Behavioral/Cognitive

Reversal Learning and Dopamine: A Bayesian Perspective

Vincent D. Costa, Valery L. Tran, Janita Turchi, and Bruno B. Averbeck
Laboratory of Neuropsychology, National Institute of Mental Health, National Institutes of Health, Bethesda Maryland 20892-4415

Reversal learning has been studied as the process of learning to inhibit previously rewarded actions. Deficits in reversal learning have been seen after manipulations of dopamine and lesions of the orbitofrontal cortex. However, reversal learning is often studied in animals that have limited experience with reversals. As such, the animals are learning that reversals occur during data collection. We have examined a task regime in which monkeys have extensive experience with reversals and stable behavioral performance on a probabilistic two-arm bandit reversal learning task. We developed a Bayesian analysis approach to examine the effects of manipulations of dopamine on reversal performance in this regime. We find that the analysis can clarify the strategy of the animal. Specifically, at reversal, the monkeys switch quickly from choosing one stimulus to choosing the other, as opposed to gradually transitioning, which might be expected if they were using a naive reinforcement learning (RL) update of value. Furthermore, we found that administration of haloperidol affects the way the animals integrate prior knowledge into their choice behavior. Animals had a stronger prior on where reversals would occur on haloperidol than on levodopa (L-DOPA) or placebo. This strong prior was appropriate, because the animals had extensive experience with reversals occurring in the middle of the block. Overall, we find that Bayesian dissection of the behavior clarifies the strategy of the animals and reveals an effect of haloperidol on integration of prior information with evidence in favor of a choice reversal.

Key words: **Bayesian; dopamine; haloperidol; L-DOPA; reinforcement learning; reversal learning**

Introduction

Reversal learning has been studied extensively as a behavior that indexes important fundamental neural processes that underlie inhibiting previously rewarded actions. For example, previous studies examined the effect of dopamine manipulations on reversal learning. This work has shown that Parkinson's disease patients have deficits in reversal learning on medication, but not off (Cools et al., 2001, 2006; Graef et al., 2010) and that manipulating dopamine D_2 receptors in the striatum can also lead to deficits (Mehta et al., 2001; Clarke et al., 2011). Similarly, studies in animals (Jones and Mishkin, 1972; Dias et al., 1996; Chudasama and Robbins, 2003; Schoenbaum et al., 2003; Izquierdo et al., 2004) and humans (Fellows and Farah, 2003; Hornak et al., 2004) show that damage to the orbitofrontal cortex leads to deficits in reversal learning.

Whether a particular neural process contributes to reversal learning may depend on how experienced subjects are with the task (Rygula et al., 2010). Most studies on reversal learning used tasks in which the subjects had limited or no experience with reversals in stimulus reward mappings before the period of data collection. In these tasks, subjects often improve as they experience more reversals (Clarke et al., 2011; Rudebeck et al., 2013). Before experiencing reversals, subjects probably assume that stimulus reward mappings are relatively stable, and the first few reversals violate this assumption. Such violations have been referred to as unexpected uncertainty (Yu and Dayan, 2005). Therefore, the behavioral performance during the first few reversals likely reflects the interplay of two processes. First, a naive model the subjects apply to the task before they learn that stimulus reward mappings change with some regularity. Second, a process whereby the subjects are learning a model of the environment that allows for reversals in stimulus reward mappings. Therefore, deficits in these tasks could reflect problems with either process. As subjects gain experience with reversals, the second process will come to dominate and the reversal in the reward mapping will become an expected uncertainty. The subject does not know when a reversal will occur, but they expect it to occur.

In the present study, we examined reversal behavior after extensive training on a reversal learning task. Therefore, reversals are expected. It is only when they will occur that is uncertain. There are two advantages of studying this regime. First, effects of manipulations can be attributed to problems applying the model to the task, as opposed to problems with learning the model. Second, we can develop a Bayesian framework for analyzing choice behavior under the assumption that the animals have learned the statistics of the task. To demonstrate the utility of this approach, we examined the effect of two dopamine manipulations—systemic administration of L-DOPA or haloperidol—shown previously to have disparate effects on reversal learning. We find that, in this regime, there are no effects of dopamine manipulations on learning. However, we do find that administration of haloperidol leads to an increased reliance on prior beliefs about when a reversal will occur.

Received May 15, 2014; revised Dec. 2, 2014; accepted Dec. 20, 2014.

Author contributions: V.D.C., J.T., and B.B.A. designed research; V.D.C. and V.L.T. performed research; V.D.C. and B.B.A. analyzed data; V.D.C. and B.B.A. wrote the paper.

This work was supported by the Intramural Research Program of the National Institute of Mental Health.

Correspondence should be addressed to Dr. Bruno B. Averbeck, Laboratory of Neuropsychology, National Institute of Mental Health/National Institutes of Health, Building 49, Room 1B80, 49 Convent Drive, MSC 4415, Bethesda, MD 20892-4415. E-mail: bruno.averbeck@nih.gov.

DOI:10.1523/JNEUROSCI.1989-14.2015

 0270-6474/15/352407-10$15.00/0

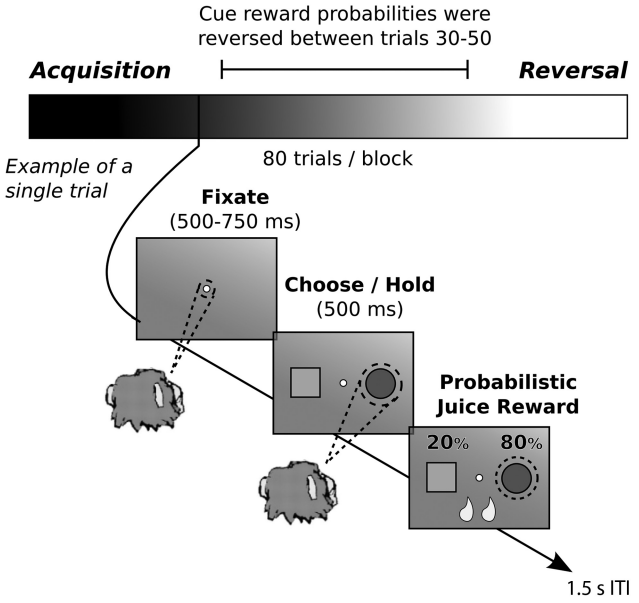

Figure 1. Trial structure of a single block and the sequence of events in a single trial of the two-arm bandit reversal learning task. Each block contained 80 trials. The stimulus reward mapping was reversed on a randomly chosen trial between trials 30 and 50. Trials before the reversal are referred to as acquisition, and trials after the reversal are referred to as reversal. The reward schedule was always constant within a block (i.e., 80/20, 70/30, or 60/40%), but it usually changed across blocks. ITI, Intertrial interval.

Materials and Methods

Subjects. Three male rhesus monkeys (*Macaca mulatta*), aged 5–6 years with weights ranging from 6.5 to 9.3 kg, were studied. All monkeys were placed on water control for the duration of the study and, on test days, earned all of their fluid through performance on the task. Experimental procedures for all monkeys were performed in accordance with the National Institutes of Health *Guide for the Care and Use of Laboratory Animals* and were approved by the Animal Care and Use Committee of the National Institute of Mental Health.

Experimental setup. The monkeys completed 4–44 (20.93 ± 0.93, mean ± SE) blocks per session of a two-arm bandit problem. Each block consisted of 80 trials and involved a single reversal of the stimulus–reward contingencies (Fig. 1). On each trial, the monkeys had to first acquire and hold a central fixation point (250–750 ms). After the monkey fixated for the required duration, two stimuli appeared to the left and right (6° visual angle) of the central fixation point. Stimuli varied in shape and color, and stimulus location (left vs right for each shape) was randomized within a block. Monkeys chose between stimuli by making a saccade to one of the two stimuli and fixating the cue for a minimum of 500 ms. One of the stimuli had a high reward probability, and one had a low reward probability. Juice rewards were probabilistically delivered at the end of each trial, followed by a fixed 1.5 s intertrial interval. A failure to acquire/hold central fixation or to make a choice within 750 ms resulted in a repeat of the previous trial. The three reward schedules used were 80/20%, 70/30%, and 60/40%. Use of these three reward schedules anticorrelates the mean reward probabilities of the bandit arms. The trial on which the cue–reward mapping reversed within each block was selected pseudorandomly from a uniform distribution across trials 30–50. The reversal trial did not depend on the monkey reaching a performance criterion. Reward schedules were always constant within a block but could (and usually did) change across blocks.

Stimuli consisted of simple images of a circle and square in one of three colors (red, green, and blue). The two choice options always differed in color and shape. This resulted in six unique stimulus combinations. When these combinations were crossed with the three reward schedules and whether a particular shape was more or less initially rewarding (e.g., whether the blue square was the best choice before or after the reversal), this resulted in 36 block combinations. Block presentations were fully randomized without replacement. This ensured that a specific stimulus–reward combination was never repeated directly until all 36 block combinations were experienced (<4% of sessions). Although combinations were potentially repeated across sessions, during inspection, there was no evidence of improved performance across sessions.

Each monkey received 10–14 d of initial training on the described reversal learning task until they were routinely completing 15–20 blocks per session. Animals first learned the structure of the task under a deterministic reward schedule. Probabilistic reward schedules were then introduced progressively until the animals exhibited stable performance on the tested reward schedules.

Stimulus presentation and behavioral monitoring were controlled by a personal computer running the Monkeylogic (version 1.1) MATLAB toolbox (Asaad and Eskandar, 2008). Eye movements were monitored using an Arrington Viewpoint eye-tracking system (Arrington Research) and sampled at 1 kHz. Stimuli were displayed on an LCD monitor (1024 × 768 resolution) situated 40 cm from the monkey's eyes. On rewarded trials, 0.085 ml of apple juice was delivered through a pressurized plastic tube gated by a computer-controlled solenoid valve (Mitz, 2005).

Drug administration. Before drug testing, monkeys were first habituated to intramuscular needle injections of saline given in conjunction with free juice (pH 7.4, 0.1 ml/kg). After this habituation period, monkeys readily presented their leg for injections. At the start of each placebo session—while chaired and outside of the test box—the monkeys received an intramuscular injection of saline (1 ml) while they drank 6 ml of apple juice from a plastic syringe. They were then head posted and placed inside the test box. The eye-tracking system was then calibrated to avoid drug-related effects on eye-tracking sensitivity. During the remainder of the wait period, the animals viewed a nature movie. This placebo procedure was consistent with the two methods of drug administration. Free juice was similarly delivered at the start of each drug session before waiting 30 min to start the task. On days the monkeys received L-DOPA, we dissolved, under sonication, a pulverized fixed dose tablet of L-DOPA (100 mg/25 mg carbidopa; Actavis) into the delivered free juice and paired it with an intramuscular injection of saline. On days the monkeys received haloperidol, free juice was delivered in conjunction with an intramuscular injection of haloperidol (6.5 μg/kg; Bedford Laboratories). This dose was consistent with doses shown previously to have behavioral effects (Turchi et al., 2010). Injections were prepared by first dissolving a fixed dose of haloperidol (100 μg) under sonication into PBS under strict sterile conditions and stored at 4°C for use within the week. On the day of the drug injections, aliquots were resonicated and allowed to reach room temperature before injection. Injections were given intramuscularly into the lateral hindlimb.

The monkeys completed multiple sessions under each drug condition. On L-DOPA, monkey E completed seven sessions comprising 138 total blocks, monkey G completed six sessions comprising 159 total blocks, and monkey M completed seven sessions comprising 193 total blocks. On haloperidol, monkey E completed seven sessions comprising 100 total blocks, monkey G completed eight sessions comprising 142 total blocks, and monkey M completed seven sessions comprising 143 total blocks. The total number of placebo sessions ranged from 15 to 24 sessions per animal (22 for E, 24 for G, and 16 for M), comprising 370–479 blocks. Haloperidol sessions were spaced a minimum of 7 d apart to facilitate washout, whereas the faster clearance of L-DOPA permitted a minimum spacing of 3 d between sessions. L-DOPA and haloperidol sessions were interleaved and counterbalanced for the day of the week to minimize routine caretaking effects on behavior. Each drug session was preceded by at least one placebo session, and all placebo sessions lagged the most recent drug session by a minimum of 2 d to minimize carryover effects.

Bayesian models. We fit three Bayesian models that estimated the posterior probability that reversals occurred on each trial, under various assumptions. To estimate the models, we fit a likelihood function given by the following:

$$f(x, y|r, p, h, M) = \prod_{k=1}^{T} q(k), \quad (1)$$

where r is the trial on which the reward mapping reversed ($r \in 0-81$), and p is the probability of reward for the high reward option (models 1 and 3) or the consistency with which the animals chose their preferred option (model 2). The variable h encodes whether option 1 or option 2 begins the block as the high reward option ($h \in 1, 2$), k indexes trial number in the block, and T is the current trial. The variable r ranges from 0 to 81 because we allowed the model to assume that reversals occurred before the block started or after the block ended. In either of these cases, there would be no switch within the block (the model estimated no reversal in <1% of the total blocks analyzed), and the posterior probability of a switch would be equally weighted for r equal to 0 or 81. The data are given by the vectors x and y, where the elements of x are the rewards ($x_i \in 0, 1$), and the elements of y are the choices ($y_i \in 1, 2$) in trial i. We fit three variants of this model indicated by M ($M \in 1, 2, 3$). $M = 1$ is the ideal observer. This model was used to estimate the evidence the animal had available to it when it made its decisions, as well as the ideal reversal point. $M = 2$ is the behavioral choice model. This model was used to estimate where the animal reversed its choice behavior. The third model, $M = 3$, is similar to the ideal observer except we parameterized a prior over reversals to better fit the animal's choice behavior instead of using the generative prior (used in model 1), which only allowed reversals during trials 30–50.

The behavioral choice model ($M = 2$) estimates the trial on which the animals switched their choice behavior. This only depends on the pattern of choices, not on whether they were rewarded. We assumed that the animal had a stable choice preference that switched at some point in the block from one stimulus to the other. Given the choice preference, the animals occasionally chose the wrong stimulus (i.e., the stimulus inconsistent with their choice preference) at some lapse rate $1 - p$. Thus, for $k < r$ and $h = 1$, choosing option 1, $q(k) = p$; and choosing option 2, $q(k) = 1 - p$. For $k \geq r$ and $h = 1$, choosing option 1, $q(k) = 1 - p$; and choosing option 2, $q(k) = p$. Correspondingly, for $k < r$ and $h = 2$, choosing option 2, $q(k) = p$, etc. Thus, this model assumed that the monkey preferred one option before switching and preferred the other option after switching. It most often chose its preferred option ($p > 0.5$), but it occasionally chose the wrong target perhaps as a result of lapses in attention. For all reported analyses, we marginalized over the correct choice rate p. Therefore, we assumed that the animals were maximizing and not doing probability matching. These values for $q(k)$ were filled in for the entire block, because we were performing this analysis *post hoc* to estimate where the animal reversed.

For models 1 and 3, we estimated whether a reversal had occurred conditioning only on outcomes before the current trial, T. (Models 1 and 3 have different priors on the reversal trial, defined below, but identical likelihood functions.) This provided an estimate of the information on which the animal was making its choice. For these models, values of $q(k)$ for each schedule were given by the following mappings from choices to outcomes. For $k < r$ and $h = 1$ (before reversal and target 1 is the high probability target), choose 1 and get rewarded $q(k) = p$; choose 1 and not get rewarded, $q(k) = 1 - p$; choose 2 and get rewarded, $q(k) = 1 - p$; and choose 2 and not get rewarded, $q(k) = p$. For $k \geq r$, these probabilities flip. Correspondingly, for $k < r$ and $h = 2$, the probabilities are also complimented. These values were filled in up to the current trial, T.

Given these mappings for $q(k)$, we could then calculate the likelihood as a function of r, p, and h for each block of trials. The posterior is given by the following:

$$p(r, p, h|x, y, M) = f(x, y|r, p, h, M)p(r|M)p(p, h|M)/p(x, y|M). \quad (2)$$

The priors on p and h were flat for all models. The prior on r, $p(r|M)$, varied by model. For model 1, the prior was given by the generative model for the data, which only allowed reversals for trials 30–50. Specifically, for $r <$ 30 or $r > 50$, $p(r|M = 1) = 0$. For $r > 29$ or $r < 51$, $p(r|M = 1) = 1/21$. Using this prior, there was general agreement between the ideal observer estimate of the reversal point and the actual programmed reversal point (mean ± SE session, $r = 0.44 \pm 0.16$, $t_{(87)} = 3.5$, $p < 0.001$). For model 2, the prior was flat on $r \in 0-81$, $p(r|M = 2) = 1/82$. For model 3, results are presented with either a flat prior [i.e., $p(r|M = 3) = 1/82$] or a Gaussian prior, given by the following:

$$p(r|M = 3) \propto \exp\left(-\frac{(r - \mu_d)^2}{2\sigma^2}\right). \quad (3)$$

Model 1 always used the flat prior restricted to have support over trials 30–50, whereas model 3 used either a flat prior with support over trials 0–81 or a parameterized Gaussian prior given by Equation 3. The different priors for model 3 are indicated explicitly in Results, and they are used to test hypotheses about the strategy the animals used to solve the task.

Given the priors, the posterior over switch trial could be calculated by marginalizing over p and h. Specifically,

$$p(r|x, y, M) = \Sigma_{p,h}\, p(r, p, h|x, y, M). \quad (4)$$

Similarly, the posterior over the probability of reward for the high probability option could be calculated by marginalizing over r and h:

$$p(p|x, y, M) = \Sigma_{r,h}\, p(r, p, h|x, y, M). \quad (5)$$

After the posterior over r for the behavioral choice ($M = 2$) was calculated, the expected reversal point was calculated as $<r|M = 2> = \sum_{r=0}^{81} r\, p(r|x, y, M = 2)$. Because the estimated reversal point was not guaranteed to be an integer, it was rounded to the nearest integer when it served as an index of summation. Trials less than $\langle r|M\rangle$ were assigned to the acquisition phase, whereas trials greater than or equal to $\langle r|M\rangle$ were assigned to the reversal phase.

To calculate a point estimate of the reversal using the ideal observer ($M = 1$), we first calculated the posterior evidence that a reversal had occurred before trial k, as follows:

$$p(r < k|x, y, M = 1) = \sum_{i=1}^{k-1} p(r = i|x, y, M = 1). \quad (6)$$

This evidence was then compared with a threshold, and the reversal trial was defined as the first trial in which the evidence exceeded the threshold, i.e., $p(r < k|x, y, M = 1) > \theta$. To compute a point estimate of the reversal trial, we assumed a distribution over thresholds, uniform on 0.51–0.99, and computed an expectation over this distribution. Thus,

$$\langle r = k|M = 1\rangle := \langle[\min(k)|p(r < k|x, y, M = 1) > \theta]\rangle_{p(\theta)}. \quad (7)$$

For model 3, we inferred the parameters of the Gaussian prior to best fit the animal's choice behavior. When we estimated these priors, we calculated the posterior evidence available to the animal when it switched, $p(r = k|x, y, M = 3)$. The animal's switch trial was given by $k = \langle r|M = 2\rangle$. This is the posterior evidence that a reversal occurred on the trial on which the animal reversed. Note that the data plotted in Figure 3, *A* and *B*, are from model 3 with a flat prior on trials 0–81. This shows the model evidence available to the animal, under an uninformative prior. When we fit the Gaussian prior for model 3, we maximized the posterior evidence conditioned on μ_d and σ^2. Thus, we maximized:

$$f(x, y|\mu_d, \sigma^2) = \Pi_N\, p(r = k - 1|x, y, M = 3), \quad (8)$$

where the product is over N blocks, with one value for each block and $k = \langle r|M = 2\rangle$.

We tested statistically whether there was a significant effect of the prior on the animals' choice behavior. We did this in two ways. First, we fit separate priors to each schedule in each session and compared the mean, μ_d, and SD, σ, of the fitted priors across schedules and drug conditions with a mixed-effects model. Specifically, either the mean or the SD was the dependent variable. We then specified drug and reward schedule as fixed effects with session specified as a random variable nested in monkey. Second, we used likelihood ratio test statistics to compare three models: (1) a model that had a flat prior over trials 0–81; (2) a model that fit an individual Gaussian to each session (two parameters per session); and (3) a model that fit an individual Gaussian to each schedule in each session (six parameters per session).

Reinforcement learning (RL) model. We first split the trials in each block into acquisition and reversal phases using the expected reversal trial calculated with model 1 and model 2. We then fit separate reinforcement learning models to each phase. This was done in each session and separately for each schedule. Thus, four reinforcement learning models were fit to each schedule (i.e., acquisition defined by model 1, acquisition defined by model 2, reversal defined by model 1, and reversal defined by model 2). We used a standard reinforcement learning model to estimate learning from positive and negative feedback, as well as the inverse temperature. Specifically, value updates were given by the following:

$$v_i(k+1) = v_i(k) + \rho_f(R - v_i(k)), \quad (9)$$

where v_i is the value estimate for option i, R is the reward feedback for the current choice, and ρ_f is the learning rate parameter, where f is either positive or negative. In other words, we fit different values of ρ for positive ($R = 1$) versus negative ($R = 0$) feedback. These value estimates were then passed through a logistic function to generate choice probability estimates:

$$d_1(k) = \left(1 + e^{\alpha(v_2(k) - v_1(k))}\right)^{-1}, \; d_2(k) = 1 - d_1(k). \quad (10)$$

The likelihood is then given by the following:

$$f(D|\alpha, \rho_{\text{pos}}, \rho_{\text{neg}}) = \Pi_k[d_1(k)c_1(k) + d_2(k)c_2(k)], \quad (11)$$

Where $c_1(k)$ had a value of 1 if option 1 was chosen on trial k and $c_2(k)$ had a value of 1 if option 2 was chosen. Otherwise, they had a value of 0. Standard function optimization techniques were used to maximize the log of the likelihood of the data given the parameters. Because the estimation can settle on local minima, we used 100 initial values for the parameters. The maximum of the log likelihood across fits was then used. When we fit the RL model to the acquisition phase, starting values were reset to 0.5 at the beginning of each block, whereas for the reversal phase, starting values for each block corresponded to end values of the corresponding acquisition phase.

Classical statistics. The final statistical analyses involved mixed-model ANOVAs. Each dependent variable was entered into full factorial, mixed-effects ANOVA models implemented in MATLAB. Drug, schedule, learning phase, feedback type, and monkey were specified as fixed effects, whereas session nested under monkey and drug was specified as a random effect. *Post hoc* analyses of significant main effects used Fisher's least significant difference test to correct for multiple comparisons, given its desirable properties when testing differences among three groups (Levin et al., 1994). *Post hoc* tests of significant interactions consisted of computing univariate ANOVAs for component effects and similarly correcting for multiple comparisons.

Results

Bayesian analysis of reversal learning

We used a Bayesian framework to estimate, for each block of trials, a posterior distribution over the trial on which an ideal observer (model = 1; see Materials and Methods) detected a reversal in the reward contingencies and also the trial on which the animal's choice behavior (model = 2) reversed. We also developed a model in which we parameterized prior assumptions about where reversals occur and fit these to the animal's choice behavior (model = 3) to infer the priors the animals were using to drive their choices. These algorithms gave us an estimate of where an ideal observer would reverse its choice behavior (model 1), where the animals reversed their choice behavior (model 2), and the prior information the animals used when they reversed their choice behavior (model 3).

We began by examining where the animals reversed on average (model 2) and comparing this with where the ideal observer reversed (model 1). We first examined schedule- and drug-related effects on the average posterior probability distribution over reversals in the monkeys' choice behavior (model 2; Fig. 2*A*,*B*). Posterior probability estimates of the reversal in the monkey's choice behavior were more dispersed for the harder schedules, and the animals also exhibited a tendency to switch earlier on more difficult schedules (e.g., a heavier left tail in the reversal distribution). Next, we examined the effect of drug on these posteriors. Inspection of the posterior probability distributions suggested that the animals were likely to reverse more efficiently on haloperidol or L-DOPA compared with placebo, because these distributions were more peaked than the saline distribution. To quantify effects of reward schedule and drug on the reversal behavior, we computed the expected value of the reversal trial (i.e., a point estimate of the trial on which the switch occurred, with the expectation taken across the posterior) for each block of the animal's choice behavior. We also computed the expected reversal trial for the ideal observer (model 1). The difference between these points quantifies where the monkeys reversed relative to the ideal observer (Fig. 2*C*), whereas the absolute value of the difference between the two estimates quantifies how closely the animals switched their behavior relative to the ideal observer (Fig. 2*D*).

On average, the animals' reversed their choice behavior earlier than expected compared with the ideal observer (mean ± SEM difference, -1.43 ± 0.29, $t_{(83)} = -4.52$, $p < 0.001$), although sufficiently close to the ideal observer estimate (mean ± SE absolute difference, 7.07 ± 0.26) to well approximate the trial window in which the programmed reversals occurred (e.g., trials 30–50). On average and in reference to the ideal observer (Fig. 2*C*), the monkeys reversed their behavior earlier during 60/40% blocks compared with 70/30% ($t_{(89)} = -4.77$, $p < 0.001$) or 80/20% blocks ($t_{(89)} = -8.77$, $p < 0.001$; effect of schedule, $F_{(2,152)} = 10.43$, $p < 0.001$). The accuracy of the animals in reversing their choice behavior relative to the ideal observer also increased with the reward schedule ($F_{(2,149)} = 38.54$, $p < 0.001$). The animals were less accurate in reversing their behavior during 60/40% blocks ($t_{(75)} = 5.06$, $p < 0.001$) and more accurate during 80/20% blocks ($t_{(80)} = 4.4$, $p = 0.014$) compared with their behavior during 70/30% blocks. These schedule effects were consistent across drug conditions (drug × schedule, both F values <1.4, $p > 0.26$).

The average difference between the estimated reversal in the monkeys' behavior and the ideal observer model did not differ by drug (Fig. 2*C*; $F_{(2,36)} < 1$, $p = 0.486$). However, there was a main effect of drug on the absolute deviation of the two reversal points (Fig. 2*D*; $F_{(2,36)} = 8.18$, $p = 0.001$). On haloperidol ($t_{(19)} = 3.66$, $p = 0.001$) or L-DOPA ($t_{(17)} = 2.97$, $p = 0.008$), estimated reversals in monkeys' choice behavior were closer to the reversal points estimated by the ideal observer model than when the monkeys received placebo. Reversal accuracy did not differ under haloperidol versus L-DOPA ($t_{(17)} = 1.65$, $p = 0.11$). Given that haloperidol and L-DOPA led to behavioral switches that were closer to the ideal observer, we reexamined the behavioral choice posterior distributions after aligning to the ideal observer switch trial (Fig. 2*E*). It could be seen that the posterior distribution peaked higher and more quickly under either drug compared with saline.

Next, we examined the evidence on which reversals were based (Fig. 3*A*,*B*). To calculate the evidence, we examined the posterior evidence that a reversal had occurred (model 3 with flat prior) before the trial on which the animals actually reversed their choice behavior (given by model 2: $\langle r|M = 2\rangle$). When the posterior of model 3 was aligned to the estimated reversal in the monkey's choice behavior, it could be seen that switches in the

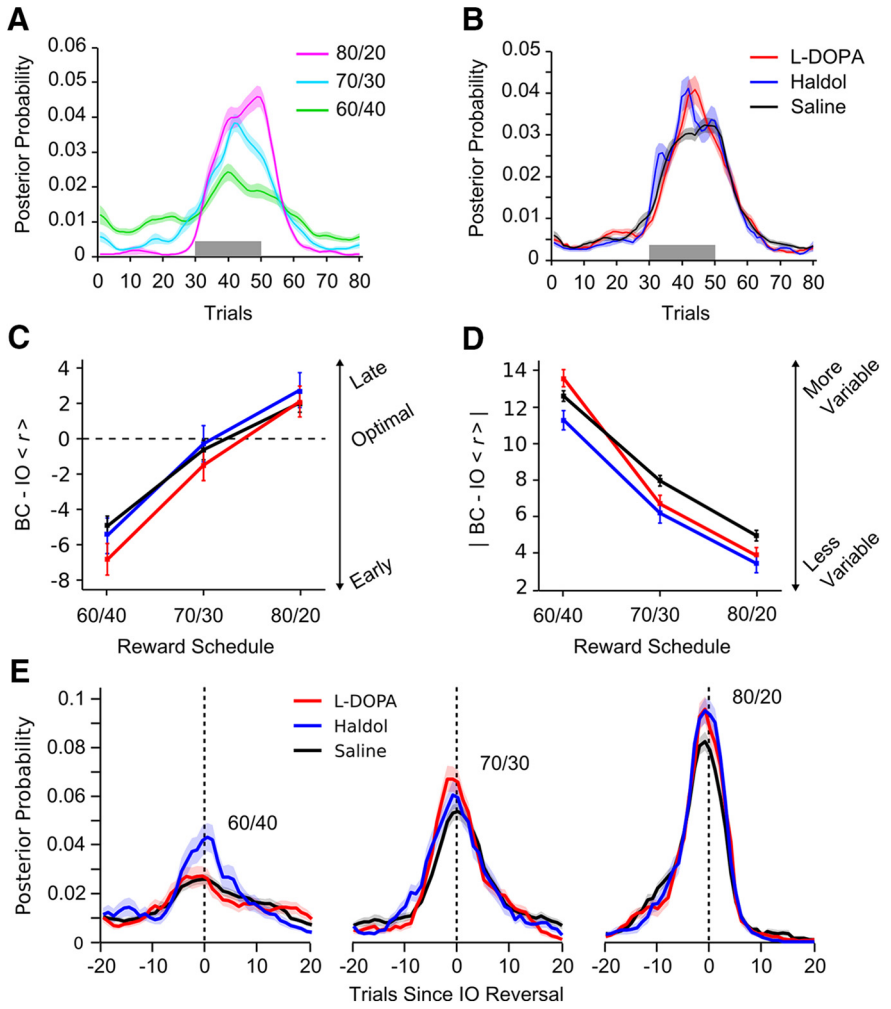

Figure 2. Bayesian estimates of reversal points by reinforcement schedule and drug condition. Error bars and shading indicate 1 SEM, and the gray windows indicate the trial range in which a reversal was programmed to occur. ***A***, The mean posterior probability by schedule that the animal reversed its choice behavior on each trial ($M = 2$; see Materials and Methods). ***B***, Same as ***A***, split out by drug condition. ***C***, Difference in the estimated reversal trial between the behavioral choice (BC; $M = 2$) and ideal observer (IO; $M = 1$) models, broken out by schedule and drug condition. ***D***, Same as ***C*** except for absolute value of difference. ***E***, Behavioral choice posterior distributions averaged after aligning the posterior from individual blocks to the ideal observer switch trial from each block.

animals' choice behavior followed a clear peak in the posterior distribution (Fig. 3*A*,*B*). This is to be expected, because the animals should switch their choice behavior after outcomes that signaled a reversal. Across the three reward schedules, reversals in choice behavior followed successively smaller peaks in the posterior distribution (Fig. 3*A*; schedule, $F_{(2,153)} = 167.31$, $p < 0.001$), consistent with the fact that evidence scales with the difference in the cue reward probabilities. There were no drug-related differences in the peak of the posterior distribution (Fig. 3*B*; drug, $F_{(2,36)} < 1$, $p = 0.862$). Thus, reversals in choice behavior were triggered by a similar feature across drug conditions.

Before examining the priors that the animals may have been using during the task, we characterized the stability of their behavior. The animals had considerable experience with the task. Across animals and conditions, data were collected in 2219 blocks (629–787 blocks per animal). Because the animals had extensive experience, we assumed they had a stable strategy, which included knowledge that reversals occurred in the middle of the block. To validate this assumption, we used the causal model ($M = 3$, flat prior) to compute the amount of evidence available to the monkeys at the time they reversed their behavior for each block and averaged this in each session, separately for each reward schedule. We then linearly regressed the sequential session number against the average accumulated evidence for each reward schedule and monkey. If the monkeys had a stable strategy, we would expect accumulated evidence at reversal to be stable across sessions. Indeed, the number of sessions completed was unrelated to the amount of evidence available at the time the monkeys reversed their behavior (session count, $F_{(1,2)} = 9.43$, $p = 0.09$). Therefore, the evidence on which animals reversed their choice behavior was stable across the period of data collection, and animals were familiar with the fact that a contingency reversal would occur during each block.

We next examined the hypothesis that a prior may have been playing a role in the choice behavior and that this prior was affected by the drug condition and/or reward schedule. The animals would be expected to have such a prior, given their experience with the task as just shown. One choice would be to specify a uniform prior between trials 30 and 50. However, the animals reversed their behavior outside this window in 40.6% of the blocks analyzed. Therefore, we used a Gaussian prior and fit the mean and SD to the data in two ways. The first prior we tested parameterized a single mean and SD per drug session, independent of the reward schedule. A likelihood ratio test indicated a significant improvement in model fit compared with specifying a flat prior, uniform on trials 1–80 (Fig. 3*C*; $\chi^2_{(515)} = 3563.88$, $p < 0.001$). We also tested whether the prior distribution varied in terms of the reward schedule by fitting a second set of Gaussian priors, in which we parameterized, per drug session, separate mean and SDs for each reward schedule. Despite an increase in the number of specified parameters, there was a significant improvement in model fit compared with when we fit a common Gaussian prior across schedules ($\chi^2_{(1236)} = 6819.44$, $p < 0.001$). This suggests that both the reward schedule and drug condition influenced the animals' prior expectation about when a reversal could occur.

A hierarchical comparison of the estimated parameters of this final set of priors allowed us to determine how the mean and SD of the prior distribution varied with the reward schedule and drug condition (Fig. 3*D*). The mean of the prior was unaffected by the reward schedule ($F_{(2,152)} = 2.48$, $p = 0.080$). The SD of the prior was inversely related to the reward schedule (Fig. 3*E*; $F_{(2,152)} = 15.48$, $p < 0.001$). The SD of the prior was smallest for 80/20% blocks ($t_{(71)} = 2.47$, $p < 0.017$) and largest for 60/40% blocks ($t_{(81)} = 2.99$, $p = 0.004$) compared with the SD of the prior for 70/30% blocks. The mean of the prior also did not differ by drug condition (Fig. 3*F*; $F_{(2,37)} = 1.76$, $p = 0.185$). The SD of the prior

did differ by drug ($F_{(2,37)} = 4.96$, $p = 0.012$). Specifically, the SD of the prior was smaller on haloperidol relative to L-DOPA ($t_{(17)} = -2.51$, $p = 0.022$) or saline ($t_{(20)} = -2.0$, $p = 0.048$). The SD of the prior did not differ on L-DOPA versus saline ($t_{(17)} < 1$, $p = 0.483$).

Choice behavior

We next examined the choice behavior of the animals, split by reward schedule and drug condition. We aligned the trials in each block around the reversal point estimated by both the ideal observer (Fig. 4*A*) and behavioral choice estimates of the reversal point (Fig. 4*B*). To quantify how realignment affected assessment of the animals' performance, we fit individual logistic regression models to choice behavior surrounding (±10 trials) the reversal points of each model, grouped by reward schedule and session. We then tested to see whether there was a significant change in the slope coefficient when choice behavior was aligned to each model. When choices were aligned according to the behavioral choice model versus the ideal observer model, there was an increase in the slope coefficient, indicating a faster switch in choice behavior before and after the reversal (model, $F_{(1,28)} = 14.59$, $p < 0.001$). This effect was consistent across the three reward schedules (model × schedule, $F_{(2,67)} < 1$, $p = 0.509$) and drug conditions (model × drug, $F_{(2,52)} = 2.53$, $p = 0.082$), although a direct comparison of the model-related change in slope (e.g., slope for model 2 − model 1) did indicate a larger increase on haloperidol (mean ± SE, 1.05 ± 0.26) compared with either saline (mean ± SE, 0.33 ± 0.17) or L-DOPA (mean ± SE, 0.16 ± 0.27; drug, $F_{(2,46)} = 3.26$, $p = 0.047$). Thus, when the animals detected a reversal in the reward contingencies, they abruptly switched their choice behavior, even in the 60/40% condition, and this was detected using the behavioral choice model. The more gradual slopes seen when the choice data were aligned to the ideal observer reversal point were not attributable to a stochastic sampling period in each block. Rather, they reflected averaging rapid switches across stimuli that occurred at different points in different blocks relative to the ideal observer.

Referencing the animals' choice behavior to an ideal observer (model 1) also allowed us to determine whether the administered drug modulated how monkeys optimized their choice behavior. The monkeys' choice behavior was more optimal (i.e., they selected the option that was most likely to be rewarded on the basis of the past choices and outcomes) on haloperidol ($t_{(20)} = 2.51$, $p = 0.025$) or L-DOPA ($t_{(17)} = 3.17$, $p = 0.025$) compared with saline (drug, $F_{(2,38)} = 6.35$, $p = 0.004$). As a consequence, the percentage of rewards earned per block also differed by drug ($F_{(2,38)} = 4$, $p = 0.026$), with more rewards earned on haloperidol ($t_{(20)} = 2.43$, $p = 0.024$) or L-DOPA ($t_{(17)} = 2.13$, $p = 0.047$) compared with saline. These results are consistent with the decreased variance between the ideal observer and the animal's choice reversals on either drug compared with saline (Fig. 2*D*).

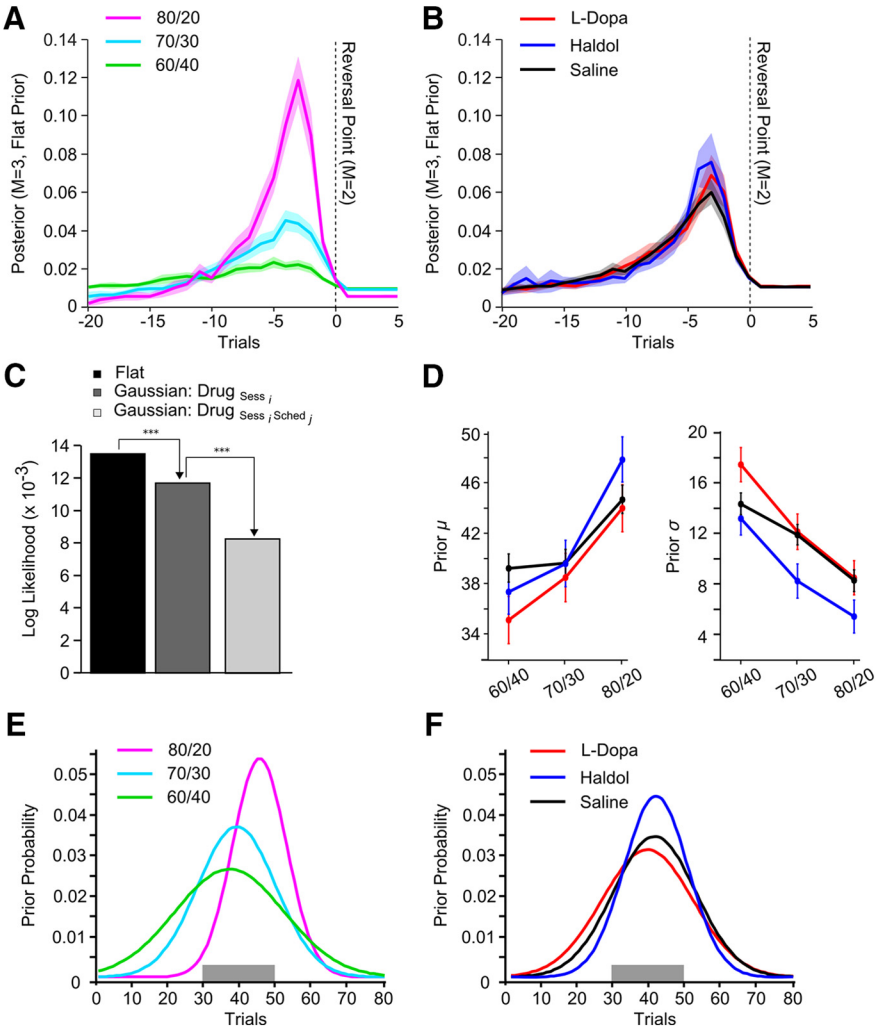

Figure 3. Causal evidence at the time the monkeys reversed their behavior. ***$p < 0.001$. Error bars and shading indicate 1 SEM, and the gray windows indicate the trial range in which a reversal was programmed to occur. ***A***, The posterior of the causal model ($M = 3$, flat prior) aligned to the estimated trial on which the monkeys switched their choice behavior, averaged by reward schedule (***A***) or drug condition (***B***). Note that the posterior in these plots was calculated with a flat prior. ***C***, Log likelihoods for different models with different priors. Sess, Session; Sched, schedule. ***D***, Mean and SD of prior distributions fit to individual sessions for each schedule and drug condition. ***E***, ***F***, Average prior distributions for each reward schedule and drug condition.

Reinforcement learning in acquisition and reversal

We next fit reinforcement learning models to the choice behavior to estimate the effects of positive and negative feedback (feedback type) on choices, as well as the consistency of the animals' decisions (the inverse temperature, a measure of choice consistency). These three parameters, one for positive feedback, one for negative feedback, and one for inverse temperature, were estimated as a function of drug condition, reward schedule, and learning phase. We first split each block into an acquisition and reversal phase according to the switch points estimated by the ideal observer or behavioral choice models. The split by either the ideal observer or behavioral choice models affected whether individual trials were included in the acquisition or the reversal phase during model estimation. A separate set of RL model parameters were then fit for each phase. Comparing analyses under the two different splits can provide additional insight into the animal's

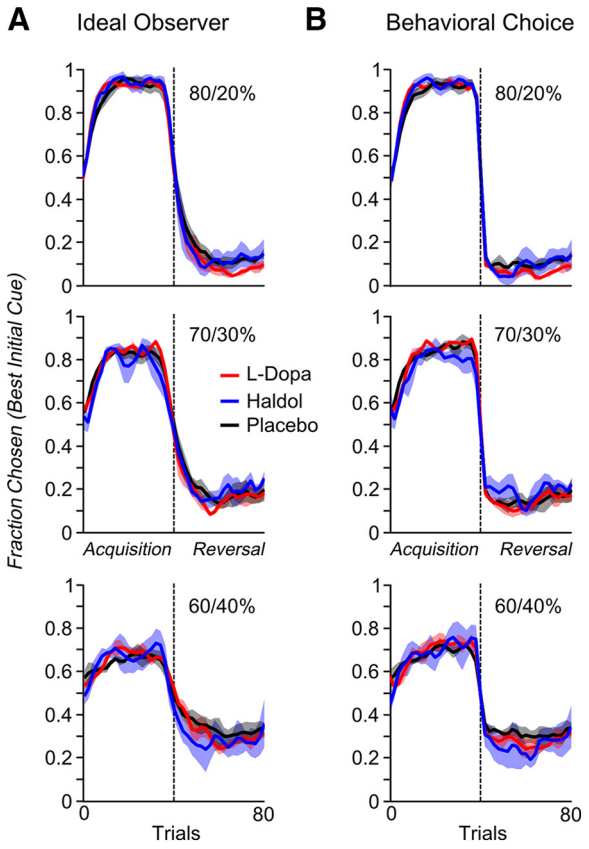

Figure 4. The fraction of times the initial high probability cue was chosen in the acquisition and reversal phases, broken out by drug and reward schedule. Curves were smoothed with a moving average window of six trials. Because the number of trials before and after acquisition varied across blocks, trial number was normalized to be between 0 and 1 within each phase and then averaged across blocks to generate the plots. ***A***, Choices aligned to reversal points estimated by the ideal observer model ($M = 1$). ***B***, Choices aligned to reversal points based on reversals in the monkeys' behavior ($M = 2$).

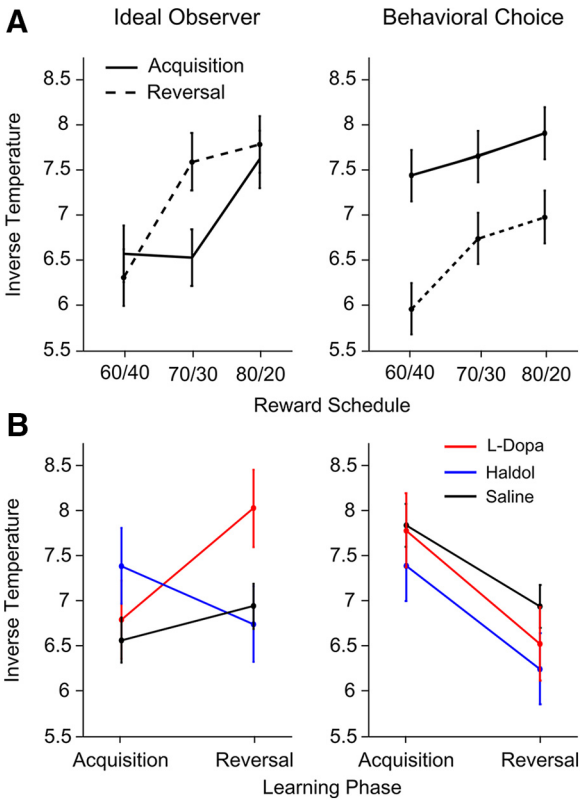

Figure 5. Effects of drug and schedule on inverse temperature estimated with phase divided by ideal observer and behavioral choice models. Error bars indicate 1 SEM. ***A***, Inverse temperature broken out by schedule when acquisition and reversal are defined by ideal observer (left) and behavioral choice (right) models. ***B***, Inverse temperature broken out by drug condition when acquisition and reversal are defined by ideal observer (left) and behavioral choice (right) models.

strategy, as well as explain effects on learning that would be attributable to improperly considering the animal's strategy and analyzing all data as if they behaved optimally.

The inverse temperature quantifies how consistently the animals chose the higher value option, particularly when value estimates have reached asymptote. As the inverse temperature parameter increases, animals more consistently choose the higher valued option for a fixed difference in values. When the reversal was estimated with blocks split by the ideal observer, the inverse temperature increased across the three reward schedules (Fig. 5*A*, left; schedule, $F_{(2,161)} = 3.22$, $p = 0.042$). In addition, the administered drug affected the inverse temperature differently for acquisition and reversal (Fig. 5*B*, left; drug × phase, $F_{(2,37)} = 4.43$, $p = 0.019$). On haloperidol, there was a decrease in the inverse temperature as the animals moved from the acquisition to the reversal phase, whereas on L-DOPA ($F_{(2,17)} = 5.01$, $p = 0.038$) or saline ($F_{(2,20)} = 4.88$, $p = 0.039$), the inverse temperature increased between the two learning phases.

We next performed the same analysis with the blocks split by reversal points estimated from monkeys' choice behavior (Fig. 5, right columns). With the data split in this manner, there was a main effect of learning phase ($F_{(1,101)} = 7.42$, $p = 0.008$), but there was no main effect of reward schedule ($F_{(2,152)} < 1$, $p = 0.591$) or an interaction with learning phase (schedule × phase, $F_{(2,143)} = 1.97$, $p = 0.143$). Thus, the schedule effect seen when the data were split by the ideal observer was not found when the data were split by the behavioral choice model and were subsumed by the phase-related changes in the inverse temperature. There were also no drug-related effects on the inverse temperature (Fig. 5*B*, right; drug, $F_{(2,38)} < 1$, $p = 0.943$; drug × phase, $F_{(2,40)} < 1$, $p = 0.553$). This is consistent with the fact that the choice data (Fig. 4) switches more abruptly when it is aligned to the behavioral choice model.

Next we analyzed the positive and negative feedback learning rate parameters. Using the reversal point estimated by the ideal observer (Fig. 6*A*, left), learning was more influenced by positive versus negative feedback (feedback, $F_{(1,95)} = 124.65$, $p < 0.001$), and learning rates were overall higher in the reversal compared with the acquisition phase (phase, $F_{(1,90)} = 4.1$, $p = 0.046$). There was no evidence that compared with saline, haloperidol, or L-DOPA modulated learning from positive versus negative feedback (drug × feedback, $F_{(2,38)} = 2.68$, $p = 0.081$). However, there was evidence that, when tested relative to each other (i.e., haloperidol vs L-DOPA), the two dopaminergic drugs modulated feedback learning (drug × feedback, $F_{(1,20)} = 6.97$, $p = 0.015$). Specifically, learning from positive feedback was heightened on L-DOPA compared with haloperidol ($t_{(18)} = 2.58$, $p = 0.018$), whereas learning from negative feedback was unaffected by drug ($t_{(18)} = 1.67$, $p = 0.221$).

When we further analyzed learning rates with blocks split according to the behavioral choice model (Fig. 6, right column), learning was again more influenced by positive versus negative feedback (feedback, $F_{(1,99)} = 189.4$, $p < 0.001$), with higher positive learning rates in the reversal compared with the acquisition

phase (feedback × phase, $F_{(1,99)} = 20.66$, $p < 0.001$). When choice data were split according to the behavioral choice model, learning rates did not differ by drug condition (Fig. 6*B*, right; drug, $F_{(2,38)} = 1.66$, $p = 0.204$; drug × feedback, $F_{(2,40)} < 1$, $p = 0.817$; drug × feedback × phase, $F_{(2,40)} < 1$, $p = 0.657$). Thus, when learning was referenced to when the animal reversed its choice behavior, there was no evidence that dopamine modulated learning rates.

Discussion

We have examined the effects of dopamine manipulation on detecting reversals in stimulus–reward associations during conditions of expected uncertainty (Yu and Dayan, 2005). Using a Bayesian analysis to characterize reversals in choice behavior, we found that the animals abruptly switched their choice behavior, consistent with a strategy well matched to the structure of the task. On haloperidol and L-DOPA, the animals were more accurate in detecting reversals, and they more often chose the optimal stimulus, earning more rewards. In addition, on haloperidol, there was a decrease in the variance of the monkeys' prior estimates about when reversals would occur. Finally, we used reinforcement learning models to analyze the behavior, splitting the data by either the ideal observer or the animals' choice behavior. There were drug-related effects on learning when the data were split by the ideal observer but not when the data were split by the behavioral choice model. These results emphasize that understanding the strategy of the animal is important for interpreting behavioral effects.

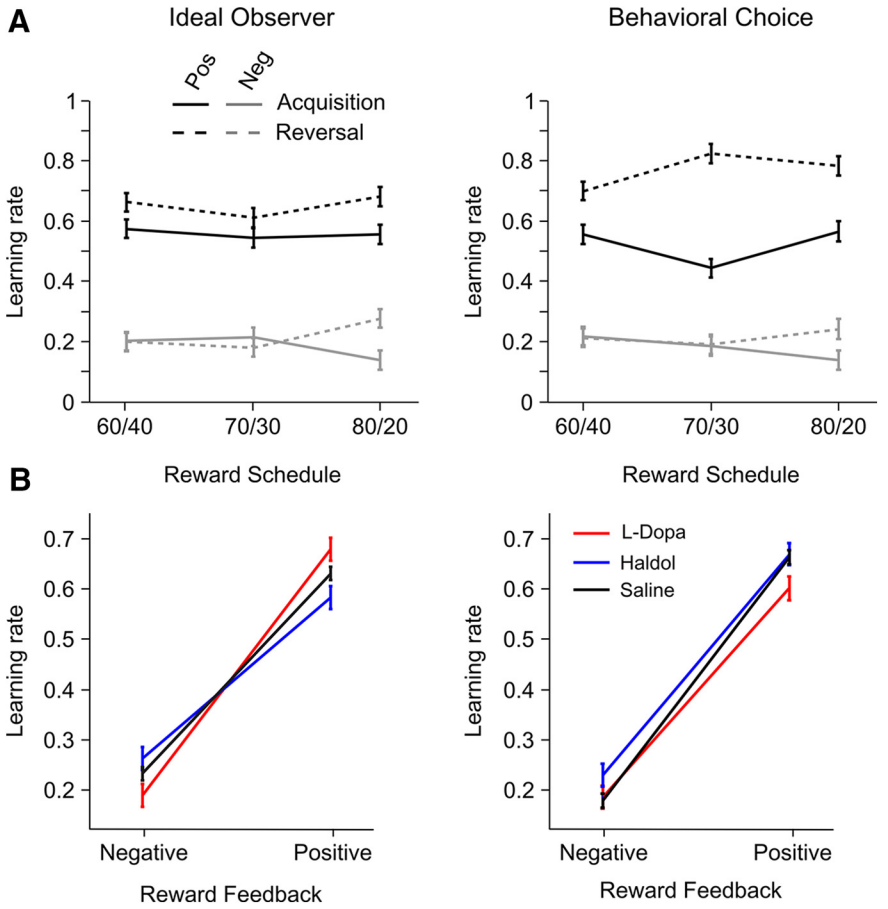

Figure 6. Effects on learning rate parameters for positive and negative feedback estimated with acquisition and reversal phase divided by ideal observer and behavioral choice models. Error bars indicate 1 SEM. ***A***, Learning rates for each form of feedback broken out by reward schedule and learning phase, averaged across drug condition. Pos, Positive; Neg, negative. ***B***, Learning rates by drug condition, averaged across learning phase and reward schedule.

Bayesian model of reversal learning

The Bayesian models provided a detailed characterization of the behavior. Specifically, we used an ideal observer (model 1) to calculate a posterior over the trial on which the stimulus reward mapping switched, as opposed to using the point at which the algorithm switched the mapping. These will not necessarily coincide, particularly in the more difficult schedules. Once we obtained a posterior distribution over switch trials, we derived a point estimate of the trial on which the reversal occurred. We performed a similar analysis for the choice behavior of the animals (model 2). This model estimated when the animal's choice behavior switched, as opposed to the ideal observer's estimate of when the stimulus reward mapping switched. Combining these approaches allowed us to examine when the animal switched relative to when the ideal observer identified a switch. This showed that, during 60/40% blocks, the animals, on average, switched before the trial that was identified by the ideal observer. Switching before the actual reversal reflects the influence of the animal's prior belief about where the switch would occur (i.e., between trials 30 and 50). When we examined the absolute deviation between the ideal observer and the animal's behavioral choice, it was largest in the most difficult schedule and smallest in the easiest schedule, again consistent with the structure of the task.

We also found that, when the behavioral choice data were aligned to the ideal observer's switch point, it looked as though the animals gradually switched their choice behavior, especially in the 60/40% condition. However, when we aligned the data using the estimate of where the animal reversed, it could be seen that the animals switched relatively abruptly. Thus, rather than gradually switching their choices as might be expected if they had used a simple reinforcement learning strategy with noisy exploration, they switched abruptly, consistent with the actual statistics of the task.

We also examined the evidence on which the choice to reverse was based. Consistent with the statistics of the task, the evidence on which the animals reversed their behavior was weaker in the 60/40% condition than in the 80/20% condition. There were no differences in the causal evidence on which animals switched as a function of drug. As already mentioned, the switching behavior of the animals suggested they used a prior in combination with causal evidence to determine when a contingency reversal occurred. When we estimated a prior over reversal trial under this assumption, we found that haloperidol led to a prior with less variance but the same mean as the other drug conditions. Thus, the animals relied more on their prior when deciding when to switch under haloperidol, because a prior with less variance more strongly affects the posterior evidence for a reversal.

Effects of dopamine manipulations on reversal learning

The relationship between dopamine and learning is often debated (Redgrave et al., 2008; Berridge et al., 2009; Nicola, 2010; Salamone and Correa, 2012). When we fit the RL algorithm to the data split by the ideal observer, there were effects of dopamine manipulations consistent with previous reports (Frank and O'Reilly, 2006). Specifically, haloperidol increased the effect of negative feedback and decreased the effect of positive feedback relative to L-DOPA. However, when we fit the RL model to the data aligned to the animals' reversal in choice behavior (which shifts trials near the reversal point from the reversal to the acquisition phase), we found that neither haloperidol nor L-DOPA differentially affected learning rates as assessed with the RL algorithm. Therefore, the drug-related effects on learning are accounted for by differences in switch points identified by the two models. Because the animals switched rather abruptly, they did not seem to be using a noisy, explorative reinforcement learning strategy. We also found evidence that they incorporated a prior into their choice process. Therefore, they appeared to be more Bayesian, having learned the statistical structure of the task.

Haloperidol did increase the consistency with which the animals reversed their choice behavior relative to the ideal observer and correspondingly decreased the variance of the prior that drove reversals. These results intersect with accumulating evidence that reversal learning in monkeys is mediated by D_2 receptor signaling. However, D_2 receptor antagonism is typically found to impair reversal learning. Deficits in reversal learning, defined in terms of the number of errors made before reaching criterion, are seen after neurochemical dopamine depletion in the medial caudate (Clarke et al., 2011), decreases in caudate D_2 receptor availability (Groman et al., 2011), and systemic injections of the D_2/D_3 receptor antagonist raclopride (Lee et al., 2007). Therefore, a straightforward interpretation of how haloperidol affected reversal learning, in terms of standard theories of dopamine, is difficult.

One possible explanation is that by disrupting striatal D_2 mechanisms, haloperidol causes an increased reliance on learned strategies driven by cortical mechanisms. For example, when monkeys are first extensively trained on serial object reversal learning, subsequent lesions to ventrolateral prefrontal cortex impair generalization of that training to reversals with novel object pairs (Rygula et al., 2010). Related deficits on serial reversal learning are also seen after disconnection lesions of prefrontal and inferotemporal cortex (Wilson and Gaffan, 2008). Also in marmosets, dopaminergic lesions of the orbitofrontal cortex cause increases in tonic striatal dopamine levels and D_2 receptor occupancy. These changes in dopaminergic tone are then correlated with increased sensitivity to probabilistic reward feedback (Clarke et al., 2014). Although this study did not assess reversal learning, this result is intriguing because it does imply that tonic blockade of striatal D_2 receptors with haloperidol might prompt less reliance on immediate feedback.

Administration of L-DOPA also caused animals to reverse closer to the ideal observer, but this effect could not be attributed to a change in the prior. A possible explanation for why haloperidol and L-DOPA both increase the accuracy of the animals in detecting reversals is that both drugs have similar effects on phasic dopamine release, because haloperidol can antagonize presynaptic D_2 autoreceptors, causing increased release in the striatum (Kuroki et al., 1999; Wu et al., 2002; Robinson et al., 2003). However, identifying potential mechanisms that explain the dissociable effects of haloperidol and L-DOPA on the monkeys' reliance on a prior is an open question.

A final point is that previous studies of reversal learning have examined performance while animals were learning that reversals occur. In these tasks, performance improves over the course of the experiment. For example, in the study by Clarke et al. (2011), there is a statistical main effect of number of reversals and no interaction of reversal and group. This shows that the animals across all conditions are learning about reversals during the study. There are two possible hypotheses for the deficits between groups in these prior studies. One is that deficits are related to effects of dopamine on a naive model that the animals use to solve the task before they learn that reversals occur. The other hypothesis is that the animals are not able to learn the correct model that solves the task more efficiently, analogous to the idea of developing a learning set (Wilson and Gaffan, 2008). Both of these hypotheses would result in similar deficits in the first few blocks, but they make different predictions about how the deficits evolve with experience.

Conclusion

We developed a Bayesian model of reversal learning to study the behavioral effects of dopamine manipulation. We found that, with extensive experience, animals developed a behavioral strategy that was well matched to the actual features of our task. In difficult schedules, animals switched their behavior, on average, earlier than the ideal observer, reflecting the influence of a prior on when a reversal would occur. In addition, when the behavioral choice data were aligned to the animals' switch trial, it could be seen that the animals switched choices abruptly, as opposed to gradually changing their behavior over a series of trials. We also found that administration of haloperidol or L-DOPA lead to increased performance on the task. Overall, the Bayesian formalism makes explicit the animal's strategy and allows for a thorough examination of the behavior and effects of causal manipulations.

References

Asaad WF, Eskandar EN (2008) A flexible software tool for temporally-precise behavioral control in Matlab. J Neurosci Methods 174:245–258. CrossRef Medline

Berridge KC, Robinson TE, Aldridge JW (2009) Dissecting components of reward: "liking," "wanting," and learning. Curr Opin Pharmacol 9:65–73. CrossRef Medline

Chudasama Y, Robbins TW (2003) Dissociable contributions of the orbitofrontal and infralimbic cortex to Pavlovian autoshaping and discrimination reversal learning: further evidence for the functional heterogeneity of the rodent frontal cortex. J Neurosci 23:8771–8780. Medline

Clarke HF, Hill GJ, Robbins TW, Roberts AC (2011) Dopamine, but not serotonin, regulates reversal learning in the marmoset caudate nucleus. J Neurosci 31:4290–4297. CrossRef Medline

Clarke HF, Cardinal RN, Rygula R, Hong YT, Fryer TD, Sawiak SJ, Ferrari V, Cockcroft G, Aigbirhio FI, Robbins TW, Roberts AC (2014) Orbitofrontal dopamine depletion upregulates caudate dopamine and alters behavior via changes in reinforcement sensitivity. J Neurosci 34:7663–7676. CrossRef Medline

Cools R, Barker RA, Sahakian BJ, Robbins TW (2001) Enhanced or impaired cognitive function in Parkinson's disease as a function of dopaminergic medication and task demands. Cereb Cortex 11:1136–1143. CrossRef Medline

Cools R, Altamirano L, D'Esposito M (2006) Reversal learning in Parkinson's disease depends on medication status and outcome valence. Neuropsychologia 44:1663–1673. CrossRef Medline

Dias R, Robbins TW, Roberts AC (1996) Primate analogue of the Wisconsin Card Sorting Test: effects of excitotoxic lesions of the prefrontal cortex in the marmoset. Behav Neurosci 110:872–886. CrossRef Medline

Fellows LK, Farah MJ (2003) Ventromedial frontal cortex mediates affective shifting in humans: evidence from a reversal learning paradigm. Brain 126:1830–1837. CrossRef Medline

Frank MJ, O'Reilly RC (2006) A mechanistic account of striatal dopamine func-

tion in human cognition: psychopharmacological studies with cabergoline and haloperidol. Behav Neurosci 120:497–517. CrossRef Medline

Graef S, Biele G, Krugel LK, Marzinzik F, Wahl M, Wotka J, Klostermann F, Heekeren HR (2010) Differential influence of levodopa on reward-based learning in Parkinson's disease. Front Hum Neurosci 4:169. CrossRef Medline

Groman SM, Lee B, London ED, Mandelkern MA, James AS, Feiler K, Rivera R, Dahlbom M, Sossi V, Vandervoort E, Jentsch JD (2011) Dorsal striatal D2-like receptor availability covaries with sensitivity to positive reinforcement during discrimination learning. J Neurosci 31:7291–7299. CrossRef Medline

Hornak J, O'Doherty J, Bramham J, Rolls ET, Morris RG, Bullock PR, Polkey CE (2004) Reward-related reversal learning after surgical excisions in orbito-frontal or dorsolateral prefrontal cortex in humans. J Cogn Neurosci 16:463–478. CrossRef Medline

Izquierdo A, Suda RK, Murray EA (2004) Bilateral orbital prefrontal cortex lesions in rhesus monkeys disrupt choices guided by both reward value and reward contingency. J Neurosci 24:7540–7548. CrossRef Medline

Jones B, Mishkin M (1972) Limbic lesions and the problem of stimulus–reinforcement associations. Exp Neurol 36:362–377. CrossRef Medline

Kuroki T, Meltzer HY, Ichikawa J (1999) Effects of antipsychotic drugs on extracellular dopamine levels in rat medial prefrontal cortex and nucleus accumbens. J Pharmacol Exp Ther 288:774–781. Medline

Lee B, Groman S, London ED, Jentsch JD (2007) Dopamine D2/D3 receptors play a specific role in the reversal of a learned visual discrimination in monkeys. Neuropsychopharmacology 32:2125–2134. CrossRef Medline

Levin JR, Serlin RC, Seaman MA (1994) A controlled, powerful multiple-comparison strategy for several situations. Psychol Bull 115:153–159. CrossRef

Mehta MA, Swainson R, Ogilvie AD, Sahakian J, Robbins TW (2001) Improved short-term spatial memory but impaired reversal learning following the dopamine D(2) agonist bromocriptine in human volunteers. Psychopharmacology 159:10–20. CrossRef Medline

Mitz AR (2005) A liquid-delivery device that provides precise reward control for neurophysiological and behavioral experiments. J Neurosci Methods 148:19–25. CrossRef Medline

Nicola SM (2010) The flexible approach hypothesis: unification of effort and cue-responding hypotheses for the role of nucleus accumbens dopamine in the activation of reward-seeking behavior. J Neurosci 30:16585–16600. CrossRef Medline

Redgrave P, Gurney K, Reynolds J (2008) What is reinforced by phasic dopamine signals? Brain Res Rev 58:322–339. CrossRef Medline

Robinson DL, Venton BJ, Heien ML, Wightman RM (2003) Detecting subsecond dopamine release with fast-scan cyclic voltammetry in vivo. Clin Chem 49:1763–1773. CrossRef Medline

Rudebeck PH, Saunders RC, Prescott AT, Chau LS, Murray EA (2013) Prefrontal mechanisms of behavioral flexibility, emotion regulation and value updating. Nat Neurosci 16:1140–1145. CrossRef Medline

Rygula R, Walker SC, Clarke HF, Robbins TW, Roberts AC (2010) Differential contributions of the primate ventrolateral prefrontal and orbitofrontal cortex to serial reversal learning. J Neurosci 30:14552–14559. CrossRef Medline

Salamone JD, Correa M (2012) The mysterious motivational functions of mesolimbic dopamine. Neuron 76:470–485. CrossRef Medline

Schoenbaum G, Setlow B, Nugent SL, Saddoris MP, Gallagher M (2003) Lesions of orbitofrontal cortex and basolateral amygdala complex disrupt acquisition of odor-guided discriminations and reversals. Learn Mem 10:129–140. CrossRef Medline

Turchi J, Devan B, Yin P, Sigrist E, Mishkin M (2010) Pharmacological evidence that both cognitive memory and habit formation contribute to within-session learning of concurrent visual discriminations. Neuropsychologia 48:2245–2250. CrossRef Medline

Wilson CR, Gaffan D (2008) Prefrontal-inferotemporal interaction is not always necessary for reversal learning. J Neurosci 28:5529–5538. CrossRef Medline

Wu Q, Reith MEA, Walker QD, Kuhn CM, Carroll FI, Garris PA (2002) Concurrent autoreceptor-mediated control of dopamine release and uptake during neurotransmission: an *in vivo* voltammetric study. J Neurosci 22:6272–6281. Medline

Yu AJ, Dayan P (2005) Uncertainty, neuromodulation, and attention. Neuron 46:681–692. CrossRef Medline