1 Title

2 The Neural Basis of Delayed Gratification

3 Short Title

4 The Neural Basis of Delayed Gratification

5 Authors

6 Zilong Gao^{1,2#}, Hanqing Wang^{3#}, Chen Lu⁴, Tiezhan Lu^{1,2}, Sean Froudish-Walsh³,
7 Ming Chen⁴, Xiao-jing Wang^{3*}, Ji Hu^{4,5*}, Wenzhi Sun^{2,6*}

8 Affiliations

9 ¹ Academy for Advanced Interdisciplinary Studies, Peking University, Beijing
10 100871, China

11 ² Chinese Institute for Brain Research, Beijing 102206, China

12 ³ Center for Neural Science, New York University, New York, NY 10003, USA.

13 ⁴ School of Life Science and Technology, ShanghaiTech University, Shanghai
14 201210, China

15 ⁵ Shanghai Key Laboratory of Psychotic Disorders, Shanghai Mental Health Center,
16 Shanghai 200030, China

17 ⁶ School of Basic Medical Sciences, Capital Medical University, Beijing 100069,
18 China

19 [#] These authors contributed equally to this work

20 ^{*} Corresponding authors. Email: xjwang@nyu.edu, huji@shanghaitech.edu.cn &
21 sunwenzhi@cibr.ac.cn

Abstract

Balancing instant gratification versus delayed, but better gratification is important for optimizing survival and reproductive success. Although delayed gratification has been studied through human psychological and brain activity monitoring as well as animal research, little is known about its neural basis. We successfully trained mice to perform a waiting-for-water-reward delayed gratification task and used these animals in physiological recording and optical manipulation of neuronal activity during the task to explore its neural basis. Our results showed that the activity of dopaminergic neurons in the ventral tegmental area increases steadily during the waiting period. Optical activation or silencing of these neurons, respectively, extends or reduces the duration of waiting. To interpret this data, we developed a reinforcement learning (RL) model that reproduces our experimental observations. Steady increases in DAergic activity signal the value of waiting and support the hypothesis that delayed gratification involves real-time deliberation.

KEYWORDS

Delayed Gratification, Dopaminergic, Ventral Tegmental Area, Ramping Activity, Reinforcement Learning, Continuous Deliberation

TEASER

Sustained ramping of dopaminergic activation helps individuals to resist impulsivity and wait for a later but larger reward.

INTRODUCTION

To optimize survival and reproductive success, animals need to balance instant gratification versus delayed, but better gratification. Repeated exposure to instant gratification may disrupt this balance, thereby increasing impulsive decisions. Such decisions contribute to numerous human disorders, such as addiction and obesity(1, 2). Delayed gratification is an important process that balances time delay with increased reward (3). It is influenced by strengths in patience, willpower, and self-control(4). Psychologists and neuroscientists have long studied this important behavior through human psychological and brain activity assessments as well as rodent-based studies. Although the dopamine system has been implicated in delayed gratification, nevertheless the precise neural activity of dopamine system that allows better gratification has not been demonstrated. Additionally, no studies to date have causally manipulated the dopamine system during delayed gratification tasks.

During a well-controlled delayed gratification task, an individual must balance the benefits vs. risks of delay in receiving an available reward. The choice to continue waiting requires suppression of the constant temptation of an immediate reward, in favor of an enhanced reward in the future (3, 5, 6). Midbrain dopaminergic (DAergic) neurons are well known to play central roles in reward-related and goal-directed

behaviors (7-12). Studies have revealed that DAergic activity signals **spatial or operational** proximity to distant rewards (7, 13, 14), which has been postulated to sustain or motivate goal-directed behaviors while resisting distractions. DAergic neurons play important roles in time judgment (15) and cost-benefit calculations which are necessary for value-based decision making (13, 16-18).

We successfully trained mice to perform a waiting-**for**-water-reward delayed gratification task. Recording and manipulation of neuronal activities during this task **allowed** us to explore the cellular regulation of delayed gratification. We found that the activity of **ventral tegmental area** (VTA) DAergic neurons ramped up consistently while mice were waiting in place for rewards. Transient activation of DAergic neurons extended, **whereas** inhibition reduced the duration of the waiting period. Then we adopted reinforcement learning (RL) computational models to predict and explain our experimental observations.

RESULTS

Mice can learn to wait for greater rewards by delayed gratification task training

First, we trained **water-restricted** mice to perform a one-arm foraging task (pre-training) in which delay did not result in **an** increased reward (19). The period **in which** the mouse **remained** in the waiting zone was defined as the waiting duration, and the time **during which the** mouse **traveled** from the waiting zone to **reach** the water reward was **defined** as **the** running duration (Fig. 1A, left panel). When **the** mouse exited the waiting zone and licked the water port in the reward zone, it could

receive a 10 μ l of water drop regardless of the time spent in the waiting zone (Fig. 1A, right panel, black line). During a week of training, the average waiting and running durations both significantly decreased from Day 1 to Day 7 (Day1, Waiting: 5.58 ± 0.63 sec; Running: 3.46 ± 0.28 sec; $p < 0.001$; Day 7, Waiting: 1.99 ± 0.19 sec; Running: 1.28 ± 0.09 sec, $p < 0.001$, $n = 7$ mice, Friedman test, Figs. 1C-E, Movie S1). All mice learned the strategy of reducing durations of both waiting and running to maximize the reward rate (defined as μ l of water per second in a given trial, Fig. S1C).

Next, we trained the same mice using a delayed gratification paradigm, where the size of the reward increased quadratically with time spent in the waiting zone (Fig. 1A, right panel, green line). Over the next three weeks, this resulted in shifting the distributions of the waiting duration towards a longer wait. The averaged waiting period significantly increased from 2.76 ± 0.15 s on Day 1 to 4.62 ± 0.30 s on Day 15 ($p < 0.001$, $n = 7$ mice, Friedman test, Figs. 1F, H, and Movie S2), Continuous training steadily increased the averaged waiting duration, whereas the training did not change the average running duration from 1.19 ± 0.13 s on Day 1 to 1.24 ± 0.10 s on Day 15 ($p = 0.97$, $n = 7$ mice, Friedman test, Figs. 1G, H). The reward rate increased steadily, indicating that the mice were successfully learning to successfully delay gratification (fig. S1D).

The activity of VTA DAergic neurons increases steadily during the waiting period

To monitor the activity of VTA DAergic neurons during the delayed gratification task, we employed fiber photometry to record the calcium signals in VTA DAergic neurons in freely moving mice for as long as one month (Fig. 2A-C, optical fiber placement illustrated in Fig S2). On the first day of pre-task training, the calcium signal rose rapidly **upon** reward and quickly reached a peak. A few days of training dramatically reshaped the response pattern. Once the mice re-entered the waiting zone, the **activity** of DAergic neurons started to rise and reached the highest level when the animal received a reward (fig. **S3A**).

We next analyzed the activity of these same neurons in the mice as they learned the delayed gratification task. The recording traces showed that training gradually reshaped the pattern and time course of activity (Fig. 2D). The activity started to ramp up once the mice entered the waiting zone, and then reached its highest level when **they** exited. To investigate carefully the dynamical properties of the ramping activity during waiting, we sorted the calcium signals **from Day 15** of one mouse by their length of waiting durations and plotted them with a heat map (Fig. 2E). We divided trials according to the trial outcome (reward volume) and calculated the calcium signals while the mouse exited waiting zone with different reward volumes. Our results showed that the Z-scored calcium signals at 0.5 sec before exit were significantly **increased as** reward volumes **increased** ($F=24.67$, $p<0.01$, $n=7$, one-way ANOVA, Fig. 2F), but the mean signal curves followed similar **trajectories** regardless of trial outcome (Fig. 2G). Then, we calculated the slopes of signal curves with different outcomes over 4 time windows (0-2, 2-4, 4-6, 6-8 s) by linear regression analysis. The slopes during the same

time window **did not differ significantly** between reward groups (0~2s, $F=0.10$, $p=0.96$; 2~4s, $F=1.03$, $p=0.38$; 4~6s, $F=1.00$, $p=0.34$, $n=7$, one-way ANOVA, Fig. 2H). We pooled and plotted the slopes of different waiting periods together and found the activity curves kept rising steadily, almost **saturating** after 6 secs from the time the mice entered the waiting zone. **The ramp-up of DAergic activity** became less variable with delayed gratification task training in our experimental data (figs. S4A-D). All these results indicated **that** VTA DAergic neurons consistently ramp up their activity during waiting in as animals are trained in the delayed gratification task.

High-frequency spiking of VTA DAergic neurons sustains waiting in the delayed gratification task

Does tonic or phasic firing of DAergic neurons underlie the ramping calcium signal? To answer this, we conducted single unit recordings when mice were performing the delayed gratification task (Fig. 3A). A custom-made head plate was placed on the skull and affixed in place with dental cement. After recording, placements of recording electrodes were confirmed with electrolytic lesion inside VTA all 5 mice (Fig. 3B). We found 17 putative DAergic neurons displayed short bursts of firing during the waiting period (Fig. 3C). On trials in which the mice waited for a short duration the firing rate was low throughout the waiting period (Fig. 3D-E). On trials in which the mice waited for a long duration the firing rate was initially low before increasing during the later waiting period (Fig. 3D-E). We averaged Peri-stimulus time histograms (PSTHs) of all trials to obtain a response curve (Fig. 3F, $n=17$ cells). Similar to calcium signal, the response curve of firing rate noticeably ramped up with increased waiting

time, before reaching a plateau at around 4 seconds. To compare with the experimental calcium signal, we used a convolution algorithm to predict calcium responses trial by trial based on the firing rate of all 17 recorded cells (Fig. 3G showed the predicted calcium responses from Fig 3E). The average of predicted calcium signal ($\Delta F/F$ %, 0.5 sec before exiting from all trials) of all cells fit the measured calcium response (Z-scored $\Delta F/F$, 0.5 sec before exit from the last week of training) curve very well (Fig. 3H, $r=0.982$, Pearson correlation). This analysis shows that the spiking activity of single DAergic neurons underlies the ramping calcium signal revealed by fiber photometry recording.

Optogenetic manipulation of VTA DAergic activity **alters the waiting durations in **the** delayed gratification task**

To determine whether VTA DAergic activity controls performance in the delayed gratification task, we manipulated VTA DAergic neurons temporally **using optogenetic tools in** 20% of pseudo-randomly chosen trials while the mice were waiting during the delayed gratification task (Figs. **4A-C**). Activating the VTA DAergic neurons shifted the cumulative probability distribution **towards a** longer waiting duration (Fig. **4D**, blue, $F=12.93$, $p=0.002$, $n=6$ mice, one-way ANOVA), while inhibiting these same neurons shifted **the** distribution significantly **in the opposite direction** (Fig. **4E**, yellow, $F=7.76$, $p=0.008$, $n=6$ mice, one-way ANOVA). The effects of optogenetic stimulation or **inhibition** on the cumulative probability distributions **for waiting duration** were only observed in the Laser-ON trials. In contrast, the Laser-OFF trials, including **those immediately** after the Laser-ON trials **treated** as a single group, were not significantly

different from the trials from the previous day ($p > 0.5$, Figs. 4D-E). The optical manipulation didn't influence the running durations in the delayed gratification task in mice that expressed ChR2 or eNpHR 3.0 (figs. S5A-B), nor did it change the waiting duration distribution of mice that expressed mCherry in DAergic neurons in delayed gratification tasks (figs. S6A-B). To rule out the possibility of optogenetic manipulation-induced memory, we performed a random place preference test with the same stimulation dosage. Neither activating nor inhibiting VTA DAergic neurons significantly changed the transient waiting duration or pattern in the location in which the laser was activated in any of the tested mice (Figs. S5E-H) or in the mCherry expressing controls (figs. S6C-F).

A reinforcement learning (RL) model suggests that ramping up VTA DAergic activity signals the value of waiting for delayed gratification

How does a mouse manage to wait longer for a larger reward while ignoring smaller but more immediate reward options? We propose two models of behavioral scenarios that exemplify possible strategies a mouse may implement to achieve extended waiting performances: (i) setting a goal of expected waiting duration before the initiation of waiting or (ii) continuously deliberating during the waiting period. According to the first hypothesis (i), we modeled a reinforcement learning (RL) agent that keeps time until a predetermined moment has passed (Fig. 5A, *Decision Ahead*); according to (ii), we modeled a second RL agent that continuously balances the values of waiting versus leaving to control the decision to wait or to leave for the reward. To implement these models, we used a version of the state-action-reward-state-action (SARSA) algorithm

with a series of states (Fig. 5A, *Continuous Deliberation*, see methods) (20, 21). The behaviors of both models were able to replicate the behavioral performance we observed in animal experiments (Figs. 5B-C). The distributions of behavioral performances between early training days and late training days from the experimental data were very different from each other, such that the Kullback-Leibler (KL) divergence was large (0.39 ± 0.06). The KL divergences between the distributions of simulated behavioral performances and experimental data were significantly small to the large KL divergence value ($p = 0.005$, $n = 7$ mice, Friedman test), and there was no difference ($p > 0.99$) between Decision Ahead RL model and Continuous Deliberation RL model in either the early or late training session. (Fig. 5D). We couldn't determine which model is better based on behavioral performance alone, given that both models reproduced the behavioral data well.

What does the ramping up of DAergic activity mean in the delayed gratification task? We tried to explain it with our RL models. In the *Decision Ahead* model, the agent keeps time until the predetermined moment has passed, which suggests the ramping DAergic activity may relate to timing in the delayed gratification task. Some studies have proposed that the ramping activity is consistent with a role in the classical model of timing, with the movement being initiated when the ramping activity reaches a nearly fixed threshold value, following an adjusted slope of ramping activity(22-25). In contrast, our results showed that the DAergic activity ramped up to different values with similar trajectories on at a nearly constant slope (Figs. 2F-H). This suggests that VTA DAergic neurons may not implement a decision variable for the *Decision Ahead*

scenario. In the *Continuous Deliberation* RL model, we compared the curves of the value of waiting and leaving with the ramping up of DAergic activity and found the behavioral performance of both animals and model agents reached the asymptote. The values of waiting (Fig. 5E, purple, $r = 0.99$, Pearson correlation) and the leaving (Fig. 5E, green, $r = 0.91$, Pearson correlation) each correlated positively with the ramp of DAergic activity during waiting (Fig. 5E, blue, Z-scored $\Delta F/F$, 0.5 sec before exit from the last week of training). This detailed analysis suggested that the *Continuous Deliberation* RL model agreed with previous studies (13, 26-28) and that ramping DAergic activity signals the value of actions, either waiting or leaving, in the delayed gratification task.

In the *Decision Ahead* RL model, if the agent keeps timing during the waiting period through ramping DAergic activity to encode the elapsed of time (29-31), extra VTA DAergic activation should represent a longer time, thus leading to earlier cessation of waiting. This prediction is contrary to our optogenetics result that DAergic activation led to a longer waiting (Fig. 4D). Instead, we reproduced the optogenetic manipulations in the *Continuous Deliberation* RL model by either increasing or decreasing the value of waiting (Q_{wait}) in a pseudo-random 20% of trials. The increase or decrease in waiting durations only occurred in the Q_{wait} – manipulated trials ($p < 0.001$, Friedman test, $n=10$), whereas the remaining trials, including the next trials after value manipulation, did not differ significantly from controls ($p > 0.999$, Friedman test, $n=10$, Figs. 5F-G). Importantly, manipulating the value of leaving (Q_{leave}) in a pseudo-random 20% of trials induced the opposite results, inconsistent with the experimental data (figs.

S8A-B). Manipulating the reward prediction error (RPE) signal in the model led to persistent changes to the waiting time on subsequent trials (figs. S8C-D). This is inconsistent with the experimental finding that optogenetic manipulation of dopamine neuron firing only affects waiting time on the current trial (Figs 4D-E). Thus the ramping dopamine signal does is not consistent with an RPE signal in the delayed gratification task. Our experimental data and the *Continuous Deliberation* RL model together indicated that the ramping up of VTA DAergic activity profoundly influenced the waiting behavioral performance in the delayed gratification task, which suggested that this activity signals the value of waiting, rather than the value of leaving or reward prediction error. Conceptually, our analysis revealed that delayed gratification involves real-time deliberation.

VTA DAergic activity during waiting predicts the behavioral performance in the delayed gratification task

Our optogenetic manipulation experiments and *Continuous Deliberation* RL model indicated that VTA DAergic activity during waiting influenced the waiting durations while the mouse performed the delayed gratification task (Figs. 4D-E and Figs. 5F-G). Although the activity of VTA DAergic neurons ramped up consistently during waiting (Figs. 2G-H), it still fluctuated to a certain extent moment by moment. Therefore, we sought next to determine whether this fluctuation of DAergic activities influences the waiting behavior in the delayed gratification task. A strong prediction made by the *Continuous Deliberation* model is that, if DAergic activity signals the value of waiting at each specific moment, then a momentary increase in the DA signal

257 will make the agent more likely to keep waiting in the next "time bin," but not in
 258 subsequent time bins (fig. S8E-F). That is to say, the value of waiting is only positively
 259 correlated with the behavior of the immediately following time bin, which indicates the
 260 Markovian nature of the model (32). We thus aimed to test the relationship between the
 261 amplitude of the momentary VTA DAergic signal and the behavior (i.e., waiting or
 262 leaving) within each time bin to determine how the momentary DAergic activity (the
 263 calcium signal amplitude in 0-1 sec, 1-2 sec, 2-3 sec, or 3-4 sec after waiting onset,
 264 shown as each cluster of bars in Fig. 6B) affects the waiting performance in the
 265 subsequent periods (behavior from 1-2 sec, 2-3 sec, 3-4 sec, and 4-5 sec for DAergic
 266 activity from 0-1 sec; behavior within 2-3 sec, 3-4 sec and 4-5 sec for DAergic from in
 267 1-2 sec, and so on; Fig. 6A). To integrate data from multiple sessions as well as multiple
 268 animals, we took the advantage of the linear mixed model analysis (LMM, see Method)
 269 (33-35). We examined the correlation between momentary DAergic activity and
 270 behavior (i.e. the momentary binary waiting decision). For the behavior, we coded
 271 sustained waiting as 1 and leaving in that time bin as 0 (trials that stopped before the
 272 examined time window were not taken into account) for different pairs of time bins.
 273 Ten independent LMM analyses were done for each activity-behavior pair, as indicated
 274 by the dash lines in Fig. 6A. The regression coefficients, as well as the confidence
 275 intervals of each of the ten pairs, are shown by corresponding bars in Fig. 6B. The
 276 correlation is only significantly positive between adjacent DAergic and waiting bins
 277 ($p < 0.001$, $n = 7$ mouse, black lines, regressed coefficient median; boxes, 50% confidence
 278 interval; whisker, 95% confidence interval, Fig. 6B, the first bar of each cluster, where

the DA activity bin is 1sec ahead of the waiting bin). There was no significant correlation between DAergic activity at 3-4 sec and behavior at 4-5 sec ($r=0.007$, $p=0.61$, $n=7$), which may result from insufficient data for those long trials. Apart from these pairs with adjacent bins, other pairs of DAergic activity and waiting behavior didn't show any significant correlation (Fig. 6B, the remaining bars of each cluster). These results indicate that the waiting decision of the current moment is only influenced by the most recent DAergic signal but not by DAergic signal further in the past, which suggests that deliberation for waiting in delayed gratification may be a Markov process as we formalized in the *Continuous Deliberation* RL model(32).

In the *Continuous Deliberation* RL model, the probability of waiting (P_w) positively correlates with the value of waiting (Q_{wait}). To explore the impact of DAergic activity on the probability of waiting in our experimental data, we binned DAergic activity of every trial and normalized data points (V_{DA}) in each momentary DAergic period (with each period lasting 1 sec and starting from 0 to 9 sec, as shown in northeast figure of Fig. 6C). Then, we divided the trials into two groups by setting a series of arbitrary thresholds (red, High DAergic activity, $V_{DA-Z} \geq Th$; green, Low DAergic activity, $V_{DA-Z} \leq -Th$, where Th was the threshold for the analysis of high/low DAergic activity) from these trials (Fig. 6C, Th was set to 0.9). We analyzed the P_w of low or high DAergic activity trials for the adjacent waiting period with different thresholds. In doing so, we found that the probability of waiting increased rapidly as the absolute value of the threshold increased. The P_w of High DA and Low DA activity trials fit well with a fifth-degree polynomial function ($R^2=0.93$, $-2.1 \leq \text{threshold} \leq 2.1$). When the absolute

values of the threshold are large enough ($|Th| \geq 1.7$), the Pw of the High DA activity trials is significantly ($p=0.04$, $F(1,12)=5.483$, Two-way ANOVA) higher than the Pwait of the Low DA-ramping activity trials in adjacent waiting periods ($|threshold|=2.0$, $p=0.02$; $|threshold|=2.1$, $p<0.001$, Sidak's multiple comparisons test, $n=7$, Fig. 6D).

Finally, we investigated the influence of fluctuations of intrinsic VTA DAergic activity on the waiting performance of mice in the delayed gratification task. There were certain trials in which DAergic activity across the whole duration of waiting was significantly higher (red, high-ramping) or lower (green, low-ramping) than the mean DAergic activity (see Methods). We then got two groups of trials and found that the cumulative distribution of waiting durations in high-ramping trials shifted to the right, with significantly higher normalized waiting durations (1.03 ± 0.01) compared to those low-ramping trials (0.98 ± 0.01 , $p=0.024$, $n=7$, paired Student's t-test, Fig. 6E), but there was no difference between the normalized waiting durations for the trials immediately following the high-ramping and low-ramping trials (next trial of H-R, 1.01 ± 0.01 ; next trial of L-R, 0.99 ± 0.01 ; $p=0.290$, $n=7$, paired Student's t-test, Fig. 6F). These results accorded with our optogenetic manipulation experiment (Figs. 4D-E), which indicated that optogenetically manipulated VTA DAergic activity transiently influences the duration of waiting in the delayed gratification task.

DISCUSSION

Here we reported a novel behavioral task in which the mice were trained to learn a foraging task with a delayed gratification paradigm. Mice learned to wait for bigger

rewards **that they received after** waiting for longer (Figs. 1F-H). **On a neural level, we**
found that the calcium signal of VTA DAergic neurons ramped up consistently when
mice waited in place before taking action to fetch the expected reward (Figs. 2G-H).
Further data analysis showed that the ramping VTA DA activity indeed influenced the
behavioral performance of waiting (Figs. 6B-E), which was confirmed with bi-
directional optogenetic manipulations of VTA DAergic activity (Figs. 4D-E). Finally,
we developed a RL model that predicted our experimental observations well and
consolidated the conclusion that the ramping up of VTA DAergic activity signaled the
value of waiting in the delayed gratification task, which involved real-time deliberation
(Figs. 5B-G).

DA release in the **nucleus accumbens** (NAc) was previously conjectured to sustain
or motivate the goal-directed behavior as well as resistance to distractions (*13, 14*).
Here, we explicitly implemented continuous less-optimal options during the delayed
gratification process, in which, to achieve better performance, mice needed to sustain
waiting as well as prevent/control impulsivity (*3, 6, 36, 37*). We found remarkable and
sustained DAergic activation when mice managed to wait longer, and further
demonstrated a causal link between DAergic activation and the increase in transient
waiting probability. Furthermore, we found DAergic activity ramps up in a consistent
manner during waiting, mimicking the value of waiting along with a series of states in
our *Continuous Deliberation* RL model, **both of which presumably contributed to**
pursuing a more valuable future goal and resisting the distraction of the less-optimal
immediate options in our task. Intriguingly, the momentary DAergic activity was found

to correlate positively with the momentary waiting probability, which also suggested DAergic activity may be involved in the continuous deliberation process. Therefore, we not only for the first time to our knowledge demonstrated the behavioral significance of DAergic activity in delayed gratification, but also depicted "a continuous deliberation" framework in which DAergic activity may participate to help achieve more flexible and sophisticated performance.

Numerous works have used Pavlovian conditioning in studying DAergic activity(10, 12, 38-40). Some studies paired the reward with a cue (or cues), such that the animals don't need to perform effortful work to obtain rewards. It is well known that this kind of DAergic activity signals the RPE via phasic firing. In the studies using operant conditioning or goal-directed behavior, the animals have to perform actions and need effortful work to obtain outcomes, and a ramping up of DAergic activity was reported to emerge while the animals were approaching the reward (13, 14, 41, 42). This ramping activity was suggested to signal the value of work (13) or a distant reward (14), but key evidence is lacking because the change of sensory input flow dramatically alters the DAergic activity over time. Under such mutual influence, it is impossible to identify the RPE or the value of work from external cues. The RPE model of ramping activity assumes that the value increases exponentially (or at least in a convex curve) as the reward is approached. Under this model, sensory feedback is suggested to result in the RPE signal ramping up (41, 43, 44), while a lack of sensory feedback is predicted to make a flat RPE signal. In contrast, the ramping up of DAergic activity is well isolated from the external sensory inputs when performing the delayed gratification task

in our model. Mice continuously deliberate about the current state and future rewards without any external sensory inputs while waiting in place. Despite the lack of external sensory inputs, we still observed the calcium signal of VTA DAergic neurons ramping up in a consistent manner, functionally mimicking an inner variable of the evolving value of waiting. This observation is consistent with the hypothesis of dopamine signaling the value that is related to time and effort investment under certain circumstances (13), but cannot be immediately explained by an RPE response to external sensory inputs.

Midbrain DAergic neurons play an important role in reinforcement learning(9, 11, 12, 45, 46), where activation of DAergic neurons usually produces a reinforcement effect on an associated action, stimulus, or place. But in our delayed gratification task, optogenetic manipulation of DAergic activity substantially influenced the ongoing behavior during the current trial without a visible reinforcement effect on later trials. Notably, this optogenetic manipulation was not sufficient to induce a reinforcement effect in the random place performance test. These results revealed the distinct and potent instantaneous effect of DAergic activity during delayed gratification. By simulating transient manipulations of variables in the RL model, we showed that manipulating dopamine activity was equivalent to manipulating the value of waiting in the model. In contrast, manipulating the value of leaving, or the RPE signal itself caused dramatically different effects on behavior. The observations and analysis in our experiments suggest that the value of waiting is represented in VTA DAergic neurons during a delayed gratification task. This significantly updates the understanding of the

coding mechanisms and fundamental functions of the DAergic system in delayed gratification. Our results suggested that DAergic neuron stimulation during the RPPT test is not rewarding, but does lead to a shift in wait time. This control experiment was performed to exclude the rewarding effect of our DAergic stimulation paradigm during the DG task. However, such a test may be not strictly comparable with stimulation during the delayed gratification task due to the differences in the behavioral contexts.

The previous finding suggested that reasonable behavior in the face of instant gratification requires suppression of reflexive reward desiring. Human brain imaging results demonstrated hemodynamic responses to conditioned (rewarding) stimuli in both the NAc and the VTA were significantly attenuated during the desire–reason dilemma(47). Such discrepancies with our results may be the consequence of different behavioral strategies, in which they measured the reward-related activation during a desire–reason dilemma, and we measured the DA activity during the waiting time. Also, the different temporal and spatial resolution of human brain imaging and fiber photometry and electrophysiology may lead to the discrepancies, and our ramping pattern of DAergic activity also indicated there are inhibitory tones in the NAc and VTA during the beginning phase of wait.

The design of our delayed gratification task recapitulates the realistic situation where immediate less-valuable choices lie in the way of pursuing a later but possibly larger benefit. A deficit in the ability to resist immediate reward for the delayed but possibly larger reward is closely related to a variety of disorders like obesity, gambling, and addiction(1, 48). The ramping VTA DAergic activity accords with a model of

410 NOW vs LATER decisions in which DAergic signals have a strong influence on the
 411 **prefrontal cortex in favoring** LATER rewards(2). We propose that the sustained phasic
 412 VTA DAergic activity during the delay period could serve as a neural basis for the
 413 power to resist a temptation close at hand, and improve reward rate or goal pursuit in
 414 the long run.

MATERIALS AND METHODS

Mice. Animal care and use strictly followed institutional guidelines and governmental regulations. All experimental procedures were approved by the IACUC at the Chinese Institute for Brain Research (Beijing) and ShanghaiTech University. Adult (8-10 weeks) DAT-IRES-Cre knock-in mice (Jax stock# 006660) were trained and recorded. Mice were housed under a reversed 12/12 day/night cycle at 22–25°C with free access to ad libitum rodent food.

Stereotaxic viral injection and optical fiber implantation. After deep anesthesia with isoflurane in oxygen, mice were placed on the stereoscopic positioning instrument. Anesthesia remained constant at 1~1.5% isoflurane supplied per anesthesia nosepiece. The eyes were coated with aureomycin eye cream. The scalp was cut open, and the fascia over the skull was removed with 3% hydrogen peroxide in saline. The Bregma and Lambda points were used to level the mouse head. A small window of 300~500µm in diameter was drilled just above the VTA (AP: -3.10 mm, ML: ±1.15mm, and DV: -4.20 mm) for viral injection and fiber implantation. 300 nl of AAV2/9-hSyn-DIO-GCamp6m (10^{12}) solution was slowly injected at 30nl /min unilaterally for fiber photometry recording. Either 300 nl of AAV2/9-EF1a-DIO-hChR2(H134R)-mCherry (10^{12}) or 300 nl of AAV2/9-EF1a-DIO-eNpHR3.0-mCherry (10^{12}) was injected bilaterally for optogenetic experiments. The injection glass pipette was tilted at an angle of 8° laterally to avoid the central sinus. After injection, the glass pipette was kept in place for 10 min and then slowly withdrawn. An optical fiber (200 µm O.D., 0.37 NA; Anilab) hold in a ceramic ferrule was slowly

inserted into the brain tissue with the tip slightly above the viral injection sites. The fiber was sealed to the skull with dental cement. Mice were transferred to a warm blanket for recovery, then housed individually in a new home until all experiments were done.

Behavioral tasks. One week after surgery, mice started a water restriction schedule to maintain 85–90% of free-drinking bodyweight for 5 days. The experimenter petted the mice 5 minutes per day for 3 days in a row and then started task training. All behavioral tasks were conducted during the dark period of the light/dark cycle.

The foraging task shuttle box had two chambers (10×10×15 cm) connected by a narrow corridor (45×5×15 cm, Fig. 1A). A water port (1.2 mm O.D. steel tube, 3 cm above the floor) was attached to the end of one chamber, defined as the reward zone, with the other as the waiting zone. The position of the mouse in the shuttle box was tracked online with a custom MATLAB (2016b, MathWorks) program through an overhead camera (XiangHaoDa, XHD-890B). The experimental procedure control and behavioral event acquisition were implemented with a custom MATLAB program and an IC board (Arduino UNO R3).

One-arm foraging task (pre-training): A water-restricted mouse was put in the shuttle box for free exploration for up to 1 hour. When the animal traveled from the waiting zone through the corridor to the reward zone to lick the water port, 10 µl water was delivered by a step motor in 100 ms as a reward. A capacitor sensor monitored the timing and duration of licking. The animal's return to the waiting zone to initiate the

next trial. Exiting from the waiting zone triggered an auditory cue (200 ms at 4 kHz sine wave with 90 dB) to signal this exit from the waiting zone. The time spent in the waiting zone was defined as the waiting duration. The training was conducted every day for a week. All mice learned to move quickly back and forth between the two chambers to maximize the reward rate within one week.

Delayed gratification task. From the second week, the volume of water reward was changed to a function proportional to the waiting time: a wait time of 0-2 s for 0 μ l; 2-4 s triggered delivery of, 2 μ l; 4-6 s, 6 μ l; 6-8 s, 18 μ l; and >8 s, 30 μ l as shown in Fig. 1A. The training was conducted five days a week, from Monday to Friday.

P_w Calculation. We divided all trials into two groups, Waiting Trials and Leaving Trials according to whether the animal remained to wait or left during a given time interval, such as 1 sec after each behavioral period. Then, we calculated the probability of waiting (P_w) in this given time interval by the number of 'Waiting Trials' (N_{w(n)}) and the number of 'Leaving Trials' (N_{L(n)}) in the time window n:

$$P_{w(n)} = \frac{N_{w(n)}}{N_{w(n)} + N_{L(n)}}$$

Then we could calculate the P_w for a given time duration:

$$P_{w(n)} = \frac{\sum_0^9 N_{w(n)}}{\sum_0^9 N_{w(n)} + \sum_0^9 N_{L(n)}}$$

Linear Mixed Model. We implemented the Linear Mixed Model Analysis using the open-source Python package "statsmodels" (https://www.statsmodels.org/stable/mixed_linear.html). The binary value of waiting or leaving during a specific behavioral period t_{beh} was set as the dependent factor

(t_{beh} =[1, 2), [2, 3), [3, 4), or [4, 5), unit: second); the fluctuation of momentary DA signal from its mean during a preceding period t_{DA} was set as a fixed-effects (t_{DA} =[0, 1), [1, 2), [2, 3), [3, 4), unit: second. Note that t_{DA} is always smaller than t_{beh}); the animal identity and session numbers were set as a random effect ($n=5$ for each animal from the third week). The parameters of the model were estimated by restricted maximum likelihood estimation (REML).

Optogenetic stimulation. Lasers, 473 nm wavelength for activation and 589 nm for inhibition, were coupled to the common end of a patchcord (200 μ m O.D., 1-m long, 0.37 NA). The patchcord split through an integrated rotatory joint into two ends connected to optical fibers implanted as described above (200 μ m O.D., 0.37 NA) for bilateral light delivery. First, the mice were trained for 3 weeks to learn the delayed gratification task. Optical stimulation was delivered pseudo-randomly in ~20% of behavioral trials in the test experiment. Square pulses of 20 ms at 10 Hz for activation or a continuous stimulation for inhibition were delivered. The laser was set to ON when the animal entered the reward zone and to OFF upon exiting from the reward zone. The maximal laser stimulation was no longer than 16 seconds, even if in the case a mouse stayed in the waiting zone longer than this time. Continuous laser power at the tip of the splitting patchcord was about 10 mW for the 473 nm laser and 8 mW for the 589 nm laser, respectively.

Random place performance test. After finishing optogenetic tests for delayed gratification, all mice took a random place performance test (RPPT). The RPPT was carried on in a rectangular apparatus consisting of two chambers (30×30×30 cm)

separated by an acrylic board. With an 8 cm wide door open, the mice could move freely between the two chambers. Before testing, each mouse was placed into the apparatus for 5-min free exploration. The RPPT consisted of two rounds of 10-min tests. First, we randomly assigned one chamber as a test chamber. Laser pulses were delivered with 20% possibility (in accord with the setting of laser pulses delivering in delayed gratification task) while the mouse entered the test chamber. The delivery of light, no longer than 16 sec, stopped while the mouse exited the test chamber. Next, we switched the chamber in which laser pulses were delivered. The laser output power and pulse length were set the same as in the optogenetic manipulations in the delayed gratification task. In this task, we analyzed the time in each chamber with or without laser delivered.

Fiber photometry recording. During the behavioral task training and test, we recorded the fluorescence signal of VTA dopaminergic (DAergic) neurons. The signal was acquired with a fiber photometry system equipped with a 488 nm excitation laser and a 505~544 nm emission filter. The GCaMP6m signal was focused on a photomultiplier tube (R3896 & C7319, Hamamatsu) and then digitalized at 1 kHz and recorded with a 1401 digitizer and Spike2 software (CED, Cambridge, UK). An optical fiber (200µm O.D., 0.37 NA, 1.5-m long, Thorlabs) was used to transfer the excitation and emission light between recording and brain tissue. The laser power output at the fiber tip was adjusted to 5~10 µW to minimize bleaching.

All data were analyzed with custom programs written in MATLAB (MathWorks). First, we sorted the continuously recorded data by behavioral trials. For each trial, the data spanned the range between 1 s before the waiting onset and 2 s after the reward.

Before hooking up the fiber to the mouse, we recorded 20 s of data and averaged as F_b as the ground baseline. For each trial, we averaged 1-sec of data before the waiting onset as the baseline F_0 and then calculated its calcium transient as:

$$\Delta F/F (\%) = (F - F_0)/(F_0 - F_b) \times 100 (\%)$$

In the correlation analysis between VTA DAergic activity before waiting and the waiting duration of mice, we used averaged 1-sec data before the waiting onset as the DAergic activity before waiting.

In the analysis of high-ramping and low-ramping DAergic activity, we compared the whole calcium signal of every trial with the average curve (the same length as the analyzed calcium signal) of all trials from one mouse in a single training day with paired **a t-test and then separated their waiting times into high-ramping group and low-ramping group.**

To facilitate presenting the data, we divided each trial data into four segments, including 1 s before waiting onset, waiting, running, and 2 s after rewarding. For comparing the rising trends, we resampled the data segments at 100, 100, 50, and 100 data points, respectively. In the delayed gratification task, the trial data were aligned to the waiting onset and presented as the mean plots with a shadow area indicating SEM of fluctuations.

In vivo electrophysiological recording. A custom-made head plate was placed on the skull of each mouse and affixed in place with dental cement. We removed the skull and dural carefully above the recording window before the implantation. Stereotrodes were

twined from 12.7 mm Ni-Cr-Fe wires (Stablohm 675, California Fine Wire, CA, USA). Then 8 stereotrodes were glued together and gold plated to reduce impedance to 250–500 kΩ. The stereotrodes were gradually lowered to a depth of 0.7 mm above the VTA. A silver wire (127 mm dia, A-M System) was attached to one of the four skull-penetrating M1 screws to serve as ground(19). Mice were allowed a recovery time of 7 days. Extracellular spiking signals were detected with 8 stereotrodes and amplified (1000×) through a custom-made 16-channel amplifier with built-in band-pass filters (0.5–6 kHz). We selected one channel that did not show spike signals and defined it as a reference ground to reduce movement artifacts. Analog signals were digitized at 25 kHz and sampled by a Power1401 digitizer and Spike2 software (Cambridge Electronic Design). Spikes recorded by the stereotrode were sorted offline using Spike2 software (Cambridge Electronic Design). Classified single units should have a high signal-to-noise ratio (>3:1), reasonable refractory period (interspike interval >1 ms), and relatively clear boundaries among different PCA clusters. The spike frequency and waveform were used to determine cell type as the DA or GABA neurons. The putative DA neurons were identified by their relatively low firing rate (the mean firing rate<15Hz), and a broad initial positive phase of >1 ms. Then, the spike trains were aligned with the waiting onset in delayed gratification task. Peri-stimulus time histograms (PSTHs) (bin width = 100 ms) for each trial were calculated and presented with averages in plots.

Calcium signal simulation. We performed a simple simulation of the “calcium fluorescence signal” contributed by each recorded unit using kernel convolution with

0.1s time bin (49). The kernel we used is comprised of a linear rising edge and exponential decay. We set the kernel parameters, namely the rise time and half-peak decay time, to 0.2s and 0.7s respectively, from the relationship between fluorescence signal and a single action potential, without concerning the non-linear effect of multiple action potentials (50)

Reinforcement learning model. We investigated two potential scenarios. One was that the mouse decided on a waiting duration before entering the waiting area, and then waited according to the decided goal. The other scenario was that the mouse entered the waiting zone, and determined whether to wait or leave as an ongoing process throughout the whole waiting period. We called these two scenarios "*Decision Ahead*" and "*Continuous Deliberation*", respectively, and formulated corresponding reinforcement learning based models for simulation using Python (Python Software Foundation, version 2.7. Available at <https://www.python.org/>).

Decision Ahead. Inspired by animal behavior, we simply set three optional "actions" with different expected waiting durations that could empirically cover the main range of animals' waiting durations seen during training ($T_{a1} = 1.65$ sec for action1, $T_{a2} = 2.72$ sec for action2, $T_{a3} = 4.48$ sec for action3). These waiting durations were equally spaced on the log-time axis, consistent with Weber's law (that is, $\ln(T_{a1}) = 0.5$, $\ln(T_{a2}) = 1$, $\ln(T_{a3}) = 1.5$). During the execution of action a_i , we imposed additional noise on the timing so that the actual waiting time τ_{ai} for action a_i followed a Gaussian distribution on the log-time axis centered at the T_{ai} , $\ln \sim \mathcal{N}(\ln(T_{ai}), 0.4^2)$, $i = 1, 2, 3$. These settings allowed us to best capture the animals' waiting performance in the model. For

each trial, the agent chose an action randomly based on the three action values and a Boltzmann distribution (Softmax):

$$P_{a_i} = \frac{e^{\beta Q_{a_i}}}{\sum_{j=1,2,3} e^{\beta Q_{a_j}}}$$

Where P_{a_i} was the probability of choosing action a_i and waiting for τ_{ai} . Q_{a_i} was the value for a_i . β was the inverse temperature constant tuned to 5 to best fit the animal experimental data. After waiting, the agent would get a reward according to the same reward schedule used in our experiment. Each action value was updated separately during the reward delivery:

$$\delta = r - Q_a$$

$$r = R/(\tau + 1)$$

$$Q_a \leftarrow Q_a + \alpha * \delta$$

Where the reward prediction error δ was calculated by the difference between the hyperbolically discounted reward r (or "reward rate", given by the absolute reward R dividing total time $\tau+1$ for obtaining the reward, where τ was the waiting duration and the additional 1sec was the estimated delay of running between the two zones) and the chosen action value Q_a . The reward prediction error was then used to update the value of the chosen action. We tuned the learning rate α to 0.002 to fit the animal behavioral data.

Continuous Deliberation. In each trial, the agent would go through a series of hidden states, each lasting for 0~2sec randomly according to a Gaussian distribution (mean at

1 sec). At each hidden state, the agent had two action options, either to keep waiting or to leave. If it chose to keep waiting, the agent would transition to the next hidden state, with the past time of the previous state cumulated to the whole waiting duration. If the choice was to leave, the cumulation would cease and a virtual reward dependent on the duration was delivered; a new trial would then begin from the initial state. The reward schedule was identical to that used for the animals during the experiments.

The action choice for the future was determined randomly by a Boltzmann distribution (SoftMax) and action values:

$$P_{a_w}^{(T+1)} = \frac{e^{\beta Q_{a_w}^{(T+1)}}}{e^{\beta Q_{a_w}^{(T+1)}} + e^{\beta Q_{a_L}^{(T+1)}}}$$

$P_{a_w}^{(T+1)}$ was the probability of choosing to wait for the next state $T + 1$. $Q_{a_w}^{(T+1)}$ and $Q_{a_L}^{(T+1)}$ were the value of waiting and leaving, respectively, for state $T + 1$. β was the inverse temperature constant tuned to 5 to best fit the animal experimental data. The action values for each hidden state T were updated by a temporal difference learning algorithm (SARSA):

$$\delta = r + \gamma * Q_{a'}^{(T+1)} - Q_a^{(T)}$$

$$r = R/(\tau + 1)$$

$$Q_a^{(T)} \leftarrow Q_a^{(T)} + \alpha * \delta$$

Where the future action a' was determined by the Boltzmann distribution in the previous step. The current action a and the future action a' could both be either waiting or leaving. The prediction error δ was calculated by the sum of reward rate r

(r remained zero until the reward R was delivered. $\tau+1$ was the total time for obtaining the reward, where τ was the waiting duration and the additional 1sec was the estimated delay of running between the two zones) and the future action value $\gamma * Q_{a'}^{(T+1)}$ discounted by γ ($\gamma = 0.9$), minus the current action value $Q_a^{(T)}$. When a was leaving, the future action value $Q_{a'}^{(T+1)}$ would always be zero. This error signal δ was used to update $Q_a^{(T)}$ with the learning rate $\alpha = 0.001$.

As a **Markovian** process, each state would be identical to the agent no matter how the state was reached or what the following actions might be. So, we extracted the learned value of waiting as a time series along all the hidden states to compare with the averaged curve of VTA DAergic activity. For each trial, we also extracted the time series of the transient waiting value for a trial-wise analysis. Apart from the value of waiting, we could also extract the time series of RPE for each trial.

We simulated optogenetic manipulation in the model after normal training was accomplished as in the animal experiments.

Value manipulation. In 20% of trials in the simulation session, the future waiting value throughout the whole waiting period was manipulated. The optogenetics activation was simulated as an extra positive value added onto the future waiting value, and the optogenetics inhibition corresponded to a proportional decrease of the future waiting value as follows:

$$Q_{a_w}^{(T+1)} \leftarrow \tilde{Q}_{a_w}^{(T+1)}, \quad \text{for the current trial}$$

$$\text{where } \tilde{Q}_{a_w}^{(T+1)} = \begin{cases} Q_{a_w}^{(T+1)} + \Delta_{value-ext}, & \text{if "Chr2 - lighton"} \\ \kappa_{value-inh} Q_{a_w}^{(T+1)}, & \text{if "eNPHR - lighton"} \end{cases}$$

$$\text{and, } \delta = r + \gamma * \tilde{Q}_{a_w}^{(T+1)} - Q_a^{(T)}, \quad \text{if } a' = a_w$$

Here we set $\Delta_{value-ext} = 0.15$, and $\kappa_{value-inh} = 0.9$, so that the change in averaged waiting duration in the simulated "light-on" trials could capture the magnitude of the instantaneous effect of optogenetic stimulations on the current trials. Using these parameters "calibrated" by the current trial effect, we were able to compare the stimulation effect on the light-off or the following trials in both real and simulated situations. Also note that if the future action was chosen as waiting, the manipulated value of waiting would be used in the RPE calculation, and thus current action value updating as well.

RPE manipulation. Under this situation, in 20% of trials in the stimulation session, instead of the future waiting value, RPE (δ) was manipulated throughout the whole waiting period as follows:

$$\tilde{\delta} = \begin{cases} \delta + \Delta_{RPE-ext}, & \text{if "Chr2 - lighton"} \\ \delta - \Delta_{RPE-inh}, & \text{if "eNPHR - lighton"} \end{cases}$$

$$\text{and, } Q_a^{(T)} \leftarrow Q_a^{(T)} + \alpha * \tilde{\delta}$$

we set $\Delta_{RPE-ext} = 15$, and $\Delta_{RPE-inh} = 20$, which was calibrated by the current trial effect of real light stimulation.

To simulate the fluctuation in real DAergic signal, we simply multiplied the future waiting value during each state by a factor of $\sigma \sim \mathcal{N}(1, 0.3^2)$ (determined by the averaged signal-dependent noise magnitude / relative standard deviation for all

momentary DAergic amplitudes), as an addition to the original model (this is only implemented for Figs. S8E~F).

Electrophysiological recordings. Adult (8-10 weeks) DAT-IRES-Cre knock-in male mice 4 weeks after injection with AAV2/9-EF1a-DIO-ChR2(H134R)-mCherry or AAV-DIO-eNpHR3.0-mCherry were anesthetized with an intraperitoneal injection of pentobarbital (100 mg kg⁻¹) and then perfused transcardially with ice-cold oxygenated (95% O₂/5% CO₂) NMDG ACSF solution (93 mM NMDG, 93 mM HCl, 2.5 mM KCl, 1.25 mM NaH₂PO₄, 10 mM MgSO₄·7H₂O, 30 mM NaHCO₃, 25 mM glucose, 20 mM HEPES, 5 mM sodium ascorbate, 3 mM sodium pyruvate, and 2 mM thiourea, pH 7.4, 295-305 mOsm). After perfusion, the brain was rapidly dissected out and immediately transferred into an ice-cold oxygenated NMDG ACSF solution. Then the brain tissue was sectioned into slices horizontally at 280 μm in the same buffer with a vibratome (VT-1200 S, Leica). The brain slices containing the VTA were incubated in oxygenated NMDG ACSF at 32°C for 10~15 min, then transferred to a normal oxygenated solution of ACSF (126 mM NaCl, 2.5 mM KCl, 1.25 mM NaH₂PO₄, 2 mM MgSO₄·7H₂O, 10 mM Glucose, 26 mM NaHCO₃, 2 mM CaCl₂) at room temperature for 1h. A slice was then transferred to the recording chamber, then submerged and superfused with ACSF at a rate of 3 ml/min at 28°C. Cells were visualized using infrared DIC and fluorescence microscopy (BX51, Olympus). VTA DAergic neurons were identified by their fluorescence and other electrophysiological characteristics. Whole-cell current-clamp recordings of VTA DAergic neurons were made using a MultiClamp 700B amplifier and Digidata 1440A interface (Molecular

Devices). Patch electrodes (3-5 MΩ) were backfilled with internal solution containing (in mM): 130 K-gluconate, 8 NaCl, 10 HEPES, 1 EGTA, 2 Mg·ATP and 0.2 Na₃·GTP (pH:7.2, 280 mOsm). Series resistance was monitored throughout the experiments. For optogenetic activation, blue light was delivered onto the slice through a 200-μm optical fiber attached to a 470 nm LED light source (Thorlabs, USA). The functional potency of the ChR2-expressing virus was validated by measuring the number of action potentials elicited in VTA DAergic neurons using blue light stimulation (20 ms, 10 Hz, 2.7 mW) in VTA slices. For optogenetic inhibition, yellow light (0.7 mW) was generated by a 590 nm LED light source (Thorlabs, USA) and delivered to VTA DAergic neurons expressing eNpHR3.0 through a 200-μm optical fiber. To assure eNpHR-induced neuronal inhibition, whole-cell recordings were carried out in current-clamp mode and spikes were induced by current injection (200 pA) with the presence of yellow light. Data were filtered at 2 kHz, digitized at 10 kHz, and acquired using pClamp10 software (Molecular Devices).

Immunostaining. Mice were deeply anesthetized with pentobarbital (100 mg/kg, i.p.), followed by saline perfusion through the heart. After blood was drained out, 4% paraformaldehyde (PFA) was used for fixation. Upon decapitation, the head was soaked in 4% PFA at room temperature overnight. The brain was harvested the next day, post-fixed overnight in 4% PFA at 4°C, and transferred to 30% sucrose in 0.1 M PBS, pH 7.4 for 24~48 h. Coronal sections (20 μm) containing the VTA were cut on a cryostat (Leica CM3050 S). The slides were washed with 0.1 M PBS, pH 7.4,

incubated in blocking buffer (0.3% Triton X-100, 5% bovine serum albumin in 0.1 M PBS, pH 7.4) for an hour, and then transferred into the primary antibody (rabbit anti-tyrosine hydroxylase antibody, 1:1,000; Invitrogen) in blocking buffer overnight at 4°C. The sections were washed three times in 0.1 M PBS, then incubated with donkey anti-rabbit IgG H&L secondary antibody (conjugated to fluor-488 or fluor-594, 1:1,000; Jackson ImmunoResearch) at room temperature for 2 h. The nucleus was stained with DAPI (4',6-diamidino-2-phenylindole). Sections were mounted in glycerine and covered with coverslips sealed in place. Fluorescent images were collected using a Zeiss confocal microscope (LSM 880).

Quantification and statistics. All statistics were performed by MATLAB (R2016b, MathWorks) and Python (V2.7, Python Software Foundation) routines. Data were judged to be statistically significant when the P-values were less than 0.05. Asterisks denote statistical significance * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$. Unless stated otherwise, values were presented as Mean $\pm$ s.e.m..

REFERENCES AND NOTES

1. D. Tomasi, N. D. Volkow, Striatocortical pathway dysfunction in addiction and obesity: differences and similarities. *Crit Rev Biochem Mol Biol* **48**, 1-19 (2013).
2. N. D. Volkow, R. D. Baler, NOW vs LATER brain circuits: implications for obesity and addiction. *Trends Neurosci* **38**, 345-352 (2015).
3. W. Mischel, Y. Shoda, M. I. Rodriguez, Delay of gratification in children. *Science* **244**, 933-938 (1989).
4. J. E. Maddux, J. P. Tangney, *Social psychological foundations of clinical psychology* (Guilford Press, New York, 2010), pp. xv, 555 p.
5. J. Grosch, A. Neuringer, Self-control in pigeons under the Mischel paradigm. *J Exp Anal Behav* **35**, 3-21 (1981).
6. B. Reynolds, H. de Wit, J. B. Richards, Delay of gratification and delay discounting in rats. *Behav Process* **59**, 157-168 (2002).

- 739 7. B. Engelhard, J. Finkelstein, J. Cox, W. Fleming, H. J. Jang, S. Ornelas, S. A. Koay,
740 S. Y. Thiberge, N. D. Daw, D. W. Tank, I. B. Witten, Specialized coding of sensory,
741 motor and cognitive variables in VTA dopamine neurons. *Nature*, (2019).
- 742 8. C. K. Starkweather, B. M. Babayan, N. Uchida, S. J. Gershman, Dopamine reward
743 prediction errors reflect hidden-state inference across time. *Nat Neurosci* **20**,
744 581-589 (2017).
- 745 9. P. W. Glimcher, Understanding dopamine and reinforcement learning: the
746 dopamine reward prediction error hypothesis. *Proc Natl Acad Sci U S A* **108**
747 **Suppl 3**, 15647-15654 (2011).
- 748 10. G. Morris, A. Nevet, D. Arkadir, E. Vaadia, H. Bergman, Midbrain dopamine
749 neurons encode decisions for future action. *Nat Neurosci* **9**, 1057-1063 (2006).
- 750 11. J. R. Hollerman, W. Schultz, Dopamine neurons report an error in the temporal
751 prediction of reward during learning. *Nat Neurosci* **1**, 304-309 (1998).
- 752 12. W. Schultz, P. Dayan, P. R. Montague, A neural substrate of prediction and
753 reward. *Science* **275**, 1593-1599 (1997).
- 754 13. A. A. Hamid, J. R. Pettibone, O. S. Mabrouk, V. L. Hetrick, R. Schmidt, C. M.
755 Vander Weele, R. T. Kennedy, B. J. Aragona, J. D. Berke, Mesolimbic dopamine
756 signals the value of work. *Nat Neurosci* **19**, 117-126 (2016).
- 757 14. M. W. Howe, P. L. Tierney, S. G. Sandberg, P. E. M. Phillips, A. M. Graybiel,
758 Prolonged dopamine signalling in striatum signals proximity and value of distant
759 rewards. *Nature* **500**, 575-+ (2013).
- 760 15. S. Soares, B. V. Atallah, J. J. Paton, Midbrain dopamine neurons control
761 judgment of time. *Science* **354**, 1273-1277 (2016).
- 762 16. M. Guitart-Masip, U. R. Beierholm, R. Dolan, E. Duzel, P. Dayan, Vigor in the face
763 of fluctuating rates of reward: an experimental examination. *J Cogn Neurosci* **23**,
764 3933-3938 (2011).
- 765 17. Y. Niv, Cost, benefit, tonic, phasic: what do response rates tell us about
766 dopamine and motivation? *Ann N Y Acad Sci* **1104**, 357-376 (2007).
- 767 18. Y. Niv, N. D. Daw, P. Dayan, Choice values. *Nat Neurosci* **9**, 987-988 (2006).
- 768 19. Y. Li, W. Zhong, D. Wang, Q. Feng, Z. Liu, J. Zhou, C. Jia, F. Hu, J. Zeng, Q. Guo,
769 L. Fu, M. Luo, Serotonin neurons in the dorsal raphe nucleus encode reward
770 signals. *Nat Commun* **7**, 10503 (2016).
- 771 20. G. A. Rummery, M. Niranjana, On-Line Q-Learning Using Connectionist
772 Systems. *Technical Report CUED/F-Infeng/TR 166*, (1994).
- 773 21. R. S. Sutton, A. G. Barto, *Reinforcement learning : an introduction* (Adaptive
774 computation and machine learning, MIT Press, Cambridge, Mass., 1998), pp.
775 xviii, 322 p.
- 776 22. M. Treisman, Temporal discrimination and the indifference interval. Implications
777 for a model of the "internal clock". *Psychol Monogr* **77**, 1-31 (1963).
- 778 23. P. R. Killeen, J. G. Fetterman, A behavioral theory of timing. *Psychol Rev* **95**, 274-
779 295 (1988).
- 780 24. W. H. Meck, Neuropharmacology of timing and time perception. *Brain Res Cogn*
781 *Brain Res* **3**, 227-242 (1996).

- 782 25. M. Jazayeri, M. N. Shadlen, A Neural Mechanism for Sensing and Reproducing a
783 Time Interval. *Curr Biol* **25**, 2599-2609 (2015).
- 784 26. Y. Niv, N. D. Daw, D. Joel, P. Dayan, Tonic dopamine: opportunity costs and the
785 control of response vigor. *Psychopharmacology (Berl)* **191**, 507-520 (2007).
- 786 27. R. S. Sutton, A. G. Barto, *Reinforcement learning : an introduction* (Adaptive
787 computation and machine learning series, The MIT Press, Cambridge,
788 Massachusetts, ed. Second edition., 2018), pp. xxii, 526 pages.
- 789 28. B. A. Bari, C. D. Grossman, E. E. Lubin, A. E. Rajagopalan, J. I. Cressy, J. Y. Cohen,
790 Stable Representations of Decision Variables for Flexible Behavior. *Neuron*,
791 (2019).
- 792 29. P. Simen, F. Balci, L. de Souza, J. D. Cohen, P. Holmes, A model of interval timing
793 by neural integration. *J Neurosci* **31**, 9238-9253 (2011).
- 794 30. D. Durstewitz, Self-Organizing Neural Integrator Predicts Interval Times through
795 Climbing Activity. **23**, 5342-5353 (2003).
- 796 31. F. Balci, P. Simen, A decision model of timing. *Current Opinion in Behavioral*
797 *Sciences* **8**, 94-101 (2016).
- 798 32. S. I. Gass, C. M. Harris, "Markov propertyMarkov property" in *Encyclopedia of*
799 *Operations Research and Management Science*, S. I. Gass, C. M. Harris, Eds.
800 (Springer US, New York, NY, 2001), pp. 490-490.
- 801 33. T. K. Koerner, Y. Zhang, Application of Linear Mixed-Effects Models in Human
802 Neuroscience Research: A Comparison with Pearson Correlation in Two Auditory
803 Electrophysiology Studies. *Brain Sci* **7**, (2017).
- 804 34. M. P. Boisgontier, B. Cheval, The anova to mixed model transition. *Neurosci*
805 *Biobehav Rev* **68**, 1004-1005 (2016).
- 806 35. S. N. Chettih, C. D. Harvey, Single-neuron perturbations reveal feature-specific
807 competition in V1. *Nature* **567**, 334-340 (2019).
- 808 36. K. Jimura, M. S. Chushak, T. S. Braver, Impulsivity and self-control during
809 intertemporal decision making linked to the neural dynamics of reward value
810 representation. *J Neurosci* **33**, 344-357 (2013).
- 811 37. B. Schmidt, C. B. Holroyd, S. Debener, J. Hewig, Why Is It So Hard to Wait? Brain
812 Responses to Delayed Gratification Predict Impulsivity and Self-Control.
813 *Psychophysiology* **52**, S42-S42 (2015).
- 814 38. C. D. Fiorillo, W. T. Newsome, W. Schultz, The temporal precision of reward
815 prediction in dopamine neurons. *Nat Neurosci* **11**, 966-973 (2008).
- 816 39. B. M. Babayan, N. Uchida, S. J. Gershman, Belief state representation in the
817 dopamine system. *Nat Commun* **9**, 1891 (2018).
- 818 40. J. Y. Cohen, S. Haesler, L. Vong, B. B. Lowell, N. Uchida, Neuron-type-specific
819 signals for reward and punishment in the ventral tegmental area. *Nature* **482**,
820 85-88 (2012).
- 821 41. J. G. Mikhael, H. R. Kim, N. Uchida, S. J. Gershman, Ramping and State
822 Uncertainty in the Dopamine Signal. 805366 (2019).
- 823 42. A. Guru, C. Seo, R. J. Post, D. S. Kullakanda, J. A. Schaffer, M. R. Warden,
824 Ramping activity in midbrain dopamine neurons signifies the use of a cognitive
825 map. 2020.2005.2021.108886 (2020).

- 826 43. S. J. Gershman, Dopamine ramps are a consequence of reward prediction errors.
827 *Neural Comput* **26**, 467-471 (2014).
- 828 44. H. R. Kim, A. N. Malik, J. G. Mikhael, P. Bech, I. Tsutsui-Kimura, F. Sun, Y. Zhang,
829 Y. Li, M. Watabe-Uchida, S. J. Gershman, N. Uchida, A Unified Framework for
830 Dopamine Signals across Timescales. *Cell* **183**, 1600-1616 e1625 (2020).
- 831 45. W. X. Pan, R. Schmidt, J. R. Wickens, B. I. Hyland, Dopamine cells respond to
832 predicted events during classical conditioning: evidence for eligibility traces in
833 the reward-learning network. *J Neurosci* **25**, 6235-6242 (2005).
- 834 46. H. C. Tsai, F. Zhang, A. Adamantidis, G. D. Stuber, A. Bonci, L. de Lecea, K.
835 Deisseroth, Phasic firing in dopaminergic neurons is sufficient for behavioral
836 conditioning. *Science* **324**, 1080-1084 (2009).
- 837 47. E. K. Diekhof, O. Gruber, When Desire Collides with Reason: Functional
838 Interactions between Anteroventral Prefrontal Cortex and Nucleus Accumbens
839 Underlie the Human Ability to Resist Impulsive Desires. *Journal of Neuroscience*
840 **30**, 1488-1493 (2010).
- 841 48. A. E. Goudriaan, M. Yucel, R. J. van Holst, Getting a grip on problem gambling:
842 what can neuroscience tell us? *Front Behav Neurosci* **8**, 141 (2014).
- 843 49. M. Pachitariu, C. Stringer, K. D. Harris, Robustness of Spike Deconvolution for
844 Neuronal Calcium Imaging. *J Neurosci* **38**, 7976-7985 (2018).
- 845 50. T. W. Chen, T. J. Wardill, Y. Sun, S. R. Pulver, S. L. Renninger, A. Baohan, E. R.
846 Schreiter, R. A. Kerr, M. B. Orger, V. Jayaraman, L. L. Looger, K. Svoboda, D. S.
847 Kim, Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature*
848 **499**, 295-300 (2013).

849

850 **Acknowledgments:** We thank Drs. M. Lou, W. Ge, Y. Rao, W Zhou, and B Min for
851 comments on the manuscript. **We thank Dr. L. Lu for the help of in vivo electrophysiology**
852 **recording. We thank the Molecular Imaging Core Facility (MICF) at the School of Life**
853 **Science and Technology, ShanghaiTech University for providing technical support.** This
854 work was supported by the National Natural Science Foundation of China (grant nos.
855 31922029, 31671086, 61890951, and 61890950 to J.H.), and a Shanghai Pujiang Talent
856 Award (grant no. 2018X0302-101-01 to W.S.). **Author contributions:** W.S. and J. H.
857 oversaw the whole project. W.S., J. H. & Z. G. designed the experiments. Z. G., C. L. & T.
858 L.. performed all animal experiments. Z. G. & H. W. analyzed the data. H. W. & S. F.

859 performed the computational modeling under the supervision of X. J. W.. M. C. performed
 860 the electrophysiological recordings. W.S., J. H., Z. G. & H. W. wrote the paper with the
 861 participation of all other authors. **Competing interests:** The authors declare no competing
 862 interests. **Data and Materials Availability:** All data needed to evaluate the conclusions in
 863 the paper are present in the paper and/or the Supplementary Materials.

864

Fig. 1. The behavioral performance of mice during learning of a delayed gratification task learning. (A) Left panel, schematic of the delayed gratification task. Right panel, the relationship between reward volumes and waiting durations in the two behavioral tasks. (B) A plot of Transistor-Transistor Logic (TTL) signals for the chronological sequence of behavioral events in the tasks. (C-E). The waiting duration and running duration both decreased during the training process in the pre-training phase ($p < 0.001$). (F) The distribution of waiting durations from the behavioral session on the last analyzed day (Day 15, light red), revealing significantly longer waiting durations compared to those from Day 1 (Day 1, gray, $n = 7$ mice). (G) The distribution

875 of running durations on Day1 and Day15 did not differ with training. **(H)** Continuous
 876 training significantly increased the average waiting duration ($p < 0.001$), whereas the
 877 training did not change the average running duration ($p = 0.97$).

878

Fig. 2. VTA DAergic activity ramps up consistently while the mice are waiting for the reward. (A) Schematic of stereotaxic virus injection procedures. (B) Confocal images illustrating GCaMP6m expression in VTA TH⁺ neurons. (C) A live-recording trace (magenta) of Ca²⁺ signal in VTA DAergic neurons and running speed (black) when the mouse was performing the delayed gratification task. Task events over time (top): the dashed vertical lines indicate waiting onset (blue), waiting termination (green), and reward onset (red). (D) The scaled Ca²⁺ signals (magenta) and GFP signals (green) curves of VTA DAergic neurons from the last day of pre-training and Days in the

delayed gratification task training (black, speed). (E) Waiting duration sorted ramping Ca^{2+} signal data from one mouse during the delayed gratification task training (150 trials). (F) Z-scored $\Delta F/F$ values at 0.5s before exit were significantly different when the reward volumes were different ($p < 0.01$). (G) Averaged Ca^{2+} signal curves with different outcomes from (E). (H) Shown separately for different trial stages (DAergic ramping periods), during the last week of training. There were no differences in the slope of the Ca^{2+} signals between trials with different reward outcomes.

Fig. 3. Single unit recordings reveal that high-frequency spiking of VTA DAergic neurons sustains waiting in the delayed gratification task. (A) Schematic of single

unit electrical recordings in the delayed gratification task. **(B)** An example image showing the placement of electrode tips in VTA with an electrolytic lesion. **(C)**. An example recording trace when a mouse performed a whole trial of the delayed gratification task. The waiting period was noted with a solid magenta line. **(D)** Raster plot of spikes trial by trial sorted in the waiting duration of the delayed gratification task. **(E)** A 3D PSTH plot for D. **(F)** Average response curve during the waiting period ($n = 17$ cells, each dashed line represents one cell). **(G)** Predicted calcium responses based on the convolution of the spike rate from (E) with a 0.1s time bin. **(H)** The predicted calcium response curve (black) is based on the spiking response curves in (F), compared to the measured calcium response curve (blue line) ($r=0.982$, Pearson correlation).

Fig. 4. Optogenetic manipulation of VTA DAergic activity altered the waiting durations. (A) Left: schematic of stereotaxic virus injection and surgical procedure. Right: the behavioral events and optogenetic manipulation protocol. (B) Images showing ChR2-mCherry expression (Top),. And whole-cell recordings showing action potentials evoked by 10-Hz laser flash sequences (50 flashes) in VTA TH⁺ neurons (Bottom). (C) Images showing eNpHR3.0-mCherry expression (Top), and whole-cell recordings showing that evoked action potentials were inhibited by continued 589nm laser in VTA TH⁺ neurons (Bottom). (D) Cumulative probabilities of waiting durations. The waiting durations of optogenetically activated trials were significantly increased (p=0.002, blue) relative to those of the previous day's trials (gray). Inset: a bar graph of the normalized waiting durations showing that the average waiting duration of unstimulated trials (p=0.96, magenta) did not differ from those of the previous day's

923 trials, or the next trials following photoactivation ($p=0.63$, green). (E) The same
924 experimental configuration as in (D), but VAT TH⁺ neurons were optogenetically
925 inhibited by a yellow laser. Optogenetic inhibition decreased the waiting duration
926 ($p=0.008$, yellow), whereas there were no differences between other trials ($p>0.5$).

927 **Fig. 5. Behavioral performances and ramping VTA DAergic activity are explained**
928 **by the RL model.** (A) Two reinforcement learning (RL) models: *Decision Ahead* and
929 *Continuous Deliberation*. $Q(a_n)$: value for action $a(n)$; $P(T_{an})$: probability of action $a(n)$.
930 $P_{wait}^{(n)}$: the probability of waiting; $Q_{wait}^{(n)}$: waiting action value; $P_{leave}^{(n)}$: probability of
931 leaving; $Q_{leave}^{(n)}$: leaving action value; $R^{(n)}$: received reward. (B-C) The distributions of
932 waiting durations from the early session and late session were simulated in the *Decision*
933 *Ahead* model (B) and *Continuous Deliberation* model (C). (D) The Kullback-Leibler
934 (KL) divergence between the distribution of early and late experimental data was
935

936 significantly larger than the KL divergences between the distributions of simulated
 937 behavioral performances and experimental data ($p = 0.005$), but no difference ($p > 0.99$)
 938 between the two RL models. (E) Plots of Z-scored $\Delta F/F$ values (DA_w , blue), scaled
 939 Q_{wait} (purple), and Q_{leave} (green) from the *Continuous Deliberation* model. The Q_{wait}
 940 and Q_{leave} both predicted the experimental calcium activity well. (F-G) Computational
 941 RL model (continuous deliberation)-simulated data, dependent on manipulating Q_{wait}
 942 in delayed gratification task. The simulation showed that manipulating Q_{wait} (i.e.,
 943 simulated dopamine activity) only influenced the waiting durations of Q_{wait}
 944 manipulated trials ((F-G), $p < 0.001$).

Fig. 6. VTA DAergic activity during waiting predicts the behavioral performance in the delayed gratification task. (A) Schematic of waiting probability (P_{wait}) in waiting periods after momentary DAergic periods. (B) Relationship between momentary VTA DAergic activity (Ca^{2+} signals) and its waiting probability. For each momentary DAergic period, its DAergic activity is only highly correlated ($p < 0.001$) with P_{wait} in the adjacent waiting period (the left bar of each cluster). (C) The distribution of Z-scored mean $\Delta F/F$ of momentary DAergic periods. Three colors are used to illustrate: high dopamine activity (High DA: red), low dopamine activity (Low DA: green), and other activity (gray). (D) The P_{wait} of High DA and Low DA activity trials for the adjacent period after the momentary DA periods. The difference between

956 the P_w of the High DA trials and Low DA trials increases with the increase of threshold
 957 (Th). (E-F) Cumulative probabilities of waiting durations for the high-DA-ramping
 958 trials (E, H-R, red), the lower-DA-ramping trials (E, L-R, green) and their "next-in-
 959 series" trials (F). Bar graph showing that the normalized waiting durations of the H-R
 960 trials are significantly longer than those of the L-R trials ((E), $p=0.024$), but did not
 961 differ significantly between their "next-in-series" trials ((F), $p=0.290$).

Fig. S1. The change of behavioral performance and reward rate during behavioral training. (A) The KL divergence of the distribution of waiting duration and running duration between the first day and the last day in pre-task and delayed gratification (DG). The KL divergence of waiting duration have no difference between pre-task and delayed gratification ($p = 0.53$, paired t-test), but KL divergence of running duration in pre-task was significantly ($p = 0.002$, paired t-test) different with the KL divergence of running duration from (B) The slopes of curves (Fig.1 E and Fig.1 H) that waiting duration and running duration changed during behavioral training. The waiting duration and running duration in pre-task were both changed more quickly ($p = 0.004$, waiting duration; $p < 0.001$, running duration, paired t-test) than those from delayed gratification. (C-D) The reward rate both increased in pre-task training (C) and delayed gratification training (D). All error bars represent the s.e.m..

Fig. S2. The histology and tip positions of optical fibre in GCaMP6m (green, n = 7), GFP (blue, n = 4), ChR2 (red, n = 12), ENpHR (yellow, n = 12), mCherry (magenta, n = 7) and electrolytic lesion (black, n = 5) were shown as coordinates in the mouse brain atlas. The circle markers showed the fiber placements from the mouse with optogenetic manipulation in the early training process. Scale bar 500um.

Fig. S3. The calcium and GFP signals of VTA DA neurons in behavioral tasks. (A) The GFP (green), calcium signals (magenta) of VTA DA neurons, and mouse running speed (black) in 7 days of pre-training. **(B)** The calcium signals of VTA DA neurons were changing with pre-training progress. The max calcium signals after the mouse received water rewards significantly decreased on days 6&7 ($p<0.01$, Friedman test). All error bars represent the s.e.m..

Fig. S4. The ramping dopamine activity became more stable with delayed gratification training. (A) The Frequency distribution of the slope of Z-scored $\Delta F/F$ during waiting (gray: day 1, red: day 15, $n = 7$). (B) The slope of $\Delta F/F$ had no difference between day 1 and day 15 ($p=0.63$, paired Student's t-test). (C) The standard deviation(σ) of the slope of $\Delta F/F$ in day 15 (0.32 ± 0.07) was significantly decreased than that in day 1 (0.17 ± 0.05 , $p=0.03$, paired Student's t-test). (D-E) The z-scored σ of $\Delta F/F$ and z-scored waiting durations were negatively correlated both in the experimental data (d, blue, $r = -0.80$, $p < 0.001$) and RL model (e, red, $r = -0.76$, $p < 0.001$). W.D. is for waiting duration. All error bars represent the s.e.m..

Fig. S5. Optical activation or inhibition of VTA DA activity didn't change the motivation of mouse in the delayed gratification tasks, the waiting duration of mouse in the early process of delayed gratification training, or the duration the mouse stayed in the given box in RPPT. (A) The running duration didn't change while the VTA DA neurons were optically activated ($F=0.20$, $p=0.82$, one-way ANOVA). (B) The result was the same as Figure A while optical inhibiting the VTA DA neurons ($F=0.12$, $p=0.88$, one-way ANOVA). (C) The waiting duration didn't change ($p = 0.37$, laser-on vs laser-off) while the VTA DA neurons were optically activated in the early training process ($F=1.31$, $p=0.30$, one-way ANOVA). (D) The result was the same ($p = 0.59$, laser-on vs laser-off) as Figure C while optical inhibiting the VTA DA neurons ($F=4.97$, $p=0.22$, one-way ANOVA). (E) The heatmap of mouse traces in RPPT in which the VTA DA neurons were optically activated pseudo-randomly in 20% probability while the mouse entered into the given box (red rectangle). (F) The Z-scored duration that the mouse stayed in the given box while the VTA DA neurons were activated (Laser-On In) had no significant difference ($F=0.75$, $p=0.44$, one-way ANOVA, $n=6$) with the un-inhibited durations (Laser-Off In) and durations in another box (Out). (G) The heatmap of mouse traces as shown in Figure C while inhibiting the VTA DA neurons in the given box (red rectangle). (H) Optical inhibiting the VTA DA neurons also didn't change the duration the mouse stayed in the given box ($F = 0.17$, $p = 0.73$, one-way ANOVA, $n=6$). All error bars represent the s.e.m..

Fig. S6. Optogenetic manipulation of DAT-Cre mouse expressed mCherry in the delayed gratification tasks and RPPT.

(A) Waiting durations in 473nm laser delivered trials (blue) are not different compared with those of all other trials ($p=0.17$, Friedman test, $n=7$). (B) Waiting durations of 589nm laser un-delivered trials (magenta) slightly increased compared with the waiting duration of the previous day ($p=0.02$, Friedman test, $n=7$). (C) Heat-map of mouse traces in RPPT in which the VTA of the mouse was delivered 473nm laser pseudo-randomly in 20% probability while the mouse entered into a randomly chosen box (red rectangle). (D) Mean durations that the mouse stayed in the chosen box while the laser delivered (Laser-On In), laser off (Laser-Off In) and the other box. There is no significant difference in waiting duration between Laser-On In, and Laser-Off In ($F=3.54$, $p=0.09$, one-way ANOVA, $n=7$). (E) Heatmap of mouse traces same as shown in (c) while the mouse was delivered 589nm laser in a randomly chosen box (red rectangle). (F) Mean durations that the mouse stayed in the chosen box while 589nm laser delivered (Laser-On In), laser off (Laser-Off In), and in the other box. 589nm laser delivering to mCherry mouse didn't alter waiting duration mouse stayed in any boxes under all experimental conditions ($F=2.64$, $p=0.14$, one-way ANOVA, $n=7$). All error bars represent the s.e.m..

Fig. S7. Reward prediction error is not significantly correlated with ramping dopaminergic activity. (A) The correlation coefficient of mean DA activity 1 s before waiting (DA_b), the value of action (Q_b), and RPE of action (RPE_b) before waiting in the Decision Ahead model, with waiting durations. The correlation of V_b ($r = 0.35 \pm 0.01$, $p = 0.001$, $n = 10$, Pearson correlation) and RPE_b ($r = 0.34 \pm 0.01$, $p = 0.001$, $n = 10$, Pearson correlation) with waiting duration were significantly ($p < 0.001$, Kruskal-Wallis test) higher than the CC of DA_b ($r = 0.01 \pm 0.02$, $p = 0.36$, $n = 7$) with waiting durations. (B) Plots of Z-scored $\Delta F/F$ values (DA_w , light blue) at 0.5s before the waiting ended and RPE of waiting (RPE_w). There were no significant correlation between DA_w and RPE_w ($r = 0.34$, $p = 0.41$, Pearson correlation). (C-D) Value of waiting (Q_w) (C) and RPE (D) changed in the Continuous Deliberation model. All error bars represent the s.e.m..

Fig. S8. Manipulation of the value of leaving and RPE in RL model. (A-B) Either increasing or decreasing the value of leaving (Q_{leave}) in the continuous deliberation model as with the manipulation of Q_{wait} induced the opposite results compared with the optogenetics manipulating DAergic activity (increasing Q_{leave} in **A**: $p < 0.001$, Friedman test, $n = 10$; decreasing Q_{leave} in **B**: $p < 0.001$, Friedman test, $n = 10$) and had no influences on other trials (A-B, $p > 0.999$, Friedman test, $n = 10$). (C-D) Either increasing (C) or decreasing (D) the RPE in the continuous deliberation model as with the experimental data, alters the waiting durations in the same direction in all trials, whether or not the RPEs-manipulation (increasing RPEs in c: $p < 0.001$, Friedman test, $n = 10$; decreasing RPE in (D): $p < 0.001$, Friedman test, $n = 10$). (E) The value of waiting is only positively correlated (0.23 ± 0.02 , $p = 0.03 \pm 0.01$, $n = 20$) with the adjacent behavior in the *Continuous Deliberation* RL model. (F) The p values of correlation coefficients in E. All error bars represent the s.e.m..