Archival Report

Biological Psychiatry

Value Modulation of Self-Defeating Impulsivity

Zhe Liu, Robert Reiner, Yonatan Loewenstein, and Eran Lottem

ABSTRACT

BACKGROUND: Impulse control is a critical aspect of cognitive functioning. Intuitively, whether an action is executed prematurely depends on its associated reward, yet the link between value and impulsivity remains poorly understood. Three frameworks for impulsivity offer contrasting views: impulsive behavior may be valuable because it is associated with hidden internal reward (e.g., reduction of mental effort). Alternatively, it can emerge from exploration, which is disadvantageous in the short term but can yield long-term benefits. Finally, impulsivity may reflect Pavlovian bias, an inherent tendency that occurs even when its outcome is negative.

METHODS: To test these hypotheses, we trained 17 male mice to withhold licking while anticipating variable rewards. We then measured and optogenetically manipulated dopamine release in the ventral striatum.

RESULTS: We found that higher reward magnitudes correlated with increased impulsivity. This behavior was well explained by a Pavlovian bias model. Furthermore, we observed negative dopamine signals during premature licking, suggesting that in this task, impulsivity is not merely an unsuccessful attempt at obtaining a reward. Rather, it is a failure to overcome the urge to act prematurely despite knowledge of the negative consequences of such impulsive actions.

CONCLUSIONS: Our findings underscore the integral role value plays in regulating impulsivity and suggest that the dopaminergic system influences impulsivity through the mediation of value learning.

https://doi.org/10.1016/j.biopsych.2024.09.017

Impulsivity is a multifaceted cognitive construct (1). Among its various forms, waiting impulsivity—failure to withhold an action in the face of delay—has gained considerable attention as a core symptom of a number of psychological disorders such as drug addiction, gambling, and attention-deficit/hyperactivity disorder (2).

Reinforcement learning (RL) provides a general framework for value-based behaviors (3). RL algorithms typically converge to a behavior that maximizes accumulated rewards, and therefore, behaviors that seem detrimental, such as impulsivity, are a challenge to this framework. One possible interpretation of impulsivity is that it does reflect maximization of accumulated rewards, taking into account hidden, internal rewards or punishments. For example, withholding an action can be associated with a punishment, such as mental effort (4). A second interpretation is that impulsive behavior reflects exploration, a crucial component of RL (5,6). Because actions that initially seem disadvantageous may turn out to be beneficial, occasionally choosing seemingly worse actions is necessary for finding optimal strategies. This opens up the possibility of interpreting impulsive decisions as components of an exploratory strategy (7). Indeed, increased impulsivity has been linked to heightened exploration, which was also suggested to underlie attention-deficit/hyperactivity disorder (8). Finally, impulsivity may be attributed to an asymmetry between action and inaction, known as Pavlovian bias (9,10). Generally speaking, it is easier to train an animal to act to obtain a reward than to withhold action for the same goal (11–14).

A key distinction between these explanations has to do with the effect that value has on impulsivity: The contribution of hidden costs and exploration to impulsivity is expected to decrease as the value of being patient increases. By contrast, a Pavlovian bias to act for reward is expected to increase with reward magnitude.

At the neurobiological level, a considerable body of research links midbrain dopamine neurons in the ventral tegmental area (VTA) to RL (15–17). These neurons encode reward prediction errors (RPEs), which are defined as the difference between received and predicted reward (3,18–20). One of the major targets of these neurons is the ventral striatum (VS) (which includes the nucleus accumbens), which, beyond its importance to Pavlovian learning (21,22), is known to be a key brain structure for controlling certain forms of impulsivity, including waiting impulsivity (23–26).

In this study, our goal was to determine the behavioral effects of changing reward size on impulsive behavior and its regulation by VS dopamine. We trained head-fixed mice in a task that required them to withhold licking while expecting outcomes of different values. We found that mice impulsivity was correlated with expected values and that this behavior was well captured by an RL model, in which a Pavlovian bias is incorporated. We also found that VS dopamine release was compatible with RPE coding as predicted by this model. Finally, we showed that pairing surprising reward omissions with optogenetic stimulation of dopaminergic axons in the VS was sufficient to block learning in this task.

ISSN: 0006-3223

METHODS AND MATERIALS

Animal Subject, Training, and Behavioral Setup

All protocols and procedures were approved by the Institutional Animal Care and Use Committees at the Hebrew University of Jerusalem and were in accordance with the National Institutes of Health Guide for the Care and Use of Laboratory Animals. Seventeen male mice, 2 to 3 months of age, were used in this study. All mice underwent stereotaxic viral injections, as well as fiber and head-bar implantation for head fixation. They were then trained in an odor-guided waiting task that required them to withhold licking for several seconds to receive a water reward. The mice were group housed under a standard 12-hour light/dark cycle. During training and experiment days, the mice had free access to food, but water was only available during the behavioral sessions, with a 24-hour period of free water on weekends.

To record dopamine release in the VS using fiber photometry, we expressed the dopamine sensor $GRAB_{DA2m}$ (27) unilaterally in 4 mice and GFP (green fluorescent protein) in 2 control mice and then implanted an optic fiber above their VS. To examine dopamine release effects on impulsivity, we expressed ChR2 (channelrhodopsin-2) bilaterally in the VTA of 5 heterozygous TH-Cre transgenic mice and GFP in 4 control mice and implanted 2 optic fibers above their VS.

All analyses were performed using custom code written in MATLAB (version 9.9.0.2037887; R2020b; The MathWorks, Inc.). In all figures, average data and error bars, or shaded patches around curves, represent mean ± SEM.

Methods are detailed in the Supplement.

RESULTS

An Odor-Guided Waiting Task to Measure Impulsivity in Mice

To investigate the effect that value has on impulsivity, we trained 14 head-fixed mice in an odor-guided waiting task (Figure 1A, B). At the start of each trial, we randomly selected one of the 3 odor cues (conditioned stimuli [CS]) and presented it to the mice for 1 second. Each CS was followed by a Gaussian-distributed waiting period (4 ± 0.25 seconds). If the mice successfully suppressed licking during the waiting period, a go tone was played to indicate that it was safe to lick. Each of the odors was associated with a different outcome (unconditioned stimulus [US]): a big reward (8-μL water drop), a small reward (4-μL water drop), and no reward. In impatient trials, in which the mice licked before the go tone, no reward was provided, and no additional feedback was provided to signal the impatient behavior. In all 3 conditions, trial duration was independent of the mice licking, thus providing no motivation for the animals to lick to move more quickly to the next trial.

After 2 to 3 weeks of training, the mice demonstrated learning of the task's structure and contingencies. Learning the CS-US association was evident in trial-type differences in licking probability. The mice almost always licked in reward trials, and this probability was significantly lower in no-reward trials, even though the auditory go cue was identical in the 3 trial types ($F_{2,26}$ = 23.32, $p < .001$, one-way repeated-measures analysis of variance followed by a Tukey-Kramer post hoc test, n = 14 mice) (Figure 1C). Furthermore, we found that the mice also learned to respond to the go tone after it was sounded. To show this, we compared actual response times (the intervals between the go tone and first lick) with a randomized dataset, in which we assigned to each trial a go tone delay that was drawn from the same Gaussian distribution, and recalculated response times. We then compared real and randomized distributions using a sliding window (0.05-second non-overlapping windows spanning 1 second after the tone) and found that real responses were significantly lower during the first 100 ms and significantly higher between 150 and 200 ms after the tone ($p < .05$, paired t test Bonferroni corrected for multiple comparisons), suggesting that mice were indeed responding to the tone (Figure 1D).

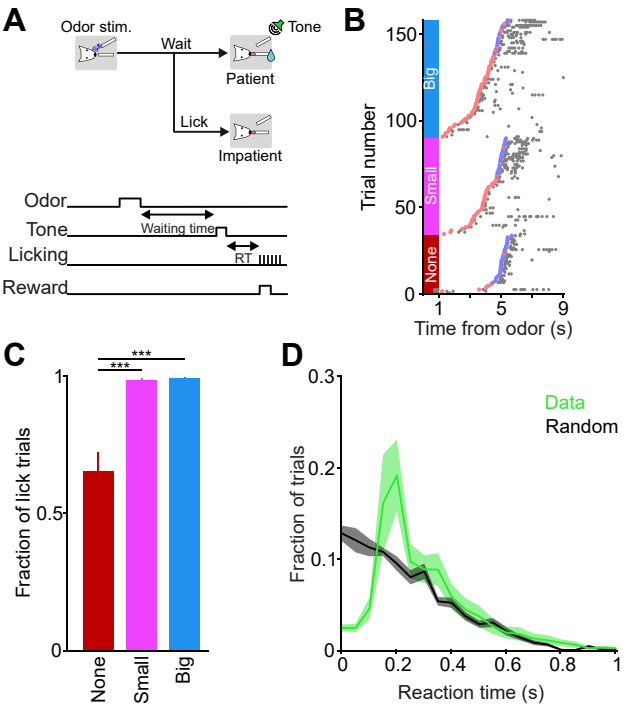

Figure 1. An odor-guided waiting task. **(A)** Diagram of trial structure (top) and events (bottom) in the task. After odor presentation, the mouse needs to wait for a randomly delayed tone (4 ± 0.25 seconds) before it can lick a waterspout for reward (patient trial), but if it licks prematurely, the reward is omitted (impatient trial). Waiting time is the interval between odor offset and either the go cue or the first premature lick in patient and premature trials, correspondingly, and RT is the interval between the go cue and the first lick. **(B)** Waiting behavior during 1 example session. Each row corresponds to a single trial. Trials are grouped according to type and sorted according to waiting time. Red and blue circles mark the ends of the waiting periods (premature licks or go tones, correspondingly), and gray dots are licks. Only trials in which the mouse licked are shown. **(C)** Bar plot showing the average fraction of trials in which the mice licked (prematurely or not) in each of the trial types (n = 14). Asterisks indicate significant difference between trial types (***$p < .001$, repeated-measures analysis of variance followed by a Tukey-Kramer post hoc test). **(D)** Distributions of the response times for rewarding trials for actual and random data. The peak of the actual data was significantly higher than that of the shuffled data ($p < .05$). RT, response time; stim., stimulation.

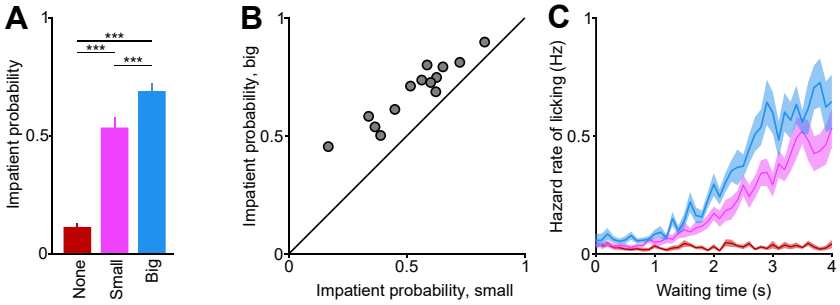

Figure 2. Value modulation of impulsivity. **(A)** Bar plot showing the mean fraction of impatient trials in each of the trial types (n = 14). Asterisks indicate significant difference between trial types (***p < .001, repeated-measures analysis of variance followed by a Tukey-Kramer post hoc test). **(B)** Impatient probability in big- vs. small-reward trial types. Each circle represents one mouse. **(C)** Average hazard rates of licking as a function of waiting time, split and colored according to trial types.

Next, to measure the effect of expected reward magnitude on impulsivity, we compared the rate of premature responses in the different trial types. We found that the mice were most impulsive in big-reward trials, moderately impulsive in small-reward trials, and mostly patient in no-reward trials ($F_{2,26}$ = 151.87, p < .001, one-way repeated-measures analysis of variance followed by a Tukey-Kramer post hoc test, n = 14 mice) (Figure 2A). We also compared the rate of premature responses of individual mice in the 2 reward trial types (big and small rewards) and found that these rates were significantly correlated within mice (r = 0.95, Pearson's correlation coefficient, p < .001) (Figure 2B). This suggested that impulsivity was a trait that generalizes across conditions.

Finally, we sought to characterize the time course of impulsive behavior within trials. We estimated the hazard rate of licking during waiting and saw that it was ramping (Figure 2C). To test the effect of reward size on the hazard, we performed Cox regression analysis (see the Supplement) and found that increasing reward size significantly increased the rate of premature licking (reward size coefficient: 1.54, p < .001, Cox regression, n = 14 mice). We concluded that both reward size and temporal proximity tended to increase impulsivity levels in mice.

An RL Model for Impulsive Behavior

The observed impulsivity is surprising if we try to interpret it in standard RL algorithms, constructed to optimize behavior. In contrast, Pavlovian bias is predicted to result in impulsivity whose magnitude increases with the size of the reward, as in Figure 2A, and with the waiting time as in Figure 2C.

Going beyond these qualitative similarities, we developed a hybrid RL-Pavlovian model that provides quantitative predictions regarding mouse impulsivity levels as a function of time (Figure 3A; see the Supplement for model derivation). According to this model, the mouse computes the expected reward's value continuously over time. This value is a product of 3 terms: 1) the utility of the water drop, which is assumed to be proportional to its volume; 2) an exponential temporal discounting term; and 3) the probability that the mouse will receive the reward.

The model is characterized by 3 parameters: the discount parameter, the utility of rewards, and the Pavlovian bias. Nevertheless, surprisingly, it generates a parameter-free prediction for the relationship between reward size and impulsivity levels. In particular, the model predicts that impulsivity odds $O = \frac{1-P_r}{P_r}$, where P_r is the reward probability (i.e., the probability of not being impulsive), is proportional to the reward size, where the proportionality constant depends of the parameters. Therefore, considering the behavior of the same animal for 2 reward sizes, the model predicts the following relationship:

$$\frac{O_{big}}{O_{small}} = \frac{R_{big}}{R_{small}} \quad (1)$$

We tested this prediction in our data, comparing impulsivity in big-reward (8 μL) versus small-reward (4 μL) trial types. Although different animals exhibited substantial variability in the impulsivity odds, their ratios were in a remarkable agreement with theory (slope: 1.80, total least squares linear regression relative to rewards ratio that is equal to 2; additionally, the average value of $\frac{O_{big}}{O_{small}}$ was 2.07 ± 0.17, not significantly different from 2, p > .05, t test, n = 14 mice) (Figure 3B).

To validate our model further, we sought to fit the model parameters to individual mouse data. While the model reproduced the general trend of the behavior, there was a qualitative discrepancy between the model's value function (and hence its hazard rate of licking during waiting) and the behavioral data near the time of the reward. In particular, the model predicted a convex hazard function, whereas the behavioral data were leveling near the time of the reward (Figure 2C). We reasoned that a simple explanation for this discrepancy may be that our model ignored timing uncertainty. This could arise from both the waiting period's variability inherent in the task's design and the subjective uncertainty of timing (28,29). To account for this variability, we added another parameter to our model that captured timing uncertainty (see the Supplement). We then fitted the model's parameters to individual behavioral hazard rates. The fitted model's behavior was in close agreement with the data, capturing the temporal structure of mouse impulsivity levels (Figure 3C and Figure S1).

Dopamine Dynamics in the VS

It is widely established that VTA dopamine neurons and dopamine release in the VS play an important role in value learning by encoding RPEs. Therefore, we wondered whether this system plays a similar role in impulsivity.

To examine VS dopamine release during impulsive behavior, we expressed the fluorescent dopamine sensor $GRAB_{DA2m}$ (27) in the VS of 4 mice (Figure 4A and Figure S2)

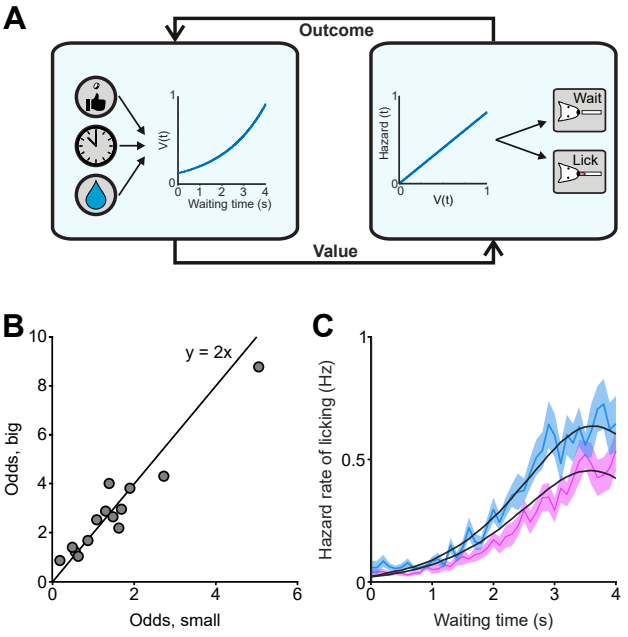

Figure 3. A Pavlovian bias reinforcement learning model for impulsivity. **(A)** Schematic diagram of the Pavlovian bias reinforcement learning model. Value functions are computed from reward probability, time, and size and then fed into an action module that converts values to licking probability. **(B)** Scatter plot showing impatient odds in big- vs. small-reward trial types. Each circle represents 1 mouse (n = 14). The black line shows a theoretically derived predication for this relation. **(C)** Average hazard rates of licking as a function of waiting time, split and colored according to trial types (same as Figure 2C). The black lines correspond to the modeled hazard rates.

and collected bulk dopamine fluorescence signals in their VS using fiber photometry for 8 to 10 days. Figure 4B, C shows dopamine fluorescence collected during one of the sessions, and Figure 4D shows the average z-scored dopamine signals in patient trials, grouped according to the trial type.

We started by analyzing dopamine signals in patient trials. As expected, VS dopamine correlated with reward size following both the CS and US. It was higher in big-reward trials and smallest in no-reward trials ($GRAB_{DA2m}$ signals from CS to 0.5 second before the US: $F_{2,6} = 20.23$, $p < .01$; $GRAB_{DA2m}$ signals 1–3 seconds after US: $F_{2,6} = 61.00$, $p < .001$; one-way repeated-measures analysis of variance followed by a Tukey-Kramer post hoc test, n = 4 mice) (Figure 4E, F). We also examined photometry signals in 2 control mice expressing GFP and found no responses in these mice (Figure S3).

We next turned to analyze VS dopamine signals in impatient trials, focusing on responses after premature licking. As expected, these responses were negative. However, our model makes another less obvious prediction that the magnitude of dopamine responses after premature licking should be modulated by action timing. Given that the value is increasing with time, both due to temporal discounting and because the probability that the animal will eventually be patient and receive the reward also increases with time, the RPE negativity should be greater in late than early responses (Figure 5A). This was the case in our data, where we found that dopamine signals in a 2-second window (1–3 seconds after the first impatient lick) were negatively correlated with the time of the action [linear mixed-effects model: dopamine ~ 1 + waiting time + (1 | MouseID), waiting time coefficient = −0.12, $p < .001$, n = 4 mice] (Figure 5B–E). This suggests that dopamine responses during impulsive behavior are sensitive to outcome timing in a manner consistent with the RPE theory of dopamine.

A Causal Role for VS Dopamine in Value Learning

Our model suggests that impulsivity is controlled by value, which likely depends on the probability of impatience. A premature trial indicates a higher probability of impatience, reducing the state's estimated value. Therefore, we hypothesized that if impatience reduces subsequent impulsivity through dopamine negativity, mice would be less impulsive after premature trials, with this effect increasing based on the timing of the impulsive action. Figure S4 supports this prediction, showing that mice adjust their reward expectations and value estimates, leading to trial-by-trial updating of impulsivity.

Based on these findings, we further hypothesized that activating VS dopamine axons during instances when a negative response would typically occur should prevent these behavioral changes.

To test this hypothesis, we used a modified version of the original task, involving repeated reversal learning of reward sizes. This design enabled us to observe and control, in a more precise manner, changes in impulsivity levels associated with each odor, as their values shift between blocks.

This task had the same trial structure as the previous one, but with only 2 trial types, reward (8 μL water drop) or no reward. However, unlike the previous task design, where CS-US associations were fixed, in this version of the task, they changed. Each session started with one of the 2 odors predicting reward and the other, no reward, and after 100 trials (the proximate middle of each session), they were reversed, such that the odor that predicted reward now predicted no reward and vice versa (Figure 6A).

Following training and behavioral data collection in the original task and after 2 to 3 days of familiarizing the mice with the new odors, we found that the mice indeed displayed rapid changes in impulsivity that were consistent with value learning. Example behavior of one session is shown in Figure 6A. To measure learning in this task, we evaluated the impulsivity levels that were associated with each odor in 6 blocks, corresponding to early, middle, and late learning during the pre- and postreversal halves of the session (Figure 6B). We then tracked the impulsivity reward bias, defined as the difference between impulsivity levels in reward and no-reward trials, across blocks. We found that the reward bias was significantly positive just before the reversal (0.19 ± 0.04, $p < .001$, t test, n = 15 mice), negative immediately after the reversal, and positive again at the end of the session. To quantify the learning process after the reversal, we compared early versus late postreversal impulsivity levels and found a significant difference between them (reward bias in early vs. late blocks after the reversal: −0.13 ± 0.03 vs. 0.07 ± 0.02, $p < .001$, paired t test, n = 15 mice) (Figure 6C), demonstrating that mouse impulsivity levels tracked the changing CS-US contingencies.

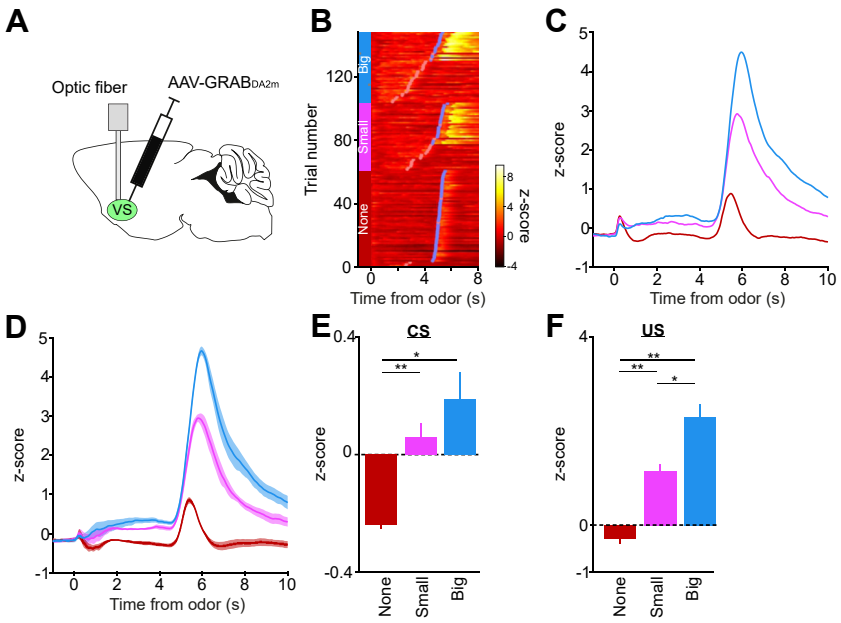

Figure 4. VS dopamine responses in patient trials. **(A)** Scheme of the locations of the dopamine sensor $GRAB_{DA2m}$ expression and optic fiber placement in the VS. **(B)** A heat map of *z*-scored dopamine signals over the course of an example session. Each row corresponds to a single trial. Trials are grouped according to type and sorted according to the waiting time. Red and blue circles mark the ends of the waiting periods (premature licks or go tones, correspondingly). **(C)** Example mouse's *z*-scored dopamine signals aligned on odor onset and split according to trial type. **(D)** Average *z*-scored dopamine signals aligned on odor onset and split according to trial type ($n = 4$). **(E, F)** Bar plots showing the mean *z*-scored dopamine responses during the waiting period **(E)** and in a 2-second window after the go cue in patient trials **(F)**; note the different y scales in the two panels. Asterisks indicate significant difference between trial types ($^{*}p < .05$, $^{**}p < .01$, repeated-measures analysis of variance followed by a Tukey-Kramer post hoc test). CS, conditioned stimulus; US, unconditioned stimulus; VS, ventral striatum.

To examine the causal contribution of dopamine to this learning process, we expressed ChR2 bilaterally in the VTA of 5 TH-Cre transgenic mice and implanted 2 optic fibers in the left and right VS. Four control mice underwent a similar surgical procedure but were infected with a virus containing GFP rather than ChR2 (Figure 7A and Figure S2).

After 1 week of training, during which we collected baseline behavioral data and observed that both groups achieved a stable reversal performance, we started a testing phase, in which in each session, photostimulation was paired with one of the odors—the one that was associated with reward at the start of the session. Photostimulation was delivered for 2 seconds only in patient trials and was triggered by the first lick that followed the go cue (Figure 7B).

Quantifying postreversal learning, we found that the ChR2-expressing group of mice did not show learning (reward bias in early vs. late blocks after the reversal: -0.17 ± 0.04 vs. -0.16 ± 0.04, $p > .05$, left-tailed paired *t* test, $n = 5$ mice) (Figure 7C, E) whereas the control groups of mice did (reward bias in early vs. late blocks after the reversal: -0.09 ± 0.05 vs.

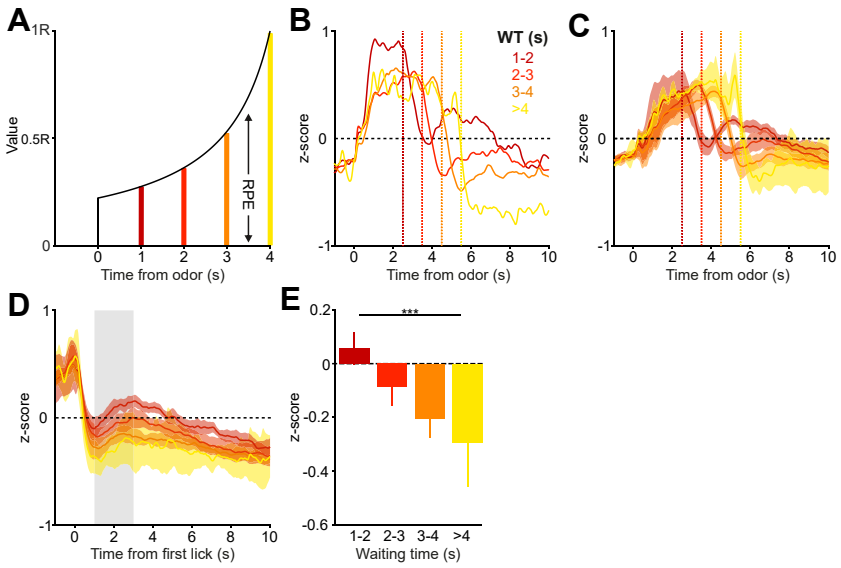

Figure 5. Ventral striatum dopamine responses in impatient trials. **(A)** Our model predicts that because value functions increase during the waiting period, the magnitude of negative RPEs following premature licking is greater in late than early licking trials. R stands for the reward utility and the lengths of the vertical lines correspond to the magnitude of the corresponding RPE had the mouse was to lick prematurely at that time. **(B)** Example mouse's *z*-scored dopamine responses aligned on odor onset and split according to WT. Vertical dashed lines mark the centers of the corresponding WT bins (measured from odor offset). **(C)** Average *z*-scored dopamine responses aligned on odor onset and split according to WT ($n = 4$). **(D)** Same as **(C)** but aligned on the time of the premature lick. **(E)** Bar plot showing the mean dopamine responses in a 2-second window after the first premature lick in impatient trials. Asterisks indicate significant linear regression slope coefficient ($^{***}p < .001$). RPE, reward prediction error; WT, waiting time.

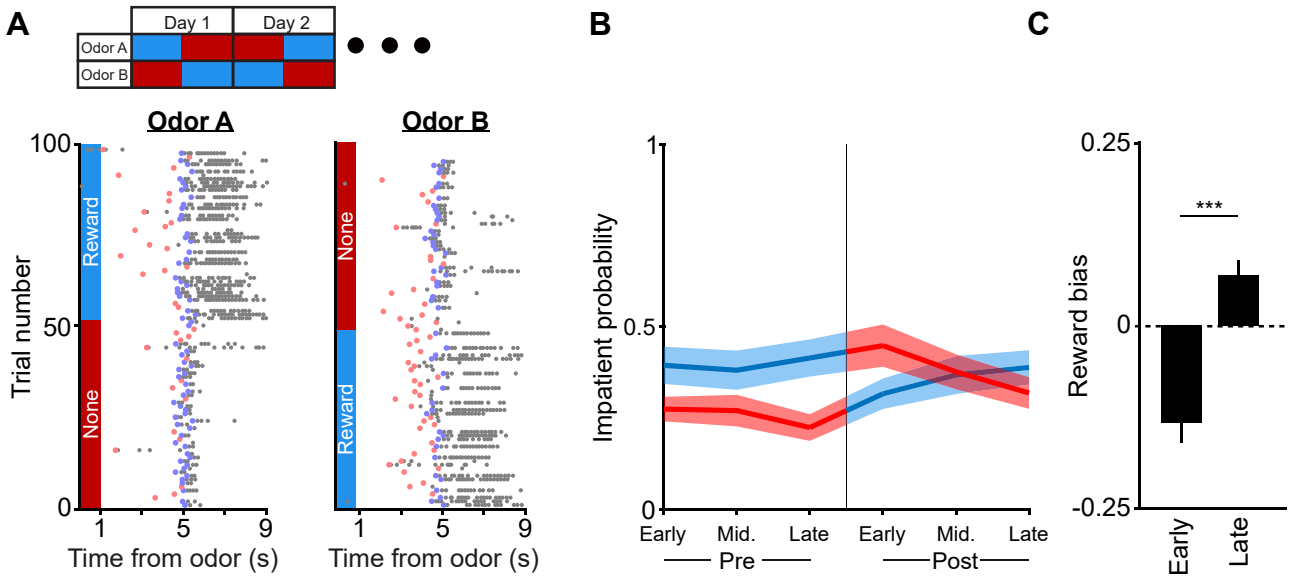

Figure 6. A reversal learning waiting task. **(A)** Top: Daily structure of the repeated reversal task. On each session, one of the 2 odors was paired with reward (blue) and the other with no reward (red). At approximately half way through the sessions (after ∼100 trials), the odor-outcome contingencies reversed, such that the odor that was paired with reward now indicated no reward and vice versa. These new associations were maintained during the first half of the following session, when they were reversed again. Bottom: Reversal behavior during one example session. Each row corresponds to a single trial. Red and blue circles mark the ends of the waiting periods (premature licks or go tones, correspondingly), and gray circles are licks. **(B)** Average impatient probability as a function of trial block (early, middle, and late pre- and postreversal blocks), split and colored according to reward size (n = 15). **(C)** Mean reward bias (the difference between impatient levels in reward vs. no-reward trials) in early and late blocks after the reversal (n = 15). Asterisks indicate significant difference between conditions (***$p < .001$, paired t test).

0.08 ± 0.03, $p < .05$, left-tailed paired t test, n = 4 mice) (Figure 7D, E). This was further confirmed by a direct comparison of the reward bias during late blocks in ChR2-expressing versus control mice, which revealed a significant difference between the 2 groups (−0.16 ± 0.04 vs. 0.08 ± 0.03 for ChR2-expressing vs. control mice, $p < .01$, t test, n = 5 ChR2-expressing mice, n = 4 control mice). To verify that this result was not due to aberrant (stimulation independent) learning behavior in the ChR2 group, we performed a similar analysis on data collected during nonstimulation days and

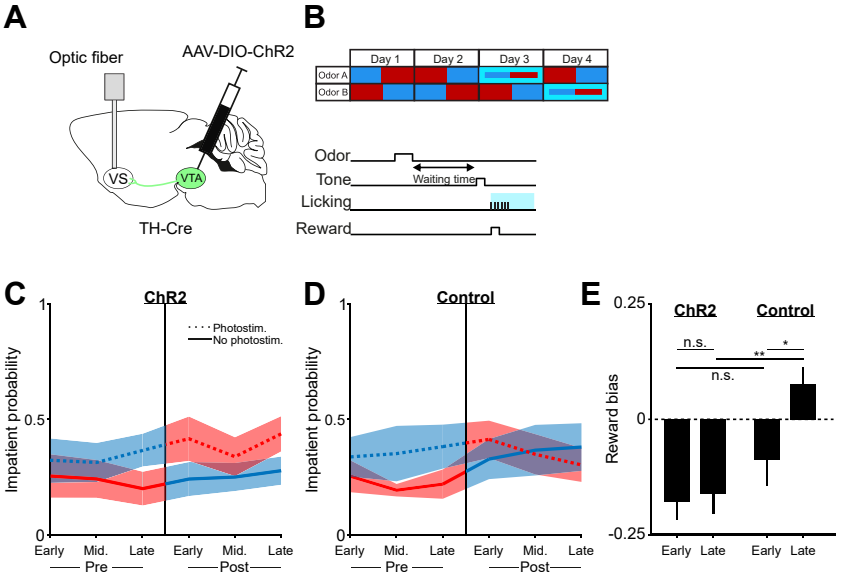

Figure 7. Optogenetic stimulation of VS dopamine axons during reversal learning. **(A)** Scheme of the locations of ChR2 expression in the VTA and optic fiber placement in the VS. **(B)** Top: Daily structure of the photostimulation protocol. On each session, the odor that predicted reward at the beginning of the session was paired with photostimulation (light blue). Bottom: Schematic of trial events. Photostimulation was triggered by the first lick after the auditory go cue (and was therefore present only in patient trials) and lasted 2 seconds. **(C)** Average impatient probability as a function of trial block (early, middle, and late pre- and postreversal blocks), split and colored according to reward size for ChR2-expressing mice during photostimulation sessions (n = 5). **(D)** Same as **(C)** for GFP-expressing control mice (n = 4). **(E)** Comparison between ChR2 and GFP mice regarding reward bias in the late after-reversal period. Asterisks indicate significant difference between blocks (*$p < .05$, **$p < .01$, paired t test). ChR2, channelrhodopsin-2; GFP, green fluorescent protein; n.s., not significant; VS, ventral striatum; VTA, ventral tegmental area.

found that both groups of mice showed learning (reward bias in early vs. late blocks after the reversal: -0.17 ± 0.05 vs. 0.07 ± 0.03, $p < .05$, left-tailed paired t test, $n = 5$ mice; reward bias in early vs. late blocks after the reversal: -0.14 ± 0.03 vs. 0.09 ± 0.04, $p < .05$, left-tailed paired t test, $n = 4$ mice) (Figure S5).

In conclusion, we found that in the ChR2-expressing mice, pairing of the reward omission to dopamine stimulation caused the reward bias to remain negative throughout the postreversal period, suggesting that VS dopamine is necessary for negative learning.

DISCUSSION

In this article, we provided theoretical and mechanistic explanations for impulsive behavior in mice. We found that mouse impulsivity is strongly linked with the expected reward size: the higher the value, the higher the impulsivity. Furthermore, we showed that this effect was well captured by a Pavlovian bias RL framework. Dopamine release in the VS was found to be consistent with RPE coding and was necessary for experience-dependent changes in impulsivity.

In general, one may consider 3 families of explanations for impulsivity. The first relates to optimality within trials. An animal may be impulsive because impulsivity has an immediate benefit, such as relieving it from mental effort (4). One could imagine that the animal weighs this immediate benefit relative to the forgone reward and chooses the option that is more beneficial in that trial. The prediction of any explanation belonging to this family is that impulsivity should decrease with expected reward. The second relates to optimality across trials through exploration. The animal believes that by forgoing the reward in one trial, it will increase its future rewards. Although there are various ways of implementing exploration in RL (5,6), they all predict that it would either decrease or remain unchanged with expected reward. The third relates to deviations from optimality. There are countless possible deviations from optimal behavior, and consistent deviations have been extensively studied in behavioral economics (30). Among these, Pavlovian bias is the only one that aligns with the waiting impulsivity observed in our experiments

Waiting impulsivity was previously studied in tasks requiring freely moving rodents to stay put while awaiting a reward [e.g., (31,32)]. One study (33) reported negative VS dopamine levels during waiting, in contrast to the small, yet positive, responses we observed (Figure 4). This finding was recently interpreted as reflecting a cognitive control signal that promotes patience by shifting the sign of the dopaminergic response (34). A possible explanation for this discrepancy is that our use of head fixation alters dopaminergic dynamics during waiting (35). Alternatively, unlike our study, this study included interleaved "go" trials (which did not require waiting), which could make the waiting trials seem worse by comparison and therefore associated with negative RPEs. A second difference between our study and studies of waiting impulsivity in freely moving rodents is that, in contrast to our findings, increasing reward size leads to decreased impulsivity (36–38). A plausible explanation for this discrepancy is that "waiting" may refer to 2 distinct cognitive functions: in our case, waiting involves inhibiting an enticing action, making impulsivity a failure of inhibition (similar to "jumping the gun"), whereas in the other studies, waiting requires persistence of inaction, with impulsivity resulting from premature giving up. Differentiating between these 2 processes is challenging, and both may occur simultaneously in the same task. Therefore, we suggest that the effects of varying reward sizes on the rate of premature responses can serve as a benchmark for identifying the type of waiting being studied.

Anticipatory licking in head-fixed mice is widely used as a robust behavioral marker for value learning (39). In a previous study (40), it was found that VTA activation after reward omission was sufficient to maintain such licking and that the inhibition of the same neurons at the time of reward delivery caused a reduction in licking [see (41) for similar results]. Here, we elaborate on these findings by providing a detailed theoretical account of this behavior. Premature licking (and by extension, anticipatory licking, which similarly does not lead to reward) occurs as a consequence of dopamine-mediated Pavlovian bias.

Previous models for the interaction between Pavlovian and instrumental behaviors, termed Pavlovian bias models, were based on a Rescorla-Wagner type framework, in which learning occurs on a trial basis (10,14,42,43), and therefore do not explain how within-trial changes in value affect the timing of behavior. In contrast, our model uses continuous-time value representations and therefore explains not only global (trial-averaged) quantities, such as overall impulsivity levels, but also the detailed temporal structure of impulsive licking during waiting.

Other Pavlovian RL models, such as temporal-difference learning, also describe within-trial value dynamics during waiting periods. These models most commonly involve partitioning the continuous waiting interval into a set of discrete microstates (19,44,45). By dispensing with discrete state representations and defining value in continuous time, our approach allowed us to derive exact, closed form equations that describe impulsive behavior. Consequently, we obtained a parameter-free quantitative prediction regarding the relationship between reward size and impulsivity that we tested and validated in our data.

Our findings on the involvement of VS dopamine in value learning and impulsivity can be compared with a recent study (46). This study revealed that dopamine signals in the dorsal striatum increased gradually during waiting, and optogenetic activation or inhibition of substantia nigra pars compacta dopamine neurons, which project to the dorsal striatum, caused actions to occur earlier or later, respectively. Taken together, these results and ours lend support to the idea of a division of labor within the basal ganglia, consistent with actor-critic RL models (3). In particular, the dorsal striatum assumes responsibility for the selection and execution of ongoing actions within a trial, while the VS and its associated dopaminergic input mediate learning and updating processes (47,48).

A potential limitation of our optogenetic experiment is that photostimulation was not calibrated to mimic naturally occurring dopamine reward responses. Recent studies suggest that supraphysiological stimulation, compared with calibrated stimulation, may have different effects on performance and learning (49,50). In one study, it was found that value-like learning occurred after supraphysiological, but not calibrated dopamine, stimulation (49). Therefore, although the precise

role of dopamine in learning is still debated, our findings nonetheless highlight the importance of value learning in regulating impulsivity in mice.

Conclusions

Our research advances our understanding of the neural mechanisms governing impulsivity in mice. The continuous-time Pavlovian bias framework presented here may have implications for the study of impulsivity and other affective processes in both animal and human behavior.

ACKNOWLEDGMENTS AND DISCLOSURES

This work was supported by a grant from the Israel Science Foundation (Grant No. 1269/20 [to EL]), a grant from the DFG (Grant No. CRC1080 [to YL]), and the Gatsby Charitable Foundation (to YL). ZL was supported by fellowships from the Edmond and Lily Safra Center for Brain Sciences. EL is the incumbent of the Sachs Family Faculty Development Chair in Brain Sciences, and YL is the incumbent of the David and Inez Myers Chair in Neural Computation.

ZL and EL designed the experiments. ZL performed the experiments with assistance from RR. ZL and EL analyzed the data with assistance from YL. YL developed the Pavlovian bias model. EL and YL wrote the article.

We thank Mati Joshua for comments on the manuscript.

The authors report no biomedical financial interests or potential conflicts of interest.

ARTICLE INFORMATION

From the Edmond and Lily Safra Center for Brain Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel (ZL, RR, YL, EL); and Alexander Silberman Institute of Life Sciences, Department of Cognitive and Brain Sciences and The Federmann Center for the Study of Rationality, The Hebrew University of Jerusalem, Jerusalem, Israel (YL).

Address correspondence to Eran Lottem, Ph.D., at eran.lottem@mail.huji.ac.il.

Received Feb 11, 2024; revised Sep 15, 2024; accepted Sep 23, 2024.

Supplementary material cited in this article is available online at https://doi.org/10.1016/j.biopsych.2024.09.017.

REFERENCES

1. Dalley JW, Robbins TW (2017): Fractionating impulsivity: Neuropsychiatric implications. Nat Rev Neurosci 18:158–171.
2. Dalley JW, Ersche KD (2019): Neural circuitry and mechanisms of waiting impulsivity: Relevance to addiction. Philos Trans R Soc Lond B Biol Sci 374:20180145.
3. Sutton RS, Barto AG (1998): Reinforcement Learning: An Introduction. Cambridge, England: MIT Press.
4. Patzelt EH, Kool W, Millner AJ, Gershman SJ (2019): The transdiagnostic structure of mental effort avoidance. Sci Rep 9:1689.
5. Fox L, Dan O, Elber-Dorozko L, Loewenstein Y (n.d.): Exploration: From machines to humans This review comes from a themed issue on Curiosity (Explore versus exploit). Curr Opin Behav Sci 2020:104–111.
6. Fox L, Dan O, Loewenstein Y (2023): On the computational principles underlying human exploration. Elife 12:RP90684.
7. Dubois M, Hauser TU (2022): Value-free random exploration is linked to impulsivity. Nat Commun 13:4542.
8. Addicott MA, Pearson JM, Schechter JC, Sapyta JJ, Weiss MD, Kollins SH (2020): Attention-deficit/hyperactivity disorder and the explore/exploit trade-off. Neuropsychopharmacology 46:614–621.
9. Rescorla RA, Solomon RL (1967): Two-process learning theory: Relationships between Pavlovian conditioning and instrumental learning. Psychol Rev 74:151–182.
10. Dayan P, Niv Y, Seymour B, Daw ND (2006): The misbehavior of value and the discipline of the will. Neural Netw 19:1153–1160.
11. Geurts DEM, Huys QJM, den Ouden HEM, Cools R (2013): Aversive Pavlovian control of instrumental behavior in humans. J Cogn Neurosci 25:1428–1441.
12. Millner AJ, Gershman SJ, Nock MK, den Ouden HEM (2018): Pavlovian control of escape and avoidance. J Cogn Neurosci 30:1379–1390.
13. Guitart-Masip M, Duzel E, Dolan R, Dayan P (2014): Action versus valence in decision making. Trends Cogn Sci 18:194–202.
14. Guitart-Masip M, Huys QJM, Fuentemilla L, Dayan P, Duzel E, Dolan RJ (2012): Go and no-go learning in reward and punishment: Interactions between affect and effect. Neuroimage 62:154–166.
15. Schultz W (2002): Getting formal with dopamine and reward. Neuron 36:241–263.
16. Saunders BT, Richard JM, Margolis EB, Janak PH (2018): Dopamine neurons create Pavlovian conditioned stimuli with circuit-defined motivational properties. Nat Neurosci 21:1072–1083.
17. Chang CY, Esber GR, Marrero-Garcia Y, Yau HJ, Bonci A, Schoenbaum G (2016): Brief optogenetic inhibition of dopamine neurons mimics endogenous negative reward prediction errors. Nat Neurosci 19:111–116.
18. Schultz W, Dayan P, Montague PR (1997): A neural substrate of prediction and reward. Science 275:1593–1599.
19. Starkweather CK, Babayan BM, Uchida N, Gershman SJ (2017): Dopamine reward prediction errors reflect hidden-state inference across time. Nat Neurosci 20:581–589.
20. Kim HR, Malik AN, Mikhael JG, Bech P, Tsutsui-Kimura I, Sun F, *et al.* (2020): A unified framework for dopamine signals across timescales. Cell 183:1600–1616.e25.
21. Day JJ, Roitman MF, Wightman RM, Carelli RM (2007): Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens. Nat Neurosci 10:1020–1028.
22. Flagel SB, Clark JJ, Robinson TE, Mayo L, Czuj A, Willuhn I, *et al.* (2011): A selective role for dopamine in stimulus-reward learning. Nature 469:53–57.
23. Murphy ER, Robinson ESJ, Theobald DEH, Dalley JW, Robbins TW (2008): Contrasting effects of selective lesions of nucleus accumbens core or shell on inhibitory control and amphetamine-induced impulsive behaviour. Eur J Neurosci 28:353–363.
24. Basar K, Sesia T, Groenewegen H, Steinbusch HWM, Visser-Vandewalle V, Temel Y (2010): Nucleus accumbens and impulsivity. Prog Neurobiol 92:533–557.
25. Sesia T, Temel Y, Lim LW, Blokland A, Steinbusch HWM, Visser-Vandewalle V (2008): Deep brain stimulation of the nucleus accumbens core and shell: Opposite effects on impulsive action. Exp Neurol 214:135–139.
26. Feja M, Hayn L, Koch M (2014): Nucleus accumbens core and shell inactivation differentially affects impulsive behaviours in rats. Prog Neuropsychopharmacol Biol Psychiatry 54:31–42.
27. Sun F, Zhou J, Dai B, Qian T, Zeng J, Li X, *et al.* (2020): Next-generation GRAB sensors for monitoring dopaminergic activity in vivo. Nat Methods 17:1156–1166.
28. Gibbon J (1977): Scalar expectancy theory and Weber's law in animal timing. Psychol Rev 84:279–325.
29. Malapani C, Fairhurst S (2002): Scalar timing in animals and humans. Learn Motiv 33:156–176.
30. Tversky A, Kahneman D (1974): Judgment under uncertainty: Heuristics and biases. Science 185:1124–1131.
31. Murakami M, Vicente MI, Costa GM, Mainen ZF (2014): Neural antecedents of self-initiated actions in secondary motor cortex. Nat Neurosci 17:1574–1582.
32. Miyazaki KW, Miyazaki K, Doya K (2012): Activation of dorsal raphe serotonin neurons is necessary for waiting for delayed rewards. J Neurosci 32:10451–10457.
33. Syed ECJ, Grima LL, Magill PJ, Bogacz R, Brown P, Walton ME (2016): Action initiation shapes mesolimbic dopamine encoding of future rewards. Nat Neurosci 19:34–36.
34. Lloyd K, Dayan P (2023): Reframing dopamine: A controlled controller at the limbic-motor interface. PLoS Comput Biol 19:e1011569.

35. Hughes RN, Bakhurin KI, Petter EA, Watson GDR, Kim N, Friedman AD, Yin HH (2020): Ventral tegmental dopamine neurons control the impulse vector during motivated behavior. Curr Biol 30:2681–2694.e5.
36. Miyazaki K, Miyazaki KW, Yamanaka A, Tokuda T, Tanaka KF, Doya K (2018): Reward probability and timing uncertainty alter the effect of dorsal raphe serotonin neurons on patience. Nat Commun 9:2048.
37. Härmson O, Grima LL, Panayi MC, Husain M, Walton ME (2022): 5-HT2C receptor perturbation has bidirectional influence over instrumental vigour and restraint. Psychopharmacology (Berl) 239:123–140.
38. Grima LL, Panayi MC, Härmson O, Syed ECJ, Manohar SG, Husain M, Walton ME (2022): Nucleus accumbens D1-receptors regulate and focus transitions to reward-seeking action. Neuropsychopharmacology 47:1721–1731.
39. Cohen JY, Haesler S, Vong L, Lowell BB, Uchida N (2012): Neuron-type-specific signals for reward and punishment in the ventral tegmental area. Nature 482:85–88.
40. Lee K, Claar LD, Hachisuka A, Bakhurin KI, Nguyen J, Trott JM, *et al.* (2020): Temporally restricted dopaminergic control of reward-conditioned movements. Nat Neurosci 23:209–216.
41. Van Zessen R, Flores-Dourojeanni JP, Eekel T, van den Reijen S, Lodder B, Omrani A, *et al.* (2021): Cue and reward evoked dopamine activity is necessary for maintaining learned Pavlovian associations. J Neurosci 41:5004–5014.
42. Huys QJM, Cools R, Gölzer M, Friedel E, Heinz A, Dolan RJ, Dayan P (2011): Disentangling the roles of approach, activation and valence in instrumental and Pavlovian responding. PLoS Comput Biol 7: e1002028.
43. Swart JC, Froböse MI, Cook JL, Geurts DEM, Frank MJ, Cools R, den Ouden HEM (2017): Catecholaminergic challenge uncovers distinct Pavlovian and instrumental mechanisms of motivated (in)action. ELife 6:e22169.
44. Ludvig EA, Sutton RS, Kehoe EJ (2008): Stimulus representation and the timing of reward-prediction errors in models of the dopamine system. Neural Comput 20:3034–3054.
45. Machado A (1997): Learning the temporal dynamics of behavior. Psychol Rev 104:241–265.
46. Hamilos AE, Spedicato G, Hong Y, Sun F, Li Y, Assad JA (2021): Slowly evolving dopaminergic activity modulates the moment-to-moment probability of reward-related self-timed movements. Elife 10:e62583.
47. O'Doherty J, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ (2004): Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science 304:452–454.
48. Botvinick MM, Niv Y, Barto AG (2009): Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. Cognition 113:262–280.
49. Coddington LT, Lindo SE, Dudman JT (2023): Mesolimbic dopamine adapts the rate of learning from action. Nature 614:294–302.
50. Long C, Lee K, Yang L, Dafalias T, Wu AK, Masmanidis SC (2024): Constraints on the subsecond modulation of striatal dynamics by physiological dopamine signaling [published online Jul 3]. Nat Neurosci.