span task (for details, see SI). Higher working memory capacity was associated with better average model fit ($p = .015$) and increases in MB control after L-DOPA (ω: $p = .003$; RxT: $p = .089$). These results indicate that the effects of L-DOPA are partly dependent on cognitive abilities such as working memory capacity. Critically, reduced MF control of behavior after administration of L-DOPA as evidenced by the effect of preceding reward or $\beta 2$ was found independently of model fit and working memory capacity.

Furthermore, we examined alternative models tested by Wunderlich et al. (2012). However, the models were inferior in terms of overall model fit and we found no consistent effect of L-DOPA on MB behavior or reward learning (see Table S1, SI). Lastly, similar effects of L-DOPA on $\beta 2$ ($p = .013$; Wilcoxon signed ranks test $Z = -2.75$; Monte Carlo $p = .006$ 95% CI [0.004-0.007]) were also obtained with a recent reparametrization of M_{7P} where ω is replaced by separate βs for MF and MB control (Doll et al., 2016; Otto et al., 2013; Sebold et al., 2016) indicating that the obtained result is robust to minor changes in the setup of the computational model.

To summarize, L-DOPA increases stochasticity of choices at the second stage of the task which is primarily under MF control. In contrast, drug effects on MB control were inconsistently found across models and not evident in the best fitting model. Furthermore, potential increases in MB control following the administration of L-DOPA were dependent on high overall model fit or working memory capacity. Hence, our results provide a partial replication of the study by Wunderlich et al. (2012) and suggest that improvements in MB control are more strongly dependent on individual characteristics. Taken together, the results of the computational models echo the results of the factorial analysis and point conclusively to reduced MF control, but largely unaffected MB control (on average) after administration of L-DOPA.

3.3. No effect of L-DOPA on BOLD correlates of RPE signals

To test for effects of L-DOPA on BOLD response correlates of RPE signals, we set up two separate second-level statistics for MF RPE and MB RPE contrast images including both conditions as repeated measures factor while controlling for order as a covariate. In line with previous studies (Daw et al., 2011), we observed a widely distributed brain network tracking MF RPE signals encompassing the ventral and dorsal striatum, the dopaminergic midbrain, the orbitofrontal cortex, the dorsolateral prefrontal cortex, and the posterior cingulate cortex (Fig. 4; Table S2). Contrary to MF RPE, MB RPE signals were found to be spatially sparser encompassing the ventral striatum (where overlap with MF RPE signals occurs) and vmPFC only. These results replicate the results reported by Daw et al. (2011).

In contrast to predictions of the reinforcement learning theory, yet in line with the value theory and the absence of behavioral effects on learning rates, we observed no modulatory effects of L-DOPA on RPE signals (Fig. 5). At a whole-brain level, no differences survived a correction for multiple comparisons (cluster-forming threshold $p < .001$) and drug effects were virtually absent even at an uncorrected cluster extent threshold (see SI). Likewise, at the level of a priori ROIs, we observed no effects of L-DOPA ($ps > .1$). This absence of a drug effect was further corroborated by an additional time course analysis: We concatenated sessions and participants to improve the estimation of potential drug effects across the group, similar to the behavior analysis, but failed to see a modulatory effect of L-DOPA on RPE signaling ($|t| < 1.21$, $p > .23$). Whereas there was no effect of L-DOPA on average parametric effects, we found an increase in interindividual variability of the MF RPE signals in the NAcc during the L-DOPA session, $F(64,64) = 1.61$, $p = .030$, which would, however, not survive correction for multiple comparisons across ROIs and/or contrasts. To summarize, across whole-brain and a priori ROIs analyses, there was no indication of an altered correspondence between BOLD response and RPE signals after administration of L-DOPA.

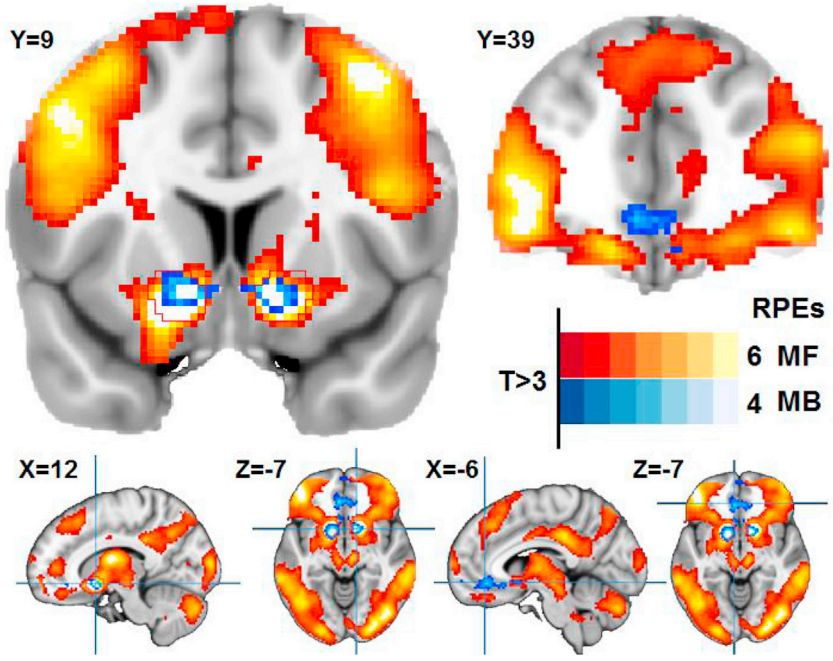

Fig. 4. Model-free (MF; red-yellow) and model-based (MB) reward prediction error signals (RPE; blue-white) pooled for placebo and L-DOPA sessions (with session order as covariate). In line with Daw et al. (2011), overlap of MF and MB reward prediction error signals were found in the nucleus accumbens (ROI outline shown in red) and ventromedial prefrontal cortex.

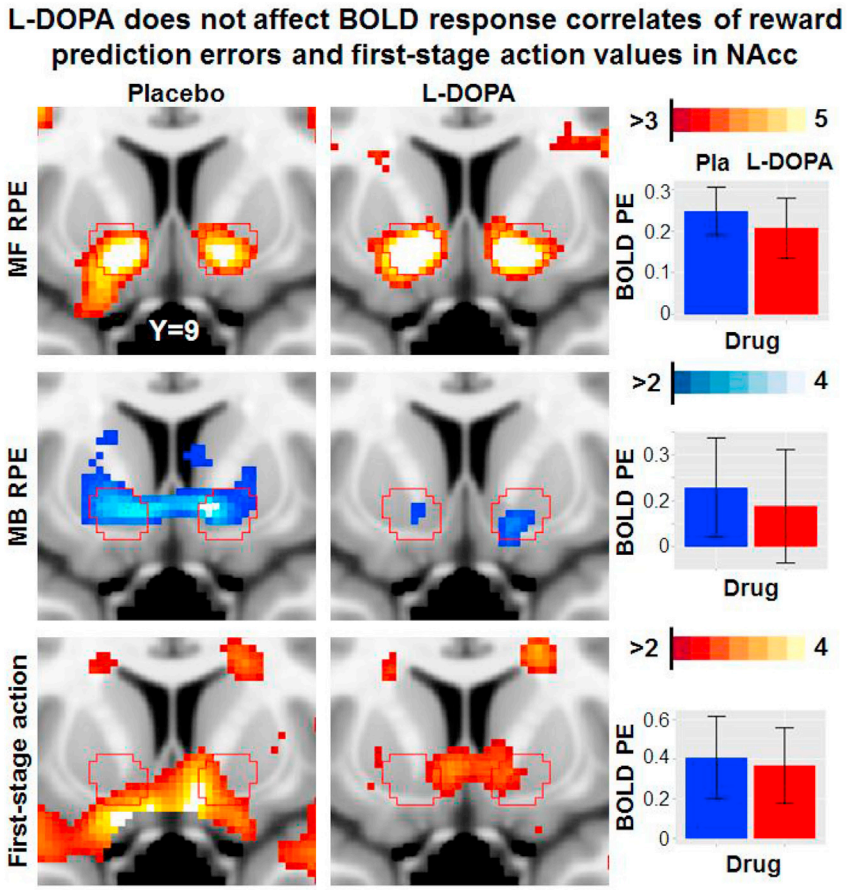

Fig. 5. A priori region-of-interest (ROI) sections of the nucleus accumbens (NAcc) for the parametric effects of interest model-free (MF) and model-based (MB) reward prediction error (RPE) signals and first-stage action values depicted separately for placebo and drug sessions. Inset graphs show extracted parameter estimates (PE) and 95% confidence intervals (CI). Likewise, non-significant differences were obtained in ROI analyses for the ventromedial prefrontal cortex and in exploratory whole-brain analyses. Pla = placebo, PE = Parameter estimate.

3.4. L-DOPA reduces the effect of reward at the outcome and subsequent trial-onset stage

In the factorial analysis and the computational model of behavior, we found that L-DOPA reduced reflexive MF control by reward without altering the encoding of RPEs. Hence, we set up an alternative first-level model of the task to assess the simple main effect of rewarded vs. unrewarded events. Second-level statistics were computed separately for the outcome stage and the subsequent trial onset, where participants may decide to stay or switch depending on the outcome and the transition of the previous trial. Based on the behavioral results, which were in line with the predictions of the thrift and value theories of dopamine, we hypothesized that L-DOPA would reduce the effect of reward (rewarded – unrewarded events) on BOLD response in the mesocorticolimbic ROIs VTA/SN, NAcc, and vmPFC. To evaluate the link between changes in reward signals and MF control, we included the individually estimated effect of reward obtained in the factorial analysis of stay/switch behavior conducted at this stage of the task as a covariate (Fig. 2b–d). This parameter captures the degree to which individuals reflexively adjust their first-stage choice after receiving a reward.

At the outcome stage, we found a significant effect of reward across the three ROIs ($p < .001$), no significant effect of drug ($p = .23$), but a significant interaction Drug x Reward ($p = .027$; Table S3). This interaction effect was driven by a reduced contrast between rewarded and unrewarded events during L-DOPA vs. placebo visits (Fig. 6a; Figure S2). Additional single ROI models indicated that the Drug x Reward effect was significant within the VTA/SN ($p = .014$; survives correction across 3 ROIs) and vmPFC ($p = .047$), but not in the NAcc alone ($p = .32$). Analogous effects were obtained with a voxel-based approach and coordinates of peak effects are reported in the SI. No other effects were seen in a whole-brain analysis that survived correction for multiple comparisons and there was no general effect of reward at the outcome stage on grey matter BOLD response, $t(64) = -0.482$, $p = .63$.

Likewise, at the onset of the following trial, we observed a significant decrease of the reward effect on BOLD response across the three a priori ROIs after administration of L-DOPA ($p = .019$). At this stage, however, the effect was not only restricted to the mesocorticolimbic system because we found a significant modulation at the whole-brain level across grey matter voxels, $t(64) = -3.034$, $p = .003$; Fig. 6b.

Lastly, to evaluate if L-DOPA affects the coupling between behavior and reward signals, we assessed if L-DOPA reduced the correspondence between MF control as captured by the reflexive effect of reward and BOLD response to reward. Here, we observed a reduced correlation between the neural and the behavioral reward effect in the ACC ($r_P(63) = 0.36$, $p = .003$; $r_D(63) = -0.18$, $p = .15$; drug interaction term in HLM $p = .004$; voxel-based $t_{max}(128) = 3.38$, $p_{SVC} = .056$; 2/38/12),

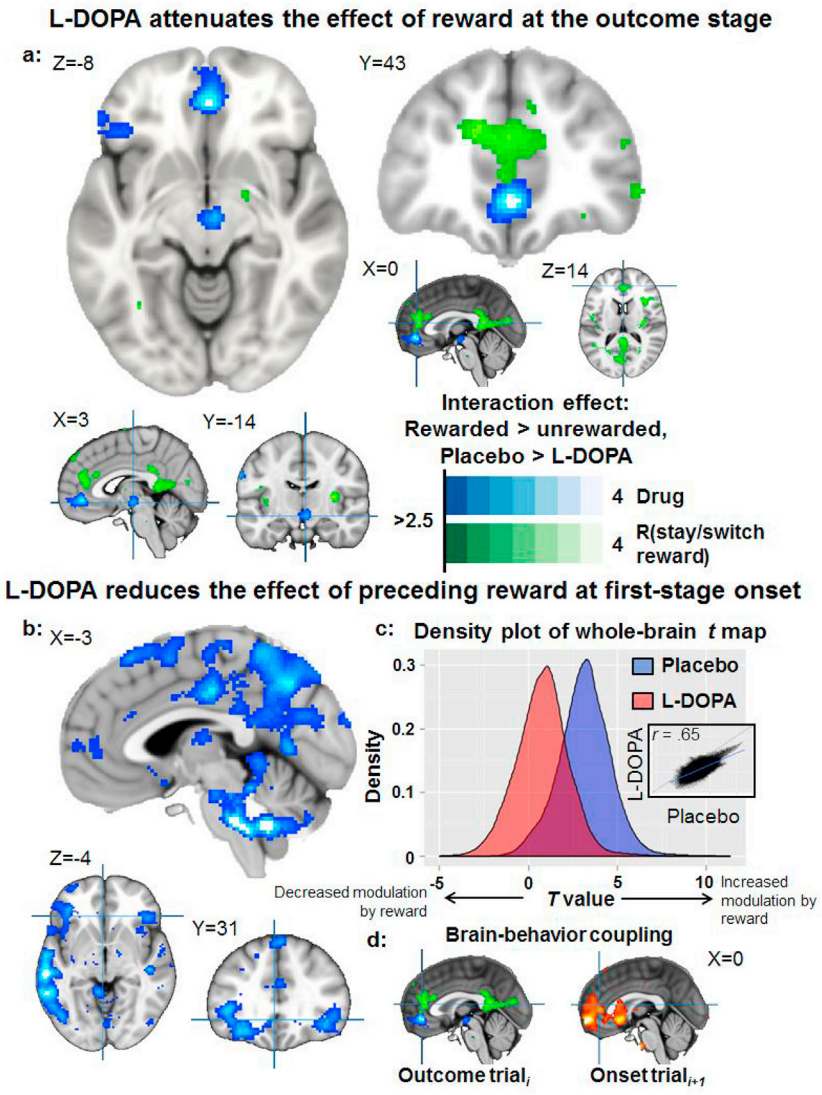

Fig. 6. A: L-DOPA reduced the effect of reward on brain activation in the ventral tegmental area and the ventromedial prefrontal cortex (blue color map) at the outcome stage. Moreover, it reduces the correlation of brain activation with the reflexive effect of reward on behavior (green color map, derived from stay/switch analyses) in an adjacent cluster of the anterior cingulate cortex (drug interaction within the ROI $p = .004$). This indicates that the transfer of model-free values to action is reduced. B: In line with the attenuating effect at the outcome stage, L-DOPA reduces the effect of preceding reward on brain activation at the onset of the next trial throughout the brain (c). The inset scatterplot depicts that the group *t*-values are reliable across conditions at a voxel-by-voxel level, but that the preceding reward effect is globally attenuated (grey matter $p = .003$). D: At first-stage onset, stronger activation in the prefrontal cortex is associated with the effect of reward on behavior regardless of drug condition (hot color map; scaled analogous to the other contrasts and plotted next to the image of the Drug × Reward Bias interaction in a).

which would survive correction for multiple comparison across ROIs and stages ($\alpha < 0.0083$). At the first stage of the following trial, we found that the neural effect of preceding reward corresponded with the reflexive effect of reward in the vmPFC regardless of the drug condition ($r_P(63) = 0.24$, $p = .053$; $r_D(63) = 0.34$, $p = .006$, no significant drug interaction; $t_{max}(128) = 4.60$, $p_{SVC} = .002$; $k = 605$, $p_{FWE.clust} = 0.002$; −4/60/-6; extending to the ACC ROI $t_{max}(128) = 3.52$, $p_{SVC} = .035$; −2/48/6). Furthermore, we observed a significant correlation in the orbitofrontal gyrus outside of our a priori ROIs, $t_{max}(128) = 4.64$, $k = 511$, $p_{FWE.clust} = 0.004$; 20/32/-8.

To summarize, L-DOPA reduced the correspondence between the neural reward effect at the outcome stage and the behavioral reward effect in the ACC. Moreover, when participants were facing the next choice, the facilitating effect of reward on BOLD response was reduced throughout the brain, but the correspondence between the neural reward effect and behavior was not affected by L-DOPA. This suggests an attenuated net effect of the reward. Hence, this corroborates the observed decrease in the reflexive effect of reward on behavior.

3.5. L-DOPA increases exploratory behavior and deliberation time

Whereas both value and thrift theories predict that increases in dopamine tone reduce MF control as we have observed, one of the key predictions of the thrift theory is that heightened dopamine tone will facilitate exploratory behavior. To test this hypothesis, we used multilevel analyses of RT as well as first- and second-stage action values to assess changes in deliberation time. We hypothesized that L-DOPA would increase deliberation time leading to slower RT for switches in addition to the previously observed less deterministic choices reflecting "noisy" exploration (Kayser et al., 2015).

In line with the hypothesized change in deliberation time, we observed that L-DOPA increased the discrepancy in RT between stay and switch trials at stage 1 (Wilcoxon signed ranks test $Z = -2.565$, Monte Carlo $p = .011$ 95% CI [0.009, 0.013]), which was mainly driven by slower RT during switch trials in L-DOPA sessions ($RT_{D.switch} = 0.781$ s vs. $RT_{P.switch} = 0.765$ s; $RT_{D.stay} = 0.739$ s vs. $RT_{P.stay} = 0.735$ s). Notably, there was no general effect of L-DOPA on RT at first stage and no shift in

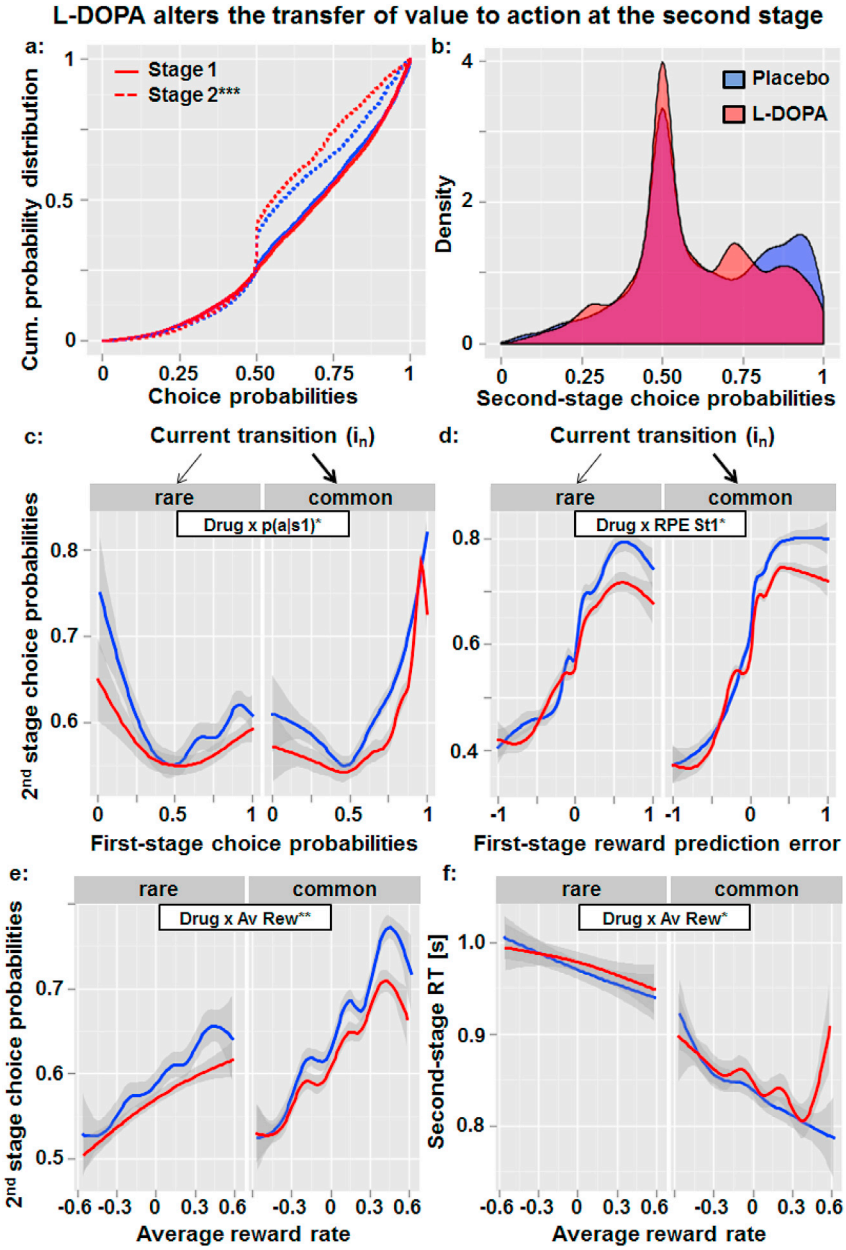

Fig. 7. A: L-DOPA reduces stochasticity of choices at the second (i.e., model free; $p < .001$), but not the first stage ($p = .52$). B: Group density plots indicate that the net effect is mainly due to a reduction of choices with very high probabilities given the model's estimation of choice values. C: The reduced correspondence of first-stage choice probabilities ($p = .014$) and (d) reward prediction errors with second-stage choice probabilities ($p = .019$) is not affected by transition and apparent for both rare and common transitions (Drug × Transition $p > .27$). E: The average reward rate (Av Rew), which reflects the recent history of wins (centered on each individual's average percentage of wins), has less of an effect on second-stage choice probabilities after administration of L-DOPA ($p = .003$). F: The increase in response time at a high average reward rate in the L-DOPA condition ($p = .048$), particularly in common trials, suggests that instead of a reflexive invigoration of behavior, a recent stream of wins may have facilitated deliberation time. This might add to the increased stochasticity of choices (noisy exploration) at the second stage. For the fitted models, see SI.

the distribution of the probability of choices (Fig. 7a).

At stage 2, we observed a decrease in choices at the deterministic end (i.e., with very high probabilities; Fig. 7a–b), which is reflected in the $\beta 2$ parameter in the computational analysis. This observation was corroborated by the fact that L-DOPA reduced the correspondence between the estimated evaluations of options and resulting actions. The increased independence of action from value with higher dopamine tone was evidenced by the reduced correspondence between first- and second-stage choice probabilities ($p = .014$; Fig. 7c) and the reduced correspondence between first-stage prediction error signals and second-stage action values ($p = .019$; Fig. 7d).

Furthermore, we assessed how increases in dopamine tone affect action control by the average reward rate, which has been linked to endogenous fluctuations in dopamine tone before (Niv et al., 2007). In line with the predictions of the thrift theory, L-DOPA reduced the correspondence between the average reward rate and choice probability at second stage ($p = .003$; Fig. 7e). Moreover, L-DOPA led to a characteristic increase in second-stage RT when the average reward rate was high and, hence, uncertainty is expected to be low ($p = .048$; Fig. 7f). This increase in deliberation time was not seen during placebo sessions. Here, RT decreased monotonically with a recent stream of wins in the task as one would expect when past success is being exploited to maximize wins. Collectively, our data suggests that L-DOPA reduces MF control by facilitating noisy exploration and deliberation during the task.

3.6. L-DOPA does not change overall task performance

As we have detailed before, L-DOPA had no overall effect on the obtained model fit of the computational model of choice and on general RT. However, given that L-DOPA increases stochasticity of choices at the second stage of the task and facilitates exploration; does it come at the cost of the number of obtained rewards? We found that L-DOPA tended to increase the rate of rewards obtained during the task, $t(64) = 1.70$, $p = .093$; $M_D = 48.8\%$ vs. $M_P = 47.7\%$, demonstrating that increased stochasticity was not detrimental for overall task performance. Thus, there was no indication that the observed reduction in MF control with L-DOPA was disadvantageous in terms of obtained reward.

4. Discussion

Tonic and phasic dopamine signaling are known to play key roles in reinforcement learning and action control. In the current study, we found that increases in dopamine tone induced by L-DOPA reduced the reflexive MF control of behavior via reduced direct reinforcement of successful actions. In contrast, deliberative MB control of behavior was unaffected by L-DOPA. The changes in MF control of behavior were not explained by changes in learning rates at different stages of the task or for positive versus negative outcomes. Moreover, these changes were not explained by differential coding of RPE signals as we did not observe differences in their correspondence with BOLD response. Nevertheless, in line with behavioral effects, L-DOPA reduced the effect of reward in the brain. First, it reduced the reward effect in the mesocorticolimbic system when the outcome (reward vs. no reward) was presented. Second, it reduced the facilitating effect of preceding reward on BOLD response in general at the beginning of the next trial. Collectively, these results conclusively indicate that L-DOPA reduces MF control of behavior by reducing the transfer of learned value to overt actions, which is in line with the recent value (Hamid et al., 2016) and thrift theories (Beeler, 2012) of dopamine function. This interpretation is corroborated by increases in noisy exploration and deliberation, which supports one of the unique predictions of the thrift theory (Beeler, 2012; Beeler et al., 2012, 2016) that has not been extensively tested in humans before. Hence, our results add to the growing evidence for a dopaminergic contribution to exploratory decisions (Beeler et al., 2012; Kayser et al., 2015).

Our main behavioral finding that L-DOPA reduces MF control without affecting MB control of behavior is in line with the observation that behavioral effects of phasic dopamine signals are dependent on their tonic baseline (Grace, 1991; Hamid et al., 2016), which prominently modulates the signal-to-noise ratio of phasic burst signaling (Zhang et al., 2009, 2012). MF control is mainly driven by phasic firing of dopaminergic neurons encoding temporal difference RPEs (Schultz, 2013, 2015; Stauffer et al., 2014). We observed that RPE signals and learning rates were not altered by an increase in dopamine tone, which suggests that reward learning *per se* was not affected. Whereas Pessiglione et al. (2006) had reported increased amplitudes of RPE signals after L-DOPA that explained improvements in reward learning, these effects were only significant compared to the haloperidol group. Furthermore, in line with our results, Wunderlich et al. (2012) did not observe significant effects of L-DOPA on learning rates. Consequently, an increased baseline of dopamine may reduce the signal-to-noise ratio of phasic dopamine release since the relative change in dopamine is attenuated. In turn, this might reduce their potential to reflexively adjust ("drive") actions. Such a decoupling between the dopaminergic mechanisms of reward learning and action control with heightened tone has been conclusively demonstrated in animal models (Hamid et al., 2016; Syed et al., 2016). When alterations in behavior occur without resultant changes in reward learning (Beeler et al., 2010; Hamid et al., 2016), the correspondence between action and learned value is effectively reduced (Beeler et al., 2010). Notably, Wunderlich et al. (2012) also reported an (non-significant) increase in choice stochasticity for L-DOPA. Since they only reported drug effects for the reduced model that did not differentiate choice stochasticity between the two stages of the task, it is conceivable that this could have masked a significant effect of L-DOPA given the marked Stage × Drug interaction on stochasticity that we observed.

Notwithstanding, the absence of L-DOPA effects on MB control appear to be in contrast with the findings reported by Sharp et al. (2016); or Wunderlich et al. (2012). However, several aspects may contribute to the observed discrepancies. First, when we included the obtained fit of the computational model during the L-DOPA session as a covariate, we found that MB control only increased in participants whose behavior was well approximated by the model. Likewise, when we entered working memory capacity as a covariate, we found that MB control only increased in participants with high working memory capacity. Thus, it is possible that our representative sample of adults performed the task less in line with the computational model or at a lower overall level, which could reduce the net effect of the drug in the studied sample. In other words, this suggests that facilitating effects of L-DOPA on MB control do not generalize to the population because they could be mediated by inter-individual differences in working memory capacity (Otto et al., 2013) or, perhaps, presynaptic dopamine levels in the ventral striatum (Deserno et al., 2015a) that are known to contribute to overall task performance. If we were to restrict our analysis to individuals with high working memory capacity, we could essentially replicate that L-DOPA increases MB control (Wunderlich et al., 2012). Thus, as Wunderlich et al. (2012) have argued, L-DOPA may enhance MB control by enhancing existing cognitive function and, as we have shown, by decreasing MF control leading to a greater net effect in employing MB control if participants have a high working memory capacity. Second, effects of L-DOPA on the MF vs. MB control weighting parameter ω were inconsistent across the alternative computational models, which may reduce reproducibility of results in general. Importantly, the reduction in MF control of behavior was observed independently of the fit and the parametrization of the computational model suggesting that this was a general effect of increased dopamine tone. Third, both previous samples were small ($N_1 = 18$; $N_2 = 22$) and given our effect-size estimate of reduced MF control ($r \sim 0.30$), the power to detect such an effect would be below 30%. In line with this interpretation, both studies reported decreased MF control which was, however, not significant. Collectively, these results suggest a complex interaction between tonic and phasic dopamine signaling on the one hand and the arbitration between MF vs. MB control on the other hand. Whereas L-DOPA may reduce MF control in general yet to a moderate degree, the net effects on MB control are perhaps mediated by interindividual

differences in task performance and working memory.

Critically, the behavioral effects of L-DOPA on MF control were also echoed in BOLD response. In line with reduced MF control of behavior, we observed an attenuation of the reward outcome signal in the mesocorticolimbic system and a global attenuation of the effect of preceding reward at the onset of the subsequent trial. Moreover, in the ACC, L-DOPA abolished the correspondence between reward outcome signals and MF control of behavior. Taken together, these findings suggest that L-DOPA reduces the transfer of value to action, which is well in line with previous studies. In patients with Parkinson's disease, L-DOPA (on medication condition) disrupts performance of reversal learning (Cools et al., 2001) and NAcc activation during reversal learning, particularly when the final reversal error occurs (i.e., preceding a switch (Cools et al., 2007)), which might be indicative of reduced transfer of value. Moreover, L-DOPA has been shown to reduce habit learning (i.e., MF control) in the weather prediction task in patients off medication (Fuhrer et al., 2014). Likewise, L-DOPA decreased the influence of a "Pavlovian controller" on instrumental action in healthy individuals such that action learning became less affected by outcome valence (Guitart-Masip et al., 2014) and it increased a value-independent propensity to gamble by making risky options more attractive to pursue (Rigoli et al., 2016b). Thus, increases in tonic dopamine may allow an individual to "escape" their previous reward history by embarking on a more effortful exploration of available options in the environment (Beeler, 2012; Beeler et al., 2012). It is plausible that this neurobiological mechanism might also account for reported increases in task switching (Cools et al., 2001) and this hypothesis calls for future research on a more comprehensive set of tasks.

The current study has limitations that will need to be addressed in future research. First, whereas it is commonly assumed that the dopamine precursor L-DOPA primarily raises dopamine levels in the striatum, it is conceivable that the effects L-DOPA are not solely attributable to the transformation into dopamine and subsequent postsynaptic activation of dopaminergic receptors (De Deurwaerdere et al., 2017). Moreover, changes in dopaminergic transmission induced by L-DOPA in other brain regions such as the prefrontal cortex are not well understood to date. Second, we observed that L-DOPA made decisions more independent of the current evaluation of options, which would be expected to impair task performance. In contrast, we found that L-DOPA tended to increase the number of wins. Thus, it remains to be determined if L-DOPA facilitates exploratory behavior to an extent that is even advantageous in environments characterized by volatile reinforcement schedules. Third, although we used a well-established two-stage decision task that has been designed to track model-based control of behavior, the task was not optimized to study trade-offs between exploration and exploitation. Thus, our results suggesting that L-DOPA facilitates exploration should be seen as preliminary evidence in line with the thrift hypothesis that call for further extension and replication in a more diverse set of reinforcement learning tasks optimized to track explorative behavior as well.

To conclude, we found that increases in dopamine tone lead to decreases in reflexive MF control of behavior by reward while deliberative MB control, which takes the learned transition structure of the environment into account, and value-based learning remained unaffected on average. These results suggest that L-DOPA reduces the transfer of learned value to action, which may result from the reduced local change in dopamine induced by phasic dopamine release when tonic levels are heightened. Our observations corroborate recent value (Hamid et al., 2016) and thrift (Beeler, 2012) theories of dopamine function pointing to an essential role of dopamine in supporting the invigoration of responses and energy expenditure (Kroemer and Small, 2016). As a result, heightened dopamine tone may allow an individual to break reflexive and habit-like chains of action, which have been established because of their previous reward history, and support exploration of available options in the environment (Beeler, 2012; Beeler et al., 2012). Hence, behavioral flexibility may arise as a consequence of increased independence of future behavior on the preceding stream of success.

Author contributions

MNS and TG were responsible for the study concept and design. NBK was responsible for running the study, YL coordinated and ran the data collection. SP set up the computational modeling. NBK performed the data analysis and YL and SP contributed to analyses. NBK wrote the manuscript. All authors contributed to the interpretation of findings, provided critical revision of the manuscript for important intellectual content and approved the final version for publication.

Financial disclosure

The authors declare no competing financial interests.

Acknowledgement

We thank Annika Kienast, Valerie Fournes, Simone Burschka, Elisabeth Kiese, and Monique Zenker for help with data acquisition as well as Michael Marxen and Dirk Müller for support in preprocessing fMRI data. Furthermore, we thank Daniel Schad for insightful discussions regarding data analysis. The study was supported by the Deutsche Forschungsgemeinschaft, grants SFB 940/1, 940/2. Nils Kroemer received salary support by the University of Tübingen's fortüne program, grant #2453-0-0.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2018.10.075.

References

Balleine, B.W., Daw, N.D., O'Doherty, J.P., 2008. Multiple forms of value learning and the function of dopamine. Neuroeconomics: Decis. Making Brain 36, 7–385.
Beeler, J.A., 2012. Thorndike's law 2.0: dopamine and the regulation of thrift. Front. Neurosci. 6, 116.
Beeler, J.A., Daw, N., Frazier, C.R., Zhuang, X., 2010. Tonic dopamine modulates exploitation of reward learning. Front. Behav. Neurosci. 4, 170.
Beeler, J.A., Frazier, C.R., Zhuang, X., 2012. Putting desire on a budget: dopamine and energy expenditure, reconciling reward and resources. Front. Integr. Neurosci. 6, 49.
Beeler, J.A., Faust, R.P., Turkson, S., Ye, H., Zhuang, X., 2016. Low dopamine D2 receptor increases vulnerability to obesity via reduced physical activity, not increased appetitive motivation. Biol. Psychiatry 79, 887–897.
Beierholm, U., Guitart-Masip, M., Economides, M., Chowdhury, R., Duzel, E., Dolan, R., Dayan, P., 2013. Dopamine modulates reward-related vigor. Neuropsychopharmacology 38, 1495–1503.
Brainard, D.H., 1997. The psychophysics toolbox. Spat. Vis. 10, 433–436.
Collins, A.G., Frank, M.J., 2014. Opponent actor learning (OpAL): modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. Psychol. Rev. 121, 337–366.
Collins, A.G., Frank, M.J., 2016. Surprise! Dopamine signals mix action, value and error. Nat. Neurosci. 19, 3–5.
Cools, R., Barker, R.A., Sahakian, B.J., Robbins, T.W., 2001. Enhanced or impaired cognitive function in Parkinson's disease as a function of dopaminergic medication and task demands. Cerebr. Cortex 11, 1136–1143.
Cools, R., Lewis, S.J., Clark, L., Barker, R.A., Robbins, T.W., 2007. L-DOPA disrupts activity in the nucleus accumbens during reversal learning in Parkinson's disease. Neuropsychopharmacology 32, 180–189.
R Core Team, 2015. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Crevoisier, C., Zerr, P., Calvi-Gries, F., Nilsen, T., 2003. Effects of food on the pharmacokinetics of levodopa in a dual-release formulation. Eur. J. Pharm. Biopharm. 55, 71–76.
Daw, N.D., Niv, Y., Dayan, P., 2005. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat. Neurosci. 8, 1704–1711.
Daw, N.D., Gershman, S.J., Seymour, B., Dayan, P., Dolan, R.J., 2011. Model-based influences on humans' choices and striatal prediction errors. Neuron 69, 1204–1215.
Deserno, L., Huys, Q.J., Boehme, R., Buchert, R., Heinze, H.J., Grace, A.A., Dolan, R.J., Heinz, A., Schlagenhauf, F., 2015a. Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. Proc. Natl. Acad. Sci. U. S. A. 112, 1595–1600.
Deserno, L., Wilbertz, T., Reiter, A., Horstmann, A., Neumann, J., Villringer, A., Heinze, H.J., Schlagenhauf, F., 2015b. Lateral prefrontal model-based signatures are reduced in healthy individuals with high trait impulsivity. Transl. Psychiatry 5, e659.

De Deurwaerdere, P., Di Giovanni, G., Millan, M.J., 2017. Expanding the repertoire of L-DOPA's actions: a comprehensive review of its functional neurochemistry. Prog. Neurobiol. 151, 57–100.
Doll, B.B., Bath, K.G., Daw, N.D., Frank, M.J., 2016. Variability in dopamine genes dissociates model-based and model-free reinforcement learning. J. Neurosci. 36, 1211–1222.
Eppinger, B., Walter, M., Heekeren, H.R., Li, S.C., 2013. Of goals and habits: age-related and individual differences in goal-directed decision-making. Front. Neurosci. 7, 253.
Fellows, I., 2012. Deducer: a data analysis GUI for R. J. Stat. Software 49, 1–15.
Frank, M.J., Seeberger, L.C., O'Reilly, R.C., 2004. By carrot or by stick: cognitive reinforcement learning in parkinsonism. Science 306, 1940–1943.
Fuhrer, H., Kupsch, A., Halbig, T.D., Kopp, U.A., Scherer, P., Gruber, D., 2014. Levodopa inhibits habit-learning in Parkinson's disease. J. Neural. Transm. 121, 147–151.
Gershman, S.J., Markman, A.B., Otto, A.R., 2014. Retrospective revaluation in sequential decision making: a tale of two systems. J. Exp. Psychol. Gen. 143, 182–194.
Gläscher, J., Daw, N., Dayan, P., O'Doherty, J.P., 2010. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron 66, 585–595.
Gorgolewski, K., Burns, C.D., Madison, C., Clark, D., Halchenko, Y.O., Waskom, M.L., Ghosh, S.S., 2011. Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. Front. Neuroinf. 5, 13.
Grace, A.A., 1991. Phasic versus tonic dopamine release and the modulation of dopamine system responsivity: a hypothesis for the etiology of schizophrenia. Neuroscience 41, 1–24.
Guitart-Masip, M., Economides, M., Huys, Q.J., Frank, M.J., Chowdhury, R., Duzel, E., Dayan, P., Dolan, R.J., 2014. Differential, but not opponent, effects of L -DOPA and citalopram on action learning with reward and punishment. Psychopharmacology 231, 955–966.
Hamid, A.A., Pettibone, J.R., Mabrouk, O.S., Hetrick, V.L., Schmidt, R., Vander Weele, C.M., Kennedy, R.T., Aragona, B.J., Berke, J.D., 2016. Mesolimbic dopamine signals the value of work. Nat. Neurosci. 19, 117–126.
Harun, R., Hare, K.M., Brough, M.E., Munoz, M.J., Grassi, C.M., Torres, G.E., Grace, A.A., Wagner, A.K., 2015. Fast-scan cyclic voltammetry demonstrates that L-DOPA produces dose-dependent regionally selective, bimodal effects on striatal dopamine kinetics in vivo. J. Neurochem. 136, 1270–1283.
Kayser, A.S., Mitchell, J.M., Weinstein, D., Frank, M.J., 2015. Dopamine, locus of control, and the exploration-exploitation tradeoff. Neuropsychopharmacology 40, 454–462.
Kleiner, M., Brainard, D.H., Pelli, D., Ingling, A., Murray, R., Broussard, C., 2007. What's new in Psychtoolbox-3. Perception 36, 1.
Kroemer, N.B., Small, D.M., 2016. Fuel not fun: reinterpreting attenuated brain responses to reward in obesity. Physiol. Behav. 162, 37–45.
Kroemer, N.B., Guevara, A., Ciocanea Teodorescu, I., Wuttig, F., Kobiella, A., Smolka, M.N., 2014. Balancing reward and work: anticipatory brain activation in NAcc and VTA predict effort differentially. Neuroimage 102 Pt 2, 510–519.
Lee, Y., Kroemer, N.B., Oehme, L., Beuthien-Baumann, B., Goschke, T., Smolka, M.N., 2018. Lower Dopamine Tone in the Striatum Is Associated with Higher Body Mass Index. European Neuropsychopharmacology.
Mannella, F., Gurney, K., Baldassarre, G., 2013. The nucleus accumbens as a nexus between values and goals in goal-directed behavior: a review and a new hypothesis. Front. Behav. Neurosci. 7, 135.
Niv, Y., Daw, N.D., Joel, D., Dayan, P., 2007. Tonic dopamine: opportunity costs and the control of response vigor. Psychopharmacology 191, 507–520.
O'Doherty, J.P., Hampton, A., Kim, H., 2007. Model-based fMRI and its application to reward learning and decision making. Ann. N. Y. Acad. Sci. 1104, 35–53.
Otto, A.R., Raio, C.M., Chiang, A., Phelps, E.A., Daw, N.D., 2013. Working-memory capacity protects model-based learning from stress. Proc. Natl. Acad. Sci. U. S. A. 110, 20941–20946.
Pessiglione, M., Seymour, B., Flandin, G., Dolan, R.J., Frith, C.D., 2006. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. Nature 442, 1042–1045.
Pötschke-Langer, M., Kahnert, S., Schaller, K., Viariso, V., 2016. Tabakatlas Deutschland. Deutsches Krebsforschungszentrum.
Raudenbush, S.W., Bryk, A.S., 2002. Hierarchical Linear Models: Applications and Data Analysis Methods. Sage.
Raudenbush, S.W., Bryk, A.S., Cheong, Y.F., Congdon, R.T., Du Toit, M., 2011. HLM 7. Lincolnwood. Scientific Software International Inc, IL.
Rigoli, F., Chew, B., Dayan, P., Dolan, R.J., 2016a. The dopaminergic midbrain mediates an effect of average reward on pavlovian vigor. J. Cognit. Neurosci. 1–15.
Rigoli, F., Rutledge, R.B., Chew, B., Ousdal, O.T., Dayan, P., Dolan, R.J., 2016b. Dopamine increases a value-independent gambling propensity. Neuropsychopharmacology 41, 2658–2667.
Schultz, W., 2013. Updating dopamine reward signals. Curr. Opin. Neurobiol. 23, 229–238.
Schultz, W., 2015. Neuronal reward and decision signals: from theories to data. Physiol. Rev. 95, 853–951.
Schultz, W., Dayan, P., Montague, P.R., 1997. A neural substrate of prediction and reward. Science 275, 1593–1599.
Sebold, M., Deserno, L., Nebe, S., Schad, D.J., Garbusow, M., Hagele, C., Keller, J., Junger, E., Kathmann, N., Smolka, M.N., Rapp, M.A., Schlagenhauf, F., Heinz, A., Huys, Q.J., 2014. Model-based and model-free decisions in alcohol dependence. Neuropsychobiology 70, 122–131.
Sebold, M., Schad, D.J., Nebe, S., Garbusow, M., Junger, E., Kroemer, N.B., Kathmann, N., Zimmermann, U.S., Smolka, M.N., Rapp, M.A., Heinz, A., Huys, Q.J., 2016. Don't think, just feel the music: individuals with strong Pavlovian-to-instrumental transfer effects rely less on model-based reinforcement learning. J. Cognit. Neurosci. 1–11.
Sharp, M.E., Foerde, K., Daw, N.D., Shohamy, D., 2016. Dopamine selectively remediates 'model-based' reward learning: a computational approach. Brain 139, 355–364.
Silvetti, M., Seurinck, R., Verguts, T., 2013. Value and prediction error estimation account for volatility effects in ACC: a model-based fMRI study. Cortex 49, 1627–1635.
Smittenaar, P., FitzGerald, T.H., Romei, V., Wright, N.D., Dolan, R.J., 2013. Disruption of dorsolateral prefrontal cortex decreases model-based in favor of model-free control in humans. Neuron 80, 914–919.
Stauffer, W.R., Lak, A., Schultz, W., 2014. Dopamine reward prediction error responses reflect marginal utility. Curr. Biol. 24, 2491–2500.
Steinberg, E.E., Keiflin, R., Boivin, J.R., Witten, I.B., Deisseroth, K., Janak, P.H., 2013. A causal link between prediction errors, dopamine neurons and learning. Nat. Neurosci. 16, 966–973.
Sutton, R.S., Barto, A.G., 1998. Reinforcement Learning: an Introduction. MIT press, Cambridge.
Syed, E.C., Grima, L.L., Magill, P.J., Bogacz, R., Brown, P., Walton, M.E., 2016. Action initiation shapes mesolimbic dopamine encoding of future rewards. Nat. Neurosci. 19, 34–36.
Verguts, T., Vassena, E., Silvetti, M., 2015. Adaptive effort investment in cognitive and physical tasks: a neurocomputational model. Front. Behav. Neurosci. 9, 57.
Wittchen, H., et al., 1997. SCID-I: Structured Clinical Interview for DSM-IV Disorders. Hogrefe, Goettingen, Germany.
Wunderlich, K., Smittenaar, P., Dolan, R.J., 2012. Dopamine enhances model-based over model-free choice behavior. Neuron 75, 418–424.
Yarkoni, T., Poldrack, R.A., Nichols, T.E., Van Essen, D.C., Wager, T.D., 2011. Large-scale automated synthesis of human functional neuroimaging data. Nat. Methods 8, 665–670.
Zhang, L., Doyon, W.M., Clark, J.J., Phillips, P.E., Dani, J.A., 2009. Controls of tonic and phasic dopamine transmission in the dorsal and ventral striatum. Mol. Pharmacol. 76, 396–404.
Zhang, L., Dong, Y., Doyon, W.M., Dani, J.A., 2012. Withdrawal from chronic nicotine exposure alters dopamine signaling dynamics in the nucleus accumbens. Biol. Psychiatry 71, 184–191.