HHS Public Access

Author manuscript

Nat Neurosci. Author manuscript; available in PMC 2016 May 23.

Published in final edited form as:

Nat Neurosci. 2016 January ; 19(1): 117–126. doi:10.1038/nn.4173.

Mesolimbic Dopamine Signals the Value of Work

Arif A. Hamid[1,5,8], **Jeffrey R. Pettibone**[1,8], **Omar S. Mabrouk**[2,3], **Vaughn L. Hetrick**[1], **Robert Schmidt**[1,6], **Caitlin M. Vander Weele**[1,7], **Robert T. Kennedy**[2,3], **Brandon J. Aragona**[1,5], and **Joshua D. Berke**[1,4,5,9]

[1]) Department of Psychology, University of Michigan, Ann Arbor, USA

[2]) Department of Chemistry, University of Michigan, Ann Arbor, USA

[3]) Department of Pharmacology, University of Michigan, Ann Arbor, USA

[4]) Department of Biomedical Engineering, University of Michigan, Ann Arbor, USA

[5]) Department of the Neuroscience Graduate Program, University of Michigan, Ann Arbor, USA

[6]) BrainLinks-BrainTools Cluster of Excellence and Bernstein Center, University of Freiburg, Germany

[7]) Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

Abstract

Dopamine cell firing can encode errors in reward prediction, providing a learning signal to guide future behavior. Yet dopamine is also a key modulator of motivation, invigorating current behavior. Existing theories propose that fast ("phasic") dopamine fluctuations support learning, while much slower ("tonic") dopamine changes are involved in motivation. We examined dopamine release in the nucleus accumbens across multiple time scales, using complementary microdialysis and voltammetric methods during adaptive decision-making. We first show that minute-by-minute dopamine levels covary with reward rate and motivational vigor. We then show that second-by-second dopamine release encodes an estimate of temporally-discounted future reward (a value function). We demonstrate that changing dopamine immediately alters willingness to work, and reinforces preceding action choices by encoding temporal-difference reward prediction errors. Our results indicate that dopamine conveys a single, rapidly-evolving decision variable, the available reward for investment of effort, that is employed for both learning and motivational functions.

[9]) *Corresponding author (jdberke@umich.edu)*

[8]) Contributed equally to this work.

Author Contributions. A.A.H. performed and analyzed both FSCV and optogenetic experiments, and J.R.P. performed and analyzed the microdialysis experiments. O.S.M. assisted with microdialysis, C.V.W. assisted with FSCV, V.L.H. assisted with optogenetics, and R.S. assisted with reinforcement learning models. B.J.A. helped supervise the FSCV experiments and data analysis, and R.T.K. helped supervise microdialysis experiments. J.D.B. designed and supervised this study, performed the computational modeling, developed the theoretical interpretation, and wrote the manuscript.

The authors declare no competing financial interests.

Introduction

Altered dopamine signaling is critically involved in many human disorders from Parkinson's Disease to drug addiction. Yet the normal functions of dopamine have long been the subject of debate. There is extensive evidence that dopamine affects learning, especially the reinforcement of actions that produce desirable results[1]. Specifically, electrophysiological studies suggest that bursts and pauses of dopamine cell firing encode the reward prediction errors (RPEs) of reinforcement learning theory (RL)[2]. In this framework RPE signals are used to update estimated values of states and actions, and these updated values affect subsequent decisions when similar situations are re-encountered. Further support for a link between phasic dopamine and RPE comes from measurements of dopamine release using fast-scan cyclic voltammetry (FSCV)[3,4] and optogenetic manipulations[5,6].

There is also extensive evidence that dopamine modulates arousal and motivation[7,8]. Drugs that produce prolonged increases in dopamine release (e.g. amphetamines) can profoundly enhance psychomotor activation, while drugs or toxins that interfere with dopamine transmission have the opposite effect. Over slow timescales (tens of minutes) microdialysis studies have demonstrated that dopamine release ([DA]) is strongly correlated with behavioral activity, especially in the nucleus accumbens[9] (i.e. mesolimbic [DA]). It is widely thought that slow (tonic) [DA] changes are involved in motivation[10–12]. However, faster [DA] changes also appear to have a motivational function[13]. Subsecond increases in mesolimbic [DA] accompany motivated approach behaviors[14,15], and dopamine ramps lasting several seconds have been reported as rats approach anticipated rewards[16], without any obvious connection to RPE. Overall, the role of dopamine in motivation is still considered "mysterious"[12].

The purpose of this study was to better understand just how dopamine contributes to motivation, and to learning, simultaneously. We demonstrate that mesolimbic [DA] conveys a motivational signal in the form of *state values*, which are moment-by-moment estimates of available future reward. These values are used for making decisions about whether to *work*, i.e. to invest time and effort in activities that are not immediately rewarded, in order to obtain future rewards. When there is an unexpected change in value, the corresponding change in [DA] not only influences motivation to work, but also serves as an RPE learning signal, reinforcing specific choices. Rather than separate functions of "phasic" and "tonic" [DA], our data support a unified view in which the same dynamically-fluctuating [DA] signal influences both current and future motivated behavior.

Results

Motivation to work adapts to recent reward history

We made use of an adaptive decision-making task (Fig.1a and Methods) that is closely related to the reinforcement learning framework (a "two-armed bandit"). On each trial a randomly-chosen nose poke port lit up (Light-On) indicating that the rat might profitably approach and place its nose in that port (Center-In). The rat had to wait in this position for a variable delay (0.75-1.25s), until an auditory white noise burst (Go cue) prompted the rat to make a brief leftward or rightward movement to an adjacent side port. Unlike previous

behavioral tasks using the same apparatus, the Go cue did not specify which way to move; instead the rat had to learn through trial-and-error which option was currently more likely to be rewarded. Left and right choices had separate reward probabilities (each either 10%, 50%, or 90%), and these probabilities changed periodically without any explicit signal. On rewarded trials only, entry into the side port (Side-In) immediately triggered an audible click (the reward cue) as a food hopper delivered a sugar pellet to a separate food port at the opposite side of the chamber.

Trained rats readily adapted their behavior in at least two respects (Fig.1b,c). Firstly, actions followed by rewards were more likely to be subsequently selected (i.e. they were reinforced), producing left/right choice probabilities that scaled with actual reward probabilities[17] (Fig.1d).

Secondly, rats were more motivated to perform the task while it was producing a higher rate of reward[18,19]. This was apparent from "latency" (the time taken from Light-On until the Center-In nose poke), which scaled inversely with reward rate (Fig.1e-g). When reward rate was higher rats were more likely to be already waiting near the center ports at Light-On ("engaged" trials; Supplementary Fig.1), producing very short latencies. Higher reward rates also produced shorter latencies even when rats were not already engaged at Light-On (Supplementary Fig.1), due to an elevated moment-by-moment probability (hazard rate) of choosing to begin work (Fig.1h,i).

These latency observations are consistent with optimal foraging theories[20], which argue that reward rate is a key decision variable ("currency"). As animals perform actions and experience rewards they construct estimates of reward rate, and can use these estimates to help decide whether engaging in an activity is worthwhile. In a stable environment, the best estimate of reward rate is simply the total magnitude of past rewards received over a long time period, divided by the duration of that period. It has been proposed that such a "long-term-average reward rate" is encoded by slow (tonic) changes in [DA][10]. However, under shifting conditions such as our trial-and-error task, the reward rate at a given time is better estimated by more local measures. Reinforcement learning algorithms use past reward experiences to update estimates of future reward from each state: a set of these estimates is called a value function[21].

Minute-by-minute dopamine correlates with reward rate

To test whether changes in [DA] accompany reward rate during adaptive decision-making, we first employed microdialysis in the nucleus accumbens combined with liquid chromatography - mass spectrometry. This method allows us to simultaneously assay a wide range of neurochemicals, including all of the well-known low-molecular weight striatal neurotransmitters, neuromodulators and their metabolites (Fig.2a), each with 1-minute time resolution. We performed regression analyses to assess relationships between these neurochemicals and a range of behavioral factors: 1) reward rate; 2) the number of trials attempted (as an index of a more general form of activation/arousal); 3) the degree of exploitation versus exploration (an important decision parameter that has been suggested to involve [DA]; see Methods), and 4) the cumulative reward obtained (as an index of progressively increasing factors such as satiety).

We found a clear overall relationship between [DA] and ongoing reward rate ($R^2 = 0.15$, $p < 10^{-16}$). Among the 19 tested analytes, [DA] had by far the strongest relationship to reward rate (Fig.2b), and this relationship was significant in 6 of 7 individual sessions, from 6 different rats (Supplementary Fig.2). Modest significant relationships were also found for the dopamine metabolites DOPAC and 3-MT. We found a weaker relationship between [DA] and the number of trials attempted, but this was entirely accounted for by reward rate - i.e. if the regression model already included reward rate, adding number of attempts did not improve model fit. We did not find support for alternative proposals that tonic [DA] is related to exploration or exploitation, since higher [DA] was not associated with an altered probability of choosing the better left/right option (Fig.2b, Supplementary Fig.2). [DA] also showed no relationship to the cumulative total rewards earned (though there was a strong relationship between cumulative reward and the dopamine metabolite HVA, among other neurochemicals; Fig.2b). Additional information about the relationships between neurochemicals and behavioral variables, and one another, is given in Supplementary Fig.3.

We conclude that higher reward rate is associated specifically with higher average [DA], rather than other striatal neuromodulators, and with increased motivation to work. This finding supports the proposal that [DA] helps to mediate the effects of reward rate on motivation[10]. However, rather than signaling an especially “long-term” rate of reward, [DA] tracked minute-by-minute fluctuations in reward rate. We therefore needed to assess whether this result truly reflects an aspect of [DA] signaling that is inherently slow (tonic), or could instead be explained by rapidly-changing [DA] levels, that signal a rapidly-changing decision variable.

Dopamine signals time-discounted available future reward

To help distinguish these possibilities we used FSCV to assess task-related [DA] changes on fast time scales (from tenths of seconds to tens of seconds; Fig.3). Within each trial, [DA] rapidly increased as rats poked their nose in the start hole (Fig.3c,d, Center-In panel) and for all rats this increase was more closely related to this approach behavior than to the onset of the light cue (for data from each of the single sessions from all 6 rats, see Supplementary Fig.4). A second abrupt increase in [DA] occurred following presentation of the Go cue (Fig. 3c,d, middle panel). If received, the reward cue prompted a third abrupt increase (Fig.3c,d, Side-In panel). [DA] rose still further as rat approached the food port (Fig.3c,d, right), then declined once the reward was obtained. The same overall pattern of task-related [DA] change was observed in all rats, albeit with some variation (Supplementary Fig.4). [DA] increases did not simply accompany movements, since on the infrequent trials in which the rat approached the food port without hearing the reward cue we observed no corresponding increase in [DA] (Fig.3c,d, right).

The overall ramping up of [DA] as rats drew progressively closer to reward suggested some form of reward expectation[16]. Specifically, we hypothesized that [DA] continuously signals a *value function*: the temporally-discounted reward predicted from the current moment. To make this more clear, consider a hypothetical agent moving through a sequence of distinct, unrewarded states leading up to an expected reward (Fig.4a, top; perhaps a rat running at

constant speed along a familiar maze arm). Since the reward is more discounted when more distant, the value function will progressively rise until the reward is obtained.

This value function describes the time-varying level of motivation. If a reward is distant (so strongly discounted), animals are less likely to choose to work for it. Once engaged however, animals are increasingly motivated, and so less likely to quit, as they detect progress towards the reward (the value function produces a "goal-gradient", in the terminology of Hull[22]). If the reward is smaller or less reliable, the value function will be lower, indicating less incentive to begin work. Moving closer to our real situation, suppose that reward is equally likely to be obtained, or not, on any given trial, but a cue indicates this outcome halfway through the trial (Fig.4a, bottom). The increasing value function should initially reflect the overall 0.5 reward probability, but if the reward cue occurs estimated value should promptly jump to that of the (discounted) full reward.

Such unpredicted sudden transitions to states with a different value produce "temporal-difference" RPEs (Fig.4b). In particular, if the value function is low (e.g. the trajectory indicating 0.25 expectation of reward), the reward cue produces a large RPE, as value jumps up to the discounted value of the now-certain reward. If instead reward expectation was higher (e.g. 0.75 trajectory), the RPE produced by the reward cue is smaller. Since temporal difference RPEs *are* rapid shifts in value, under some conditions they can be challenging to dissociate from value itself. However, RPE and value signals are not identical. In particular, as reward gets closer, the state value progressively increases but RPE remains zero unless events occur with unpredicted value or timing.

Our task includes additional features, such as variable timing between events and many trials. We therefore considered what the "true" value function should look like - on average - based on actual times to future rewards (Fig.4c). At the beginning of a trial, reward is at least several seconds away, and may not occur at all until a later trial. During correct trial performance each subsequent, variably-timed event indicates to the rat that rewards are getting closer and more likely, and thus causes a jump in state value. For example, hearing the Go cue indicates both that reward is closer, and that the rat will not lose out by moving too soon (an impulsive procedural error). Hearing the reward cue indicates that reward is now certain, and only a couple of seconds away.

To assess how the intertwined decision variables - state value and RPE - are encoded by phasic [DA], we compared our FSCV measurements to the dynamically varying state value and RPE of a reinforcement learning model (see Methods). This simplified model consisted of a set of discrete states (Supplementary Fig.5), whose values were updated using temporal-difference RPEs. When given as input the actual sequence of behavioral events experienced by the rat, the model's value function consisted of a series of increases within each trial (Fig. 4d,e), resembling the observed time course of [DA] (Fig.3c).

Consistent with the idea that state value represents motivation to work, model state value early in each trial significantly correlated with behavioral latencies for all rats (across a wide range of model parameter settings; Supplementary Fig.5). We identified model parameters (learning rate = 0.4, discount factor = 0.95) that maximized this behavioral correlation

across all rats combined, and examined the corresponding within-trial correlation between [DA] and model variables. For all of the 6 FSCV rats we found a clear and highly significant positive correlation between phasic [DA] and state value V (Fig. 4f). [DA] and RPE were also positively correlated, as expected since V and RPE partially covary. However, in every case [DA] had a significantly stronger relationship to V than to RPE (Fig.4f, Supplementary Fig.5). We emphasize that this result was *not* dependent on specific model parameters; in fact, even if parameters were chosen to maximize the [DA] : RPE correlation, the [DA] : V correlation was stronger (Supplementary Fig.5).

Correlations were maximal when V was compared to the [DA] signal measured ~0.4-0.5s later (Fig.4g). This small delay is consistent with the known brief lag associated with the FSCV method using acute electrodes[23], and prior observations that peak [DA] response occurs ~0.5s after cue onset with acute FSCV recordings[3]. As an alternative method of incorporating temporal distortion that might be produced by FSCV and/or the finite speeds of DA release and update, we convolved model variables with a kernel consisting of an exponential rise and fall, and explored the effect of varying kernel time constants. Once again [DA] always correlated much better with V than with RPE, across a wide range of parameter values (Supplementary Fig.6). We conclude that state value provides a more accurate description of the time course of [DA] fluctuations than RPE alone, even though RPEs can be simultaneously signaled as changes in state value.

Abrupt dopamine changes encode reward prediction errors

FSCV electrode signals tend to drift over a timescale of minutes, so standard practice is to assess [DA] fluctuations relative to a pre-trial “baseline” of unknown concentration (as in Fig.3). Presented this way, reward cues appeared to evoke a higher absolute [DA] level when rewards were less common (Fig.5a,b), consistent with a conventional RPE-based account of phasic [DA]. However, our model implies a different interpretation of this data (Fig. 4b, 5c). Rather than a jump from a fixed to a variable [DA] level (that encodes RPE), we predicted that the reward cue actually causes a [DA] jump from a variable [DA] level (reflecting variable estimates of upcoming reward) to a fixed [DA] level (that encodes the time-discounted value of the now certain reward).

To test these competing accounts, we compared [DA] levels between consecutive pairs of rewarded trials with Side-In events < 30s apart (i.e. well within the accepted stable range of FSCV measurements[24]; for included pairs of trials the average time between Side-In events was 11.5s). If the [DA] level evoked by the reward cue reflects RPE, then this level should tend to decline as rats experience consecutive rewards (Fig.5d,e). However, if [DA] represents state value then “baseline” [DA] should asymptotically increase with repeated rewards while reward cue-evoked [DA] remains more stable (Fig.5f,g). The latter proved correct (Fig.5h,i). These results provide clear further evidence that [DA] reflects reward expectation (the value function), not just RPE.

Considering the microdialysis and FSCV results together, a parsimonious interpretation is that, across multiple measurement time scales, [DA] simply signals estimated availability of reward. The higher minute-by-minute [DA] levels observed with greater reward rate reflect both the higher values of states distal to rewards (including “baseline” periods between

active trial performance) and the greater proportion of time spent in high-value states proximal to rewards.

By conveying an estimate of available reward, mesolimbic [DA] could be used as a motivational signal, helping to decide whether it is worthwhile to engage in effortful activity. At the same time, abrupt *relative* changes in [DA] could be detected and used as an RPE signal for learning. But is the brain actually using [DA] to signal motivation, or learning, or both, within this task?

Dopamine both enhances motivation and reinforces choices

To address this question we turned to precisely-timed, bidirectional, optogenetic manipulations of dopamine. Following an approach validated in previous studies[6], we expressed channelrhodopsin-2 (ChR2) selectively in dopamine neurons by combining *TH-Cre*$^{+}$ rats with DIO-ChR2 virus injections and bilateral optic fibers in the ventral tegmental area (Supplementary Fig.7). We chose optical stimulation parameters (10ms pulses of blue light at 30Hz, 0.5s total duration; Fig.6a,b) that produced phasic [DA] increases of similar duration and magnitude to those naturally observed with unexpected reward delivery. We provided this stimulation at one of two distinct moments during task performance. We hypothesized that enhancing [DA] coincident with Light-On would increase the estimated motivational value of task performance; this would make the rat more likely to initiate an approach, leading to shorter latencies on the same trial. We further hypothesized that enhancing [DA] at the time of the major RPE (Side-In) would affect learning, as reflected in altered behavior on subsequent trials. In each session stimulation was given at only one of these two times, and on only 30% of trials (randomly selected) to allow within-session comparisons between stimulated and unstimulated trials.

Providing phasic [DA] at Side-In reinforced choice behavior: it increased the chance that the same left or right action was repeated on the next trial, whether or not the food reward was actually received (Fig.6c, left; n=6 rats; two-way ANOVA yielded significant main effects for LASER, $F(1,5)=224.0$, $p=2.4\times10^{-5}$ and for REWARD, $F(1,5)=41.0$, $p=0.0014$, without a significant LASER * REWARD interaction; see also Supplementary Fig.8c). No reinforcing effect was seen if the same optogenetic stimulation was given in littermate controls (Fig.6c middle; n=6 *TH-Cre*$^{-}$ rats; LASER main effect $F(1,5)=2.51$, $p=0.174$). For a further group of *TH-Cre*$^{+}$ animals (n=5) we instead used the inhibitory opsin Halorhodopsin (eNpHR3.0). Inhibition of dopamine cells at Side-In reduced the probability that the same left/right choice was repeated on the next trial (LASER main effect $F(1,4)=18.7$, $p=0.012$, without a significant LASER * REWARD interaction). A direct comparison between these three rat groups also demonstrated a group-specific effect of Side-In laser stimulation on choice reinforcement (two-way ANOVA, LASER * GROUP interaction $F(2,14)=69.4$, $p=5.4\times10^{-8}$). These observations support the hypothesis that abrupt [DA] fluctuations serve as an RPE learning signal, consistent with prior optogenetic manipulations[7]. However, extra [DA] at Side-In did not affect subsequent trial latency (Supplementary Fig.8a,b), indicating that our artificial [DA] manipulations reproduced some, but not all, types of behavioral change normally evoked by rewarded trials.

Optogenetic effects on reinforcement were temporally-specific: providing extra [DA] at Light-On (instead of Side-In) on trial *n* did not affect the probability that rats made the same choice on trial *n+1* (LASER main effect F(1,5) = 0.031, p = 0.867; see also Supplementary Fig.8c) nor did it affect the probability that choice on trial *n* was the same as trial *n-1* (LASER main effect F(1,5) = 0.233, p=0.649).

By contrast, extra [DA] at Light-On dramatically affected latency for that very same trial (Fig. 6d, S8). The effect on latencies depended on what the rat was doing at the time of Light-On (two-way ANOVA yielded a significant LASER * ENGAGED interaction, F(1,3) = 28.1, p=0.013). If the rat was already engaged in task performance, the very short latencies became slightly longer on average (median control latency = 0.45s, median stimulated latency=0.61s; simple main effect of LASER, F(1,3) = 10.4, p = 0.048). This effect apparently resulted from additional laser-evoked orienting movements on a subset of trials (see Supplementary Fig.9 for more detailed analysis). By contrast, for non-engaged trials extra [DA] significantly reduced latencies (Fig. 6d; median control latency=2.64s, median stimulated latency=2.16s; simple main effect of LASER, F(1,3) = 32.5, p=0.011). These optogenetic results are consistent with the idea that mesolimbic [DA] is less important for the initiation of simple, cue-evoked responses when a task is already underway[25], but is critical for motivating "flexible approach" behaviors[26].

The shorter latencies produced by extra [DA] was not the result of rats approaching the Center-In port at faster speeds, since the average approach trajectory was unaffected (Supplementary Fig.9). Instead, extra [DA] transiently increased the probability that rats initiated the approach behavior. As the approach itself lasted ~1-2s (Supplementary Fig.9), the result was an increased rate of Center-In events ~1-2s after the laser pulse train (Fig. 6e; see Supplementary Fig.10 for hazard rate time courses in individual rats). This effect of Light-On laser stimulation on hazard rates was dependent on rat group (two-way ANOVA, LASER * GROUP interaction F(2,14) = 26.28, p = 0.000018). Post-hoc pairwise comparison of simple laser effects showed a significant increase in hazard rate for *TH-Cre*$^{+}$ / ChR2 rats (F(1,14) = 62.06, p = 1.63×10^{-6}) and a significant reduction in hazard rate for *TH-Cre*$^{+}$ / eNpHR3.0 rats (F(1,14) = 6.31, p = 0.025), with no significant change in *TH-Cre-* / ChR2 rats (F(1,14) = 2.81, p = 0.116). Overall we conclude that, beyond just correlating with estimates of reward availability, mesolimbic [DA] helps translate those estimates into decisions to work for reward.

Discussion

A dopamine value signal used for both motivation and learning

Our results help confirm a range of disparate prior ideas, while placing them within a newly integrated theoretical context. First, phasic [DA] has been previously related to motivated approach[14,15] reward expectation[16] and effort-based decision-making[27], but our demonstration that [DA] specifically conveys the temporally-discounted value of future rewards grounds this motivational aspect of dopamine fluctuations within the quantitative frameworks of machine learning and optimal foraging theory. This idea is also consistent with findings using other techniques - for example, fMRI signals in ventral striatum (often

argued to reflect dopamine signaling) encode reward expectation in the form of temporally-discounted subjective value[28].

Second, using the complementary method of microdialysis to assess slower changes we partly confirmed proposals that reward rate is reflected specifically in increased [DA], which in turn enhances motivational vigor[10]. However, our critical, novel argument is that this motivational message of reward availability can dynamically change from moment to moment, rather than being an inherently slow (tonic) signal. Using optogenetics we confirmed that phasic changes in [DA] levels immediately affect willingness to engage in work, supporting the idea that sub-second [DA] fluctuations promptly influence motivational decision-making[13,29]. This dynamic [DA] motivation signal can help account for detailed patterns of time allocation[30]. For example, animals take time to reengage in task performance after getting a reward (the "post-reinforcement pause"), and this pause is longer when the next reward is smaller or more distant. This behavioral phenomenon has been a long-standing puzzle[31] but fits well with our argument that the time-discounted value of future rewards, conveyed by [DA], influences the moment-by-moment probability (hazard rate) of engaging in work.

Third, we confirmed the vital role of fast [DA] fluctuations, including transient dips, in signaling RPEs to affect learning[4-6]. However, a striking result from our analyses is that RPEs are conveyed by fast *relative* changes in the [DA] value signal, rather than deviations from a steady (tonic) baseline. This interpretation explains for the first time how [DA] can simultaneously provide both learning and motivational signals, an important gap in prior theorizing. Our results also highlight the importance of not assuming a consistent "baseline" [DA] level across trials in voltammetry studies.

One interesting implication is that among the many postsynaptic mechanisms that are affected by dopamine, some are concerned more with absolute levels and others with fast relative changes. This possibility needs to be investigated further, together with the natural working hypothesis that [DA] effects on neuronal excitability are closely involved in motivational functions[32] while [DA] effects on spike-timing-dependent-plasticity are responsible for reinforcement-driven learning[1]. It is also intriguing that a pulse of increased [DA] sufficient to immediately affect latency, or to alter left/right choice on subsequent trials, does not appear sufficient to alter latency on subsequent trials. This suggests that state values and left/right action values[17] may be updated via distinct mechanisms, or at different times within the trial.

Though dopamine is often labeled a "reward" transmitter, [DA] levels dropped during reward consumption, consistent with findings that dopamine is relatively less important for consuming - and apparently enjoying - rewards[7,33]. Mesolimbic [DA] has also been shown not to be required for performance of simple actions that are immediately followed by reward, such as pressing a lever once to obtain food[34]. Rather, loss of mesolimbic [DA] reduces motivation to work, in the sense of investing time and effort in activities that are not inherently rewarding or interesting, but may eventually lead to rewards[12]. Conversely, increasing [DA] with drugs such as amphetamines increases motivation to engage in

prolonged work, in both normal subjects and those with attention-deficit hyperactivity disorder[35,36].

Dopamine and decision dynamics

Our interpretation of mesolimbic [DA] as signaling the value of work is based upon rat decisions to perform our task rather than alternative "default" behaviors, such as grooming or local exploration. In this view mesolimbic [DA] helps determine *whether* to work, but not *which* activity is most worthwhile (i.e. it is "activational" more than "directional"[12]). It may be best considered to signal the overall motivational excitement associated with reward expectation, or equivalently, the perceived opportunity cost of sloth[10,30].

Based on prior results[27] we expect that [DA] signals reward availability without factoring in the costs of effortful work, but we did not parametrically vary such costs here. Other notable limitations of this study are that we only examined [DA] in the nucleus accumbens, and we did not selectively manipulate [DA] within various striatal subregions (and other dopamine targets). Our functional account of [DA] effects on behavioral performance is undoubtedly incomplete, and it will be important to explore alternative descriptions, especially more generalizable accounts that apply throughout the striatum. In particular, our observation that mesolimbic [DA] affects the hazard rate of decisions to work seems compatible with a broader influence of striatal [DA] over decision-making - for example, by setting "thresholds" for decision process completion[27,37,38]. Within sensorimotor striatum dopamine influences the vigor (and learning) of more elemental actions[38,39], and it has been shown that even saccade speed in humans is best predicted by a discounting model that optimizes the rate of reward[40]. In this way the activational / invigorating role of [DA] on both simple movements and motivation may reflect the same fundamental, computational-level mechanism applied to decision-making processes throughout striatum, affecting behaviors across a range of timescales.

Activational signals are useful, but not sufficient for adaptive decision-making in general. Choosing between alternative, simultaneously available courses of action requires net value representations for the specific competing options[27,41]. Although different subpopulations of dopamine neurons may carry somewhat distinct signals[42], the aggregate [DA] message received by target regions is unlikely to have sufficient spatial resolution to represent multiple competing values simultaneously[43] or sufficient temporal resolution to present them for rapid serial consideration[44]. By contrast, distinct ensembles of GABAergic neurons within the basal ganglia can dynamically encode the value of specific options, including through ramps-to-reward[45,46] that may reflect escalating bids for behavioral control. Such neurons are modulated by dopamine, and in turn provide key feedback inputs to dopamine cells that may contribute to the escalating [DA] patterns observed here.

Relationship between dopamine cell firing and release

Firing rates of presumed dopamine cells have been previously reported to escalate within trials under some conditions[47], but this has not been typically reported with reward anticipation. Several factors may contribute to this apparent discrepancy with our [DA] measures. The first is the nature of the behavioral task. Many important prior studies of

dopamine[2,3] (though not all[41]) used Pavlovian situations, in which outcomes are not determined by the animal's actions. When effortful work is not required to obtain rewards, the learned value of work may be low and corresponding decision variables may be less apparent.

Secondly, a moving rat receives constantly-changing sensory input, and may thus more easily define and discriminate a set of discrete states leading up to reward, compared to situations in which elapsed time is the sole cue of progress. When such a sequence of states can be more readily recognized, it may be easier to assign a corresponding set of escalating values as reward gets nearer in time. Determining subjects' internal state representations, and their development during training, is an important challenge for future work. It has been argued that ramps in [DA] might actually reflect RPE if space is non-linearly represented[48], or if learned values rapidly decay in time[49]. However, these suggestions do not address the critical relationship between [DA] and motivation that we aim to account for here.

Finally, release from dopamine terminals is strongly influenced by local microcircuit mechanisms within striatum[50] producing a dissociation between dopamine cell firing and [DA] in target regions. This dissociation is not complete - the ability of unexpected sensory events to drive a rapid, synchronized burst of dopamine cell firing is still likely to be of particular importance for abrupt RPE signaling at state transitions. More detailed models of dopamine release, incorporating dopamine cell firing, local terminal control, and uptake dynamics will certainly be needed to understand to how [DA] comes to convey a value signal.

Online Methods

Animals and Behavioral Task

All animal procedures were approved by the University of Michigan Committee on Use and Care of Animals. Male rats (300-500g, either wild-type Long-Evans or *TH-Cre*$^{+}$ with a Long-Evans background[51] were maintained on a reverse 12:12 light:dark cycle and tested during the dark phase. Rats were mildly food deprived, receiving 15g of standard laboratory rat chow daily in addition to food rewards earned during task performance. Training and testing was performed in computer-controlled Med Associates operant chambers (25cm × 30cm at widest point) each with a 5-hole nose-poke wall, as previously described[52–54]. Training to perform the trial-and-error task typically took ~2 months, and included several pretraining stages (2 days-2 weeks each, advancing when ~85% of trials were performed without procedural errors). First, any one of the 5 nosepoke holes was illuminated (at random), and poking this hole caused delivery of a 45 mg fruit punch flavored sucrose pellet into the Food Port (FR1 schedule). Activation of the food hopper to deliver the pellet caused a audible click (the reward cue). In the next stage, the hole illuminated at trial start was always one of the three more-central holes (randomly-selected), and rats learned to poke and maintain hold for a variable interval (750-1250ms) until Go cue onset (250ms duration white noise, together with dimming of the start port). Next, Go cue onset was also paired with illumination of both adjacent side ports. A leftward or rightward poke to one of these ports was required to receive a reward (each at 50% probability), and initiated the inter-trial interval (5-10s randomly selected from a uniform distribution). If the rat poked an unlit

center port (wrong start) or pulled out before the end of the hold period (false start), the house light turned on for the duration of an inter-trial-interval. During this stage (only), to discourage development of a side bias, a maximum of three consecutive pokes to the same side were rewarded. Finally, in the complete trial-and-error task left and right choices had independent reward probabilities, each maintained for blocks of 40-60 trials (randomly selected block length and sequence for each session). All combinations of 10%, 50% and 90% reward probability were used except 10:10 and 90:90. There was no event that indicated to the rat that a trial would be unrewarded other than the omission of the Reward cue and the absence of the pellet.

For a subset of ChR2 optogenetic sessions, overhead video was captured at 15 frames/s. The frames immediately preceding the Light-On events were extracted, and the positions of the nose tip and neck were marked (by scorers blind to whether that trial included laser stimulation). These positions were used to determine rat distance and orientation to the center port (the one that will be illuminated on that trial). Each trial was classified as "engaged" or "unengaged", using cutoff values of distance (10.6cm) and orientation (84°) that minimized the overlap between aggregate pink and green distributions. To assess how path length was affected by optogenetic stimulation, rat head positions were scored for each video frame between Light-On and Center-Nose-In. Engaged trials were further classified by whether the rat was immediately adjacent to one of the three possible center ports, and if that port was the one that became illuminated at Light-On or not (i.e. lucky, unlucky guesses).

Smoothing of latency (and other) time series for graphical display (Fig.1B,C) was performed using the MATLAB *filtfilt* function with a 7-trial window. To quantify the impact of prior trial rewards on current trial latency, we used a multiple regression model:

$$log_{10}\left(latency\right) = \beta_1 r_1 + \beta_2 r_{t-2} + \ldots \beta_{10} r_{t-10}$$

where r = 1 if the corresponding trial was rewarded. All latency analyses excluded trials of zero latency (i.e. those for which the rat's nose was already inside the randomly-chosen center port at Light-On). For analysis of prior trial outcomes on left/right choice behavior we used another multiple regression model, just as previously described[55].

Latency survivor curves were calculated simply as the proportion of trials for which the Center-In event had not yet occurred, at each 250ms interval after Light-On (i.e. a inverted cumulative latency distribution), smoothed with a 3-point moving average ($x_t' = 0.25x_{t-1} + 0.5x_t + 0.25x_{t+1}$). These survivor curves were then used to calculate hazard rates, as the fraction of the remaining latencies that occurred in each 250ms bin (i.e. the number of Center-In events that happened, divided by the number that could have happened).

We defined reward rate as the exponentially-weighted moving average of individual rewards (i.e., a leaky integrator[56–58]). For each session the integrator time constant was chosen to maximize the (negative) correlation between reward rate and behavioral latency. If instead we defined reward rate as simply the number of rewards during each minute (i.e. ignoring the contributions of trials in previous minutes to current reward rate), the relationship

between microdialysis-measured [DA] in that minute and reward rate was lower, though still significant (R^2=0.084, p=5.5×10^{-10}).

An important parameter in reinforcement learning is the degree to which agents choose the option that is currently estimated to be the best (exploitation) versus trying alternatives to assess whether they are actually better (exploration), and dopamine has been proposed to mediate this trade-off[59,60]. To assess this we examined left/right choices in the second half of each block, by which time choices have typically stabilized (Fig. 1D; this behavioral pattern was also seen for the microdialysis sessions). We defined an Exploitation Index as the proportion of trials for which rats choose the better option in these second block halves (so values close to 1 would be fully exploitative, and values close to 0.5 would be random/exploratory). As an alternative metric of exploration/exploitation, we examined the number of times that the rat switched between left and right choices in each minute; this metric also showed no significant relationship to any neurochemical assayed in our microdialysis experiments.

Microdialysis

After 3-6 months of behavioral training rats were implanted with guide cannulae bilaterally above the nucleus accumbens core (NAcc; +1.3-1.9mm AP, 1.5mm ML from bregma) and allowed to recover for at least one week before retraining. On test days (3-5 weeks after cannula implantation) a single custom made microdialysis probe (300μm diameter) with polyacrylonitrile membrane (Hospal, Bologna, Italy; 20kD molecular weight cutoff) was inserted into NAcc, extending 1mm below the guide cannula. Artificial CSF (composition in mM: $CaCl_2$ 1.2; KCl 2.7, NaCl 148, $MgCl_2$ 0.85, 0.25 ascorbate) was perfused continuously at 2μl/min. Rats were placed in the operant chamber with the house light on for an initial 90min period of probe equilibration, after which samples were collected once every minute. Following five baseline samples the house light was extinguished to indicate task availability.

For chemical analyses, we employed a modified version of our benzoyl chloride derivatization and HPLC-MS analysis method[61]. Immediately after each 2μl sample collection, we added 1.5μl of buffer (sodium carbonate monohydrate 100 mM), 1.5μl of 2% benzoyl chloride in acetonitrile, and 1.5μl of a 13C-labeled internal standard mixture (total mixture volume 6.5μl). The mixture was vortexed for 2s between each reagent addition. Since ACh is a quaternary amine and thus not derivatized by benzoyl chloride, it was directly detected in its native form (transition 146->87). Deuterated ACh (d4-ACh) was also added to the internal standard mixture for improved ACh quantification[62]. Five μl of the sample mixture was automatically injected by a Thermo Accela HPLC system (Thermo Fisher Scientific, Waltham, MA) onto a reverse-phase Kinetex biphenyl HPLC column (2.1mm × 100 mm; 1.7 particle size; Phenomenex, Torrance CA). The HPLC system was interfaced to a HESI II ESI probe and Thermo TSQ Quantum Ultra (Thermo Scientific) triple quadrupole mass spectrometer operating in positive mode. Sample run times for all analytes were 3 min. To quantify neurochemicals in dialysate samples, we constructed 6-point external calibration curves encompassing known physiological concentrations. Thermo Xcalibur 2.1 software (Thermo Fisher Scientific) automatically detected chromatographic

peaks and quantified concentrations. To reduce noise each resulting minute-by-minute time series was smoothed with a 3-point moving average (as above), then converted to Z-scores to facilitate comparison between subjects.

Regression analysis of microdialysis data was performed stepwise. We first constructed models with only one behavioral variable as predictor and one outcome (analyte). If two behavioral variables showed a significant relationship to a given analyte, we constructed a model with both behavioral variables and an interaction term, and examined the capacity of each variable to explain analyte variance without substantial multicollinearity.

To determine cross-correlogram statistical thresholds we first shuffled the time series for all sessions 200,000 times, and calculated the average Pearson correlation coefficients (i.e. the zero-lag cross-correlation) for each shuffled pair of time series. Thresholds were based on the tails of the resulting distribution: i.e. for uncorrected two-tailed alpha=0.05 we would find the levels for which 2.5% of the shuffled values lay outside these thresholds. As we wished to correct for multiple comparisons we divided alpha by the number of tests (276; number of cross-correlograms = 23 timeseries * 22 timeseries divided by two, since the crosscorrelograms are just mirror-reversed when the order is changed, plus 23 autocorrelograms).

Voltammetry

FSCV electrode construction, data acquisition and analysis were performed as described[63]. Rats were implanted with a guide cannula above the right NAcc (+1.3-2.0 mm AP, 1.5 mm ML from bregma), a Ag/AgCl reference electrode (in the contralateral hemisphere) and a bipolar stimulation electrode aimed at the VTA (−5.2 mm AP, 0.8 mm ML, 7.5 mm DV). Carbon fiber electrodes were lowered acutely into the NAcc. Dopaminergic current was quantified offline using principal component regression (PCR)[24] using training data for dopamine and pH from electrical stimulations. Recording time points that exceeded the PCR residual analysis threshold (Qα) were omitted from further processing or analysis. Current to [DA] conversion was based on *in vitro* calibrations of electrodes constructed in the same manner with the same exposed fiber length. On many days data was not recorded due to electrode breakage or obvious movement-related electrical noise. FSCV recordings were made from 41 sessions (14 rats total). We excluded those sessions for which the rat failed to complete at least three blocks of trials, and those in which electrical artifacts caused >10% of trials to violate the assumptions of PCR residual analysis. The remaining 10 sessions came from 6 different rats. To avoid aggregate results being overly skewed by a single animal, we only included one session from each of the six rats (the session with the largest reward-evoked [DA] increase). Upon completion of FSCV testing, animals were deeply anesthetized and electrolytic lesions were created (40 μA for 15 seconds at the same depth as recording site) using stainless steel electrodes with 500 μm of exposed tip (AM Systems, USA). Lesion locations were later reconstructed in Nissl stained sections.

For between-session comparisons we normalized [DA] to the average [DA] difference between the pre-trial baseline and Food-Port-In aligned peak levels. To visualize the reward-history-dependence of [DA] change between consecutive trials (Fig.5H), we first extracted time series of normalized [DA] from consecutive pairs of rewarded trials (Side-In event to

subsequent Side-In event separated by less than 30s). For each session we divided these traces into “low-reward-rate” and “high-reward-rate” groups, using the (# of rewarded trials in the last 10) that best approximated a median-split (i.e. so low- and high- reward-rate groups had similar trial numbers). We then averaged all low-reward-rate traces, and separately all high-reward-rate traces.

Reinforcement Learning model

To estimate the time-varying state value and RPE within each trial we used a Semi-Markov Decision Process[64] with temporal difference learning, implemented in MATLAB. The model consisted of a set of states, with rat behavioral events determining the times of transitions between states (Supplementary Fig.5). Each state was associated with a stored (‘cached’) value of entering that state, V(s). At each state transition a reward prediction error δ was calculated using:

$$\delta_t = r_t + V_t(s_t) - \gamma^{-n} V_t(s_{t-n})$$

where n is the number of timesteps since the last state transition (a timestep of 50ms was used throughout), r is defined as one at reward receipt and zero otherwise, and γ specifies the rate at which future rewards are discounted at each timestep ($\gamma < 1$). The V terms in the equation compare the cached value of the new state to the value predicted, given the prior state value and the elapsed time since the last transition (as illustrated in Fig.4C). Each state also had e(s), an eligibility trace that decayed with the same time parameter γ (following the terminology of ref. 21, this is a TD(1) model with replacing traces). RPEs updated the values of the states encountered up to that point, using

$$V'(s) = V(s) + \alpha.\delta. \quad e_t(s)$$

where γ is the learning rate. V and γ were defined only at state transitions, and V was constrained to be non-negative. The model was “episodic” as all eligibilities were reset to zero at trial outcome (reward receipt, or omission). V is therefore a estimate of the time-discounted value of the next reward, rather than total aggregate future reward; with exponential discounting and best-fit parameters subsequent time-discounted rewards are negligible (but this would not necessarily be the case if hyperbolic discounting was used).

We also examined the effect of calculating prediction errors slightly differently:

$$\delta_t = r_t + \gamma^n (s_t) - V_t(s_{t-n})$$

This version compares a discounted version of the new state value to the previous state value. As expected, the results were the same. Specifically, overall [DA] correlation to V remained ~0.4, overall δ correlation was ~0.2, and each individual session [DA] was significantly better correlated to V than to δ, across the full parameter space.

We present results using γ in the 0.9 to 1 range, because 0.9 is already a very fast exponential discount rate when using 50ms timesteps. However we also tested smaller γ

(0.05-0.9) and confirmed that the [DA]: δ correlation only diminished in this lower range (not shown).

To compare within-trial [DA] changes to model variables, we identified all epochs of time (3s before to 3s after Center-In) with at least 6 state transitions (this encompasses both rewarded and unrewarded trials). Since the model can change state value instantaneously, but our FSCV signal cannot[65], we included an offset lag (so we actually compared V and δ to [DA] a few measurements later). The size of the lag affected the magnitude of the observed correlations (Fig.4f) but not the basic result. Results were also unchanged if (instead of a lag) we convolved model variables with a kernel consisting of an exponential rise and fall (Supplementary Fig.6), demonstrating that our results are not a simple artifact of time delays associated with the FSCV method or sluggish reuptake. Finally, we also tried using the SMDP model with hyperbolic (instead of exponential) discounting[66–69], and again found a consistently stronger correlation between [DA] and V than between [DA] and δ (not shown).

Code availability

custom MATLAB code for the SMDP model is available upon request.

Optogenetics—We used three groups of rats to assess the behavioral effects of VTA DA cell manipulations (first *TH-Cre*$^{+}$ with AAV-EF1α-DIO-ChR2-EYFP virus, then littermate *TH-Cre*$^{-}$ with the same virus, then *TH-Cre*$^{+}$ with AAV-EF1α-DIO-eNpHR3.0-EYFP). All virus was produced at the University of North Carolina vector core. In each case rats received bilateral viral injections (0.5 or 1μl per hemisphere at 50 nL/min) into the VTA (same coordinates as above). After 3 weeks, we placed bilateral optic fibers (200 μm diameter) under ketamine/xylazine anesthesia with FSCV guidance, at an angle of 6° from the sagittal plane, stopping at a location that yielded the most laser-evoked [DA] release in NAc. Once cemented in place, we used FSCV to test multiple sets of stimulation parameters from a 445nm blue laser diode (Casio) with Arroyo Instruments driver under LabView control. The parameters chosen for behavioral experiments (0.5s train of 10ms pulses at 30Hz, 20mW power at tip) typically produced [DA] increases in *TH-Cre*$^{+}$ / ChR2 rats comparable to those seen with unexpected reward delivery. All rats were allowed to recover from surgery and retrained to pre-surgery performance. Combined behavioral / optogenetic experiments began 5 weeks after virus injection. On alternate days, sessions either included bilateral laser stimulation (on a randomly selected 30% of trials, regardless of block or outcome), or not. In this manner, each rat received 3 sessions of Light-On stimulations and 3 sessions of Side-In stimulation, interleaved with control (no laser) sessions, over a two-week period. Halorhodopsin rats were tested with 1s of constant 20mW illumination from a 589nm (yellow/orange) laser (OEM Systems), starting either at Light-On or Side-In as above. One *TH-Cre*$^{+}$ /ChR2 rat was excluded from analyses due to misplaced virus (no viral expression directly below the optic fiber tips).

For statistical analysis of optogenetic effects on behavior we used repeated measure ANOVA models, in SPSS. For each rat we first averaged data across the 3 sessions with the same optogenetic conditions. Then, to assess reinforcing effects we examined the two factors of LASER (off vs on) and REWARD (rewarded vs omission), with the dependent measure the

probability that the same action was repeated on the next trial. For assessing effects on median latency we examined the two factors of LASER (off vs on) and ENGAGED (yes vs no). For assessing group-dependent effects on hazard rate we examined the factors of LASER (off vs on) and GROUP (*TH-Cre*$^{+}$/ChR2; *TH-Cre*$^{-}$/ChR2; *TH-Cre*$^{+}$/eNpHR3.0), with the dependent measure the average hazard rate during the epoch 1-2.5s after Light-On. This epoch was chosen since it is 1-2s after the laser stimulation period (0-0.5s) and approach behaviors have a consistent duration of ~1-2s (Supplementary Fig.9). Post-hoc tests were Bonferroni-corrected for multiple comparisons.

A supplementary methods checklist is available.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Kent Berridge, Terry Robinson, Roy Wise, Peter Redgrave, Peter Dayan, Daniel Weissman, Anatol Kreitzer, Niki Sanderson, Daniel Leventhal, Satinder Singh, Mark Walton, Saleem Nicola and members of the Berke Lab for critical reading of various manuscript drafts, Nico Mallet for initial assistance with viral injections, and Kirsten Porter-Stransky for initial assistance with microdialysis procedures. *TH-Cre*$^{+}$ rats were developed by Karl Deisseroth and Ilana Witten and made available for distribution through RRRC (www.rrrc.us). This work was supported by the National Institute on Drug Abuse (DA032259, training grant DA007281), the National Institute of Mental Health (MH093888, MH101697), the National Institute on Neurological Disorders and Stroke (NS078435, training grant NS076401), and the National Institute of Biomedical Imaging and Bioengineering (EB003320). R.S. was supported by the BrainLinks-BrainTools Cluster of Excellence funded by the German Research Foundation (DFG grant number EXC1086).

References

1. Reynolds JN, Hyland BI, Wickens JR. A cellular mechanism of reward-related learning. Nature. 2001; 413:67–70. [PubMed: 11544526]
2. Schultz W, Dayan P, Montague PR. A neural substrate of prediction and reward. Science. 1997; 275:1593–9. [PubMed: 9054347]
3. Day JJ, Roitman MF, Wightman RM, Carelli RM. Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens. Nat Neurosci. 2007; 10:1020–8. doi:nn1923 [pii] 10.1038/nn1923. [PubMed: 17603481]
4. Hart AS, Rutledge RB, Glimcher PW, Phillips PE. Phasic dopamine release in the rat nucleus accumbens symmetrically encodes a reward prediction error term. J Neurosci. 2014; 34:698–704. doi:10.1523/JNEUROSCI.2489-13.2014. [PubMed: 24431428]
5. Kim KM, Baratta MV, Yang A, Lee D, Boyden ES, Fiorillo CD. Optogenetic mimicry of the transient activation of dopamine neurons by natural reward is sufficient for operant reinforcement. PLoS One. 2012; 7:e33612. doi:10.1371/journal.pone.0033612. [PubMed: 22506004]
6. Steinberg EE, Keiflin R, Boivin JR, Witten IB, Deisseroth K, Janak PH. A causal link between prediction errors, dopamine neurons and learning. Nat Neurosci. 2013 doi:10.1038/nn.3413.
7. Berridge KC. The debate over dopamine's role in reward: the case for incentive salience. Psychopharmacology (Berl). 2007; 191:391–431. doi:10.1007/s00213-006-0578-x. [PubMed: 17072591]
8. Beierholm U, Guitart-Masip M, Economides M, Chowdhury R, Düzel E, Dolan R, Dayan P. Dopamine Modulates Reward-Related Vigor. Neuropsychopharmacology. 2013 doi:10.1038/npp.2013.48.
9. Freed CR, Yamamoto BK. Regional brain dopamine metabolism: a marker for the speed, direction, and posture of moving animals. Science. 1985; 229:62–65. [PubMed: 4012312]

10. Niv Y, Daw N, Dayan P. How fast to work: Response vigor, motivation and tonic dopamine. Advances in neural information processing systems. 2006; 18:1019.
11. Cagniard B, Balsam PD, Brunner D, Zhuang X. Mice with chronically elevated dopamine exhibit enhanced motivation, but not learning, for a food reward. Neuropsychopharmacology. 2006; 31:1362–70. doi:10.1038/sj.npp.1300966. [PubMed: 16319913]
12. Salamone JD, Correa M. The mysterious motivational functions of mesolimbic dopamine. Neuron. 2012; 76:470–85. doi:10.1016/j.neuron.2012.10.021. [PubMed: 23141060]
13. Satoh T, Nakai S, Sato T, Kimura M. Correlated coding of motivation and outcome of decision by dopamine neurons. J Neurosci. 2003; 23:9913–23. [PubMed: 14586021]
14. Phillips PE, Stuber GD, Heien ML, Wightman RM, Carelli RM. Subsecond dopamine release promotes cocaine seeking. Nature. 2003; 422:614–8. doi:10.1038/nature01476. [PubMed: 12687000]
15. Roitman MF, Stuber GD, Phillips PE, Wightman RM, Carelli RM. Dopamine operates as a subsecond modulator of food seeking. J Neurosci. 2004; 24:1265–71. doi:10.1523/JNEUROSCI.3823-03.2004. [PubMed: 14960596]
16. Howe MW, Tierney PL, Sandberg SG, Phillips PE, Graybiel AM. Prolonged dopamine signalling in striatum signals proximity and value of distant rewards. Nature. 2013; 500:575–9. doi:10.1038/nature12475. [PubMed: 23913271]
17. Samejima K, Ueda Y, Doya K, Kimura M. Representation of action-specific reward values in the striatum. Science. 2005; 310:1337–40. [PubMed: 16311337]
18. Guitart-Masip M, Beierholm UR, Dolan R, Duzel E, Dayan P. Vigor in the face of fluctuating rates of reward: an experimental examination. J Cogn Neurosci. 2011; 23:3933–8. doi:10.1162/jocn_a_00090. [PubMed: 21736459]
19. Wang AY, Miura K, Uchida N. The dorsomedial striatum encodes net expected return, critical for energizing performance vigor. Nat Neurosci. 2013; 16:639–47. doi:10.1038/nn.3377. [PubMed: 23584742]
20. Stephens, DW. Foraging theory. Foraging theory. Princeton University Press; Princeton, N.J.: 1986.
21. Sutton, RS.; Barto, AG. Reinforcement learning: an introduction. Reinforcement learning: an introduction. MIT Press; Cambridge, Massachusetts: 1998.
22. Hull CL. The goal-gradient hypothesis and maze learning. Psychological Review. 1932; 39:25.
23. Venton BJ, Troyer KP, Wightman RM. Response Times of Carbon Fiber Microelectrodes to Dynamic Changes in Catecholamine Concentration. Anal. Chem. 2002; 74:539–546. doi:10.1021/ac010819a. [PubMed: 11838672]
24. Heien ML, Khan AS, Ariansen JL, Cheer JF, Phillips PE, Wassum KM, Wightman RM. Real-time measurement of dopamine fluctuations after cocaine in the brain of behaving rats. Proceedings of the National Academy of Sciences of the United States of America. 2005; 102:10023–10028. [PubMed: 16006505]
25. Nicola SM. The flexible approach hypothesis: unification of effort and cue-responding hypotheses for the role of nucleus accumbens dopamine in the activation of reward-seeking behavior. J Neurosci. 2010; 30:16585–600. doi:10.1523/JNEUROSCI.3958-10.2010. [PubMed: 21147998]
26. Ikemoto S, Panksepp J. The role of nucleus accumbens dopamine in motivated behavior: a unifying interpretation with special reference to reward-seeking. Brain Res Brain Res Rev. 1999; 31:6–41. doi:S0165017399000235 [pii]. [PubMed: 10611493]
27. Gan JO, Walton ME, Phillips PE. Dissociable cost and benefit encoding of future rewards by mesolimbic dopamine. Nat Neurosci. 2010; 13:25–7. doi:10.1038/nn.2460. [PubMed: 19904261]
28. Kable JW, Glimcher PW. The neural correlates of subjective value during intertemporal choice. Nat Neurosci. 2007; 10:1625–33. doi:10.1038/nn2007. [PubMed: 17982449]
29. Adamantidis AR, Tsai HC, Boutrel B, Zhang F, Stuber GD, Budygin EA, Touriño C, Bonci A, Deisseroth K, de Lecea L. Optogenetic interrogation of dopaminergic modulation of the multiple phases of reward-seeking behavior. J Neurosci. 2011; 31:10829–35. doi:10.1523/JNEUROSCI.2246-11.2011. [PubMed: 21795535]
30. Niyogi RK, Breton YA, Solomon RB, Conover K, Shizgal P, Dayan P. Optimal indolence: a normative microscopic approach to work and leisure. J R Soc Interface. 2014; 11:20130969. doi:10.1098/rsif.2013.0969. [PubMed: 24284898]

31. Schlinger HD, Derenne A, Baron A. What 50 years of research tell us about pausing under ratio schedules of reinforcement. The Behavior Analyst. 2008; 31:39. [PubMed: 22478501]
32. du Hoffmann J, Nicola SM. Dopamine invigorates reward seeking by promoting cue-evoked excitation in the nucleus accumbens. J Neurosci. 2014; 34:14349–64. doi:10.1523/JNEUROSCI.3492-14.2014. [PubMed: 25339748]
33. Cannon CM, Palmiter RD. Reward without dopamine. The Journal of neuroscience. 2003; 23:10827–10831. [PubMed: 14645475]
34. Ishiwari K, Weber SM, Mingote S, Correa M, Salamone JD. Accumbens dopamine and the regulation of effort in food-seeking behavior: modulation of work output by different ratio or force requirements. Behav Brain Res. 2004; 151:83–91. doi:10.1016/j.bbr.2003.08.007. [PubMed: 15084424]
35. Rapoport JL, Buchsbaum MS, Weingartner H, Zahn TP, Ludlow C, Mikkelsen EJ. Dextroamphetamine. Its cognitive and behavioral effects in normal and hyperactive boys and normal men. Arch Gen Psychiatry. 1980; 37:933–43. [PubMed: 7406657]
36. Wardle MC, Treadway MT, Mayo LM, Zald DH, de Wit H. Amping. up effort: effects of d-amphetamine on human effort-based decision-making. J Neurosci. 2011; 31:16597–602. doi:10.1523/JNEUROSCI.4387-11.2011. [PubMed: 22090487]
37. Nagano-Saito A, Cisek P, Perna AS, Shirdel FZ, Benkelfat C, Leyton M, Dagher A. From anticipation to action, the role of dopamine in perceptual decision making: an fMRI-tyrosine depletion study. J Neurophysiol. 2012; 108:501–12. doi:10.1152/jn.00592.2011. [PubMed: 22552189]
38. Leventhal DK, Stoetzner C, Abraham R, Pettibone J, DeMarco K, Berke JD. Dissociable effects of dopamine on learning and performance within sensorimotor striatum. Basal Ganglia. 2014; 4:43–54. doi:10.1016/j.baga.2013.11.001. [PubMed: 24949283]
39. Turner RS, Desmurget M. Basal ganglia contributions to motor control: a vigorous tutor. Curr Opin Neurobiol. 2010; 20:704–16. doi:10.1016/j.conb.2010.08.022. [PubMed: 20850966]
40. Haith AM, Reppert TR, Shadmehr R. Evidence for hyperbolic temporal discounting of reward in control of movements. J Neurosci. 2012; 32:11727–36. doi:10.1523/JNEUROSCI.0424-12.2012. [PubMed: 22915115]
41. Morris G, Nevet A, Arkadir D, Vaadia E, Bergman H. Midbrain dopamine neurons encode decisions for future action. Nat Neurosci. 2006; 9:1057–63. [PubMed: 16862149]
42. Matsumoto M, Hikosaka O. Two types of dopamine neuron distinctly convey positive and negative motivational signals. Nature. 2009; 459:837–41. doi:nature08028 [pii] 10.1038/nature08028. [PubMed: 19448610]
43. Dreyer JK, Herrik KF, Berg RW, Hounsgaard JD. Influence of phasic and tonic dopamine release on receptor activation. J Neurosci. 2010; 30:14273–83. doi:10.1523/JNEUROSCI.1894-10.2010. [PubMed: 20962248]
44. McClure SM, Daw ND, Montague PR. A computational substrate for incentive salience. Trends Neurosci. 2003; 26:423–8. [PubMed: 12900173]
45. Tachibana Y, Hikosaka O. The primate ventral pallidum encodes expected reward value and regulates motor action. Neuron. 2012; 76:826–37. doi:10.1016/j.neuron.2012.09.030. [PubMed: 23177966]
46. van der Meer MA, Redish AD. Ventral striatum: a critical look at models of learning and evaluation. Curr Opin Neurobiol. 2011; 21:387–92. doi:10.1016/j.conb.2011.02.011. [PubMed: 21420853]
47. Fiorillo CD, Tobler PN, Schultz W. Discrete coding of reward probability and uncertainty by dopamine neurons. Science. 2003; 299:1898–902. doi:10.1126/science.1077349 299/5614/1898 [pii]. [PubMed: 12649484]
48. Gershman SJ. Dopamine ramps are a consequence of reward prediction errors. Neural Comput. 2014; 26:467–71. doi:10.1162/NECO_a_00559. [PubMed: 24320851]
49. Morita K, Kato A. Striatal dopamine ramping may indicate flexible reinforcement learning with forgetting in the cortico-basal ganglia circuits. Front Neural Circuits. 2014; 8:36. doi:10.3389/fncir.2014.00036. [PubMed: 24782717]

50. Threlfell S, Lalic T, Platt NJ, Jennings KA, Deisseroth K, Cragg SJ. Striatal dopamine release is triggered by synchronized activity in cholinergic interneurons. Neuron. 2012; 75:58–64. doi: 10.1016/j.neuron.2012.04.038. [PubMed: 22794260]
51. Witten IB, Steinberg EE, Lee SY, Davidson TJ, Zalocusky KA, Brodsky M, Yizhar O, Cho SL, Gong S, Ramakrishnan C, Stuber GD, Tye KM, Janak PH, Deisseroth K. Recombinase-driver rat lines: tools, techniques, and optogenetic application to dopamine-mediated reinforcement. Neuron. 2011; 72:721–33. doi:10.1016/j.neuron.2011.10.028. [PubMed: 22153370]
52. Gage GJ, Stoetzner CR, Wiltschko AB, Berke JD. Selective activation of striatal fast-spiking interneurons during choice execution. Neuron. 2010; 67:466–79. doi:S0896-6273(10)00518-0 [pii] 10.1016/j.neuron.2010.06.034. [PubMed: 20696383]
53. Leventhal DK, Gage GJ, Schmidt R, Pettibone JR, Case AC, Berke JD. Basal ganglia beta oscillations accompany cue utilization. Neuron. 2012; 73:523–36. doi:10.1016/j.neuron. 2011.11.032. [PubMed: 22325204]
54. Schmidt R, Leventhal DK, Mallet N, Chen F, Berke JD. Canceling actions involves a race between basal ganglia pathways. Nat Neurosci. 2013; 16:1118–24. doi:10.1038/nn.3456. [PubMed: 23852117]
55. Lau B, Glimcher PW. Dynamic Response-by-Response Models of Matching Behavior in Rhesus Monkeys. Journal of the Experimental Analysis of Behavior. 2005; 84:555–579. doi:10.1901/jeab. 2005.110-04. [PubMed: 16596980]
56. Simen P, Cohen JD, Holmes P. Rapid. decision threshold modulation by reward rate in a neural network. Neural Netw. 2006; 19:1013–26. doi:10.1016/j.neunet.2006.05.038. [PubMed: 16987636]
57. Daw ND, Kakade S, Dayan P. Opponent interactions between serotonin and dopamine. Neural Netw. 2002; 15:603–16. [PubMed: 12371515]
58. Sugrue LP, Corrado GS, Newsome WT. Matching behavior and the representation of value in the parietal cortex. Science. 2004; 304:1782–7. doi:10.1126/science.1094765. [PubMed: 15205529]
59. Humphries MD, Khamassi M, Gurney K. Dopaminergic Control of the Exploration-Exploitation Trade-Off via the Basal Ganglia. Front Neurosci. 2012; 6:9. doi:10.3389/fnins.2012.00009. [PubMed: 22347155]
60. Beeler JA, Frazier CR, Zhuang X. Putting desire on a budget: dopamine and energy expenditure, reconciling reward and resources. Front Integr Neurosci. 2012; 6:49. doi:10.3389/fnint. 2012.00049. [PubMed: 22833718]
61. Song P, Mabrouk OS, Hershey ND, Kennedy RT. In vivo neurochemical monitoring using benzoyl chloride derivatization and liquid chromatography--mass spectrometry. Analytical chemistry. 2011; 84:412–419. [PubMed: 22118158]
62. Song P, Hershey ND, Mabrouk OS, Slaney TR, Kennedy RT. Mass spectrometry “sensor” for *in vivo* acetylcholine monitoring. Analytical chemistry. 2012; 84:4659–4664. [PubMed: 22616788]
63. Aragona BJ, Day JJ, Roitman MF, Cleaveland NA, Mark Wightman R, Carelli RM. Regional specificity in the real-time development of phasic dopamine transmission patterns during acquisition of a cue--cocaine association in rats. European Journal of Neuroscience. 2009; 30:1889–1899. [PubMed: 19912327]
64. Daw ND, Courville AC, Touretzky DS. Representation and timing in theories of the dopamine system. Neural computation. 2006; 18:1637–1677. [PubMed: 16764517]
65. Kile BM, Walsh PL, McElligott ZA, Bucher ES, Guillot TS, Salahpour A, Caron MG, Wightman RM. Optimizing the temporal resolution of fast-scan cyclic voltammetry. ACS chemical neuroscience. 2012; 3:285–292. [PubMed: 22708011]
66. Mazur JE. Tests of an equivalence rule for fixed and variable reinforcer delays. Journal of Experimental Psychology: Animal Behavior Processes. 1984; 10:426.
67. Ainslie G. Précis of breakdown of will. Behavioral and Brain Sciences. 2005; 28:635–649. [PubMed: 16262913]
68. Kobayashi S, Schultz W. Influence of reward delays on responses of dopamine neurons. J Neurosci. 2008; 28:7837–46. doi:10.1523/JNEUROSCI.1600-08.2008. [PubMed: 18667616]
69. Kacelnik A. Normative and descriptive models of decision making: time discounting and risk sensitivity. Characterizing human psychological adaptations. 1997; 208:51–66.

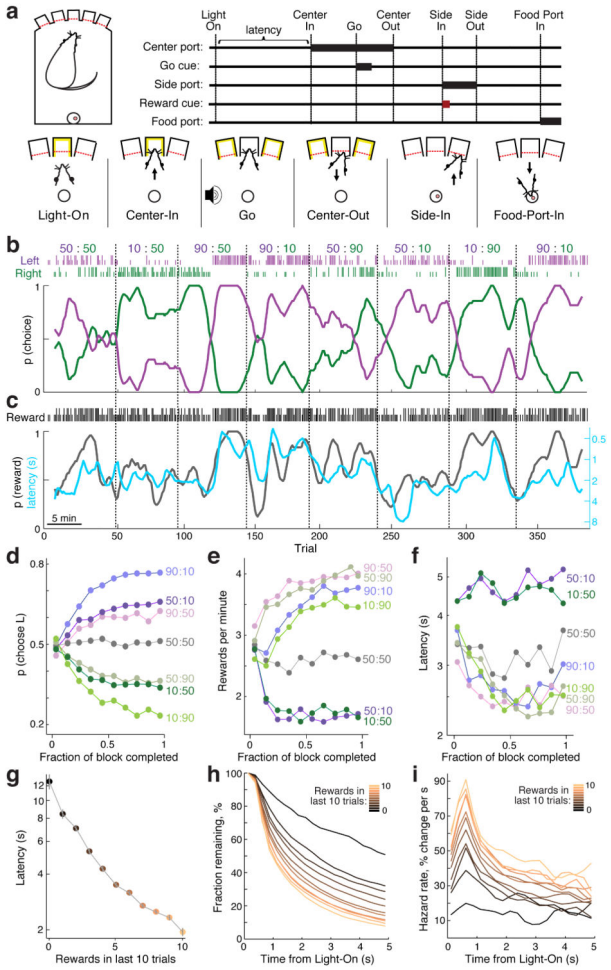

Figure 1. Adaptive choice and motivation in the trial-and-error task

(a) Sequence of behavioral events (in rewarded trials). (b) Choice behavior in a representative session. Numbers at top denote nominal block-by-block reward probabilities for left (purple) and right (green) choices. Tick marks indicate actual choices and outcomes on each trial (tall ticks indicate rewarded trials, short ticks unrewarded). The same choice data is shown below in smoothed form (thick lines; 7-trial smoothing). (c) Relationship between reward rate and latency for the same session. Here tick marks are used to indicate only whether trials were rewarded or not, regardless of choice. Solid black line shows reward rate, and cyan line shows latency (on inverted log scale), both smoothed in the same way as B. (d) Choices progressively adapt towards the block reward probabilities (data set for panels d-i: n = 14 rats, 125 sessions, 2738 +/− 284 trials per rat). (e) Reward rate breakdown by block reward probabilities. (f) Latencies by block reward probabilities. Latencies become rapidly shorter when reward rate is higher. (g) Latencies by proportion of recent trials rewarded. Error bars represent s.e.m. (h) Latency distributions presented as survivor curves (i.e. the average fraction of trials for which the Center-In event has not yet happened, by time elapsed from Light-On) broken down by proportion of recent trials rewarded. (i) Same latency distributions as panel h, but presented as hazard rates (i.e. the instantaneous probability that the Center-In event will happen, if it has not happened yet).

The initial bump in the first second after Light-On reflects engaged trials (see Supplementary Fig.1), after that hazard rates are relatively stable and continue to scale with reward history.

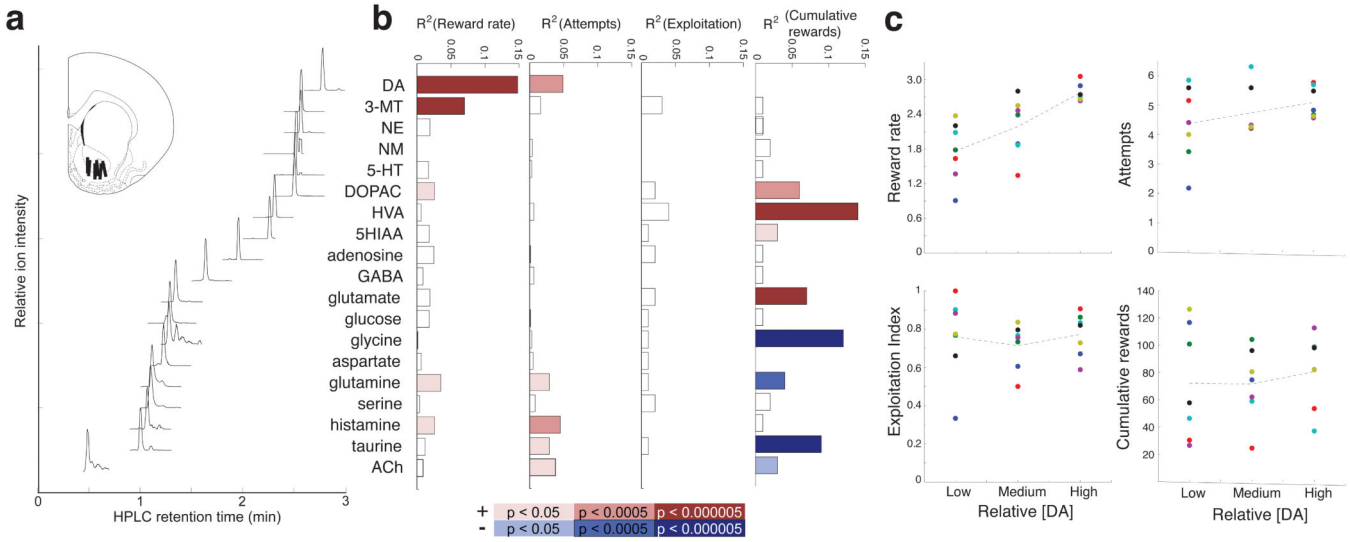

Figure 2. Minute-by-minute dopamine levels track reward rate

(a) Total ion chromatogram of a single representative microdialysis sample, illustrating the set of detected analytes in this experiment. X-axis indicates HPLC retention times, y-axis indicates intensity of ion detection for each analyte (normalized to peak values). (Inset) locations of each microdialysis probe within the nucleus accumbens (all data shown in the same Paxinos atlas section; six were on the left side and one on the right). Abbreviations: DA, dopamine; 3-MT, 3-methoxytyramine; NE, norepinephrine; NM, normetanephrine; 5-HT, serotonin; DOPAC, 3,4-dihydroxyphenylacetate acid; HVA, homovanillic acid; 5HIAA, 5-hydroxyindole-3-acetic acid, GABA, γ-aminobutyric acid; ACh, acetylcholine. (b) Regression analysis results indicating strength of linear relationships between each analyte and each of four behavioral measures (reward rate; number of attempts; exploitation index; and cumulative rewards). Data are from 6 rats (7 sessions, total of 444 one-minute samples). Color scale shows p-values, Bonferroni-corrected for multiple comparisons (4 behavioral measures * 19 analytes), with red bars indicating a positive relationship and blue bars a negative relationship. Since both reward rate and attempts showed significant correlations with [DA], we constructed a regression model that included these predictors and an interaction term. In this model R^2 remained at 0.15 and only reward rate showed a significant partial effect ($p < 2.38\times10^{-12}$). (c) An alternative assessment of the relationship between minute-long [DA] samples and behavioral variables. Within each of the seven sessions [DA] levels were divided into three equal-sized bins (LOW, MEDIUM, HIGH); different colors indicate different sessions. For each behavioral variable, means were compared across [DA] levels using one-way ANOVA. There was a significant main effect of reward rate ($F(2,18)=10.02$, $p=0.0012$), but no effect of attempts ($F(2,18)=1.21$, $p=0.32$), exploitation index ($F(2,18)=0.081$, $p=0.92$), or cumulative rewards ($F(2,18)=0.181$, $p=0.84$). Post-hoc comparisons using the Tukey test revealed that the mean reward rates of LOW and HIGH [DA] differed significantly ($p=0.00082$). See also Supplementary Figs. 2,3.

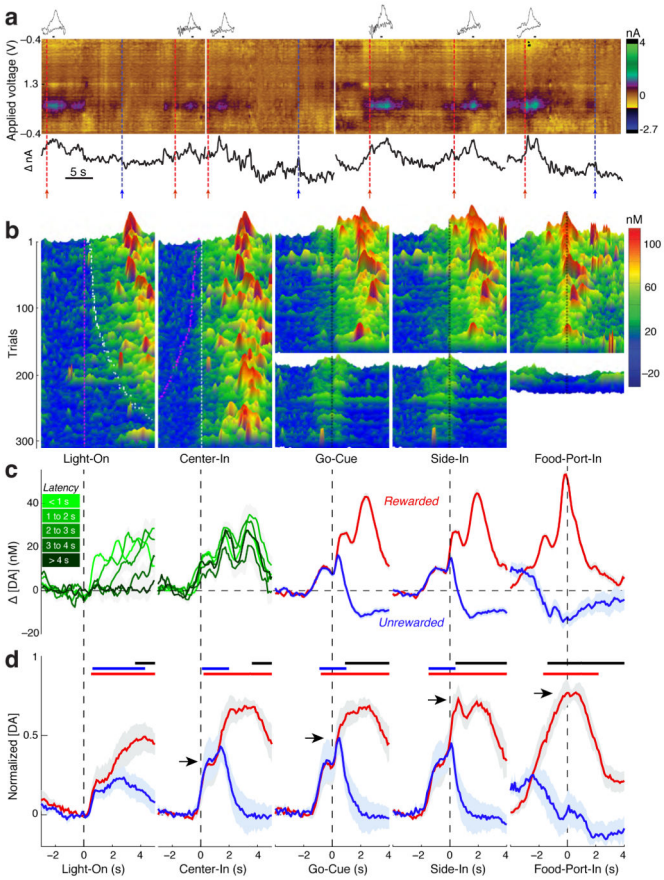

Figure 3. A succession of within-trial dopamine increases

(a) Examples of FSCV data from a single session. Color plots display consecutive voltammograms (every 0.1s) as a vertical colored strip; examples of individual voltammograms are shown at top (taken from marked time points). Dashed vertical lines indicate Side-In events for rewarded (red) and unrewarded (blue) trials. Black traces below indicate raw current values, at the applied voltage corresponding to the dopamine peak. (b) [DA] fluctuations for each of the 312 completed trials of the same session, aligned to key behavioral events. For Light-On and Center-In alignments, trials are sorted by latency (pink dots mark Light-On times; white dots mark Center-In times). For the other alignments rewarded (top) and unrewarded (bottom) trials are shown separately, but otherwise in the order in which they occurred. [DA] changes aligned to Light-On were assessed relative to a 2s baseline period, ending 1s before Light-On. For the other alignments, [DA] is shown relative to a 2s baseline ending 1s before Center-In. (c) Average [DA] changes during a single session (same data as b; shaded area represents s.e.m.). (d) Average event-aligned [DA] change across all six animals, for rewarded and unrewarded trials (see Supplementary Fig.4 for each individual session). Data are normalized by the peak average rewarded [DA] in each session, and are shown relative to the same baseline epochs as in b. Black arrows indicate increasing levels of event-related [DA] during the progression through rewarded trials. Colored bars at top indicate time periods with statistically significant differences (red, rewarded trials greater than baseline, one-tailed t-tests for each 100ms time point

individually; blue, same for unrewarded trials; black, rewarded trials different to unrewarded trials, 2-tailed t-tests; all statistical thresholds set to p=0.05, uncorrected).

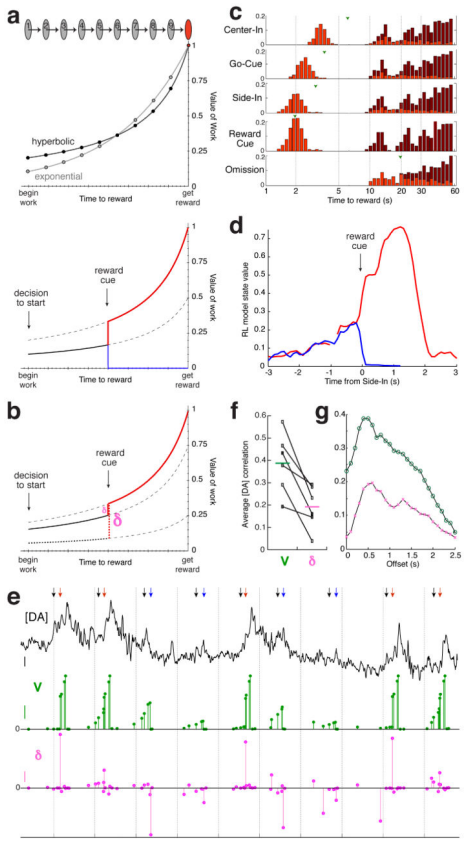

Figure 4. Within-trial dopamine fluctuations reflect state value dynamics

(a) *Top,* Temporal discounting: the motivational value of rewards is lower when they are distant in time. With the exponential discounting commonly used in RL models, value is lower by a constant factor γ for each time step of separation from reward. People and other animals may actually use hyperbolic discounting which can optimize reward rate (since rewards/time is inherently hyperbolic). Time parameters are here chosen simply to illustrate the distinct curve shapes. *Bottom.* Effect of reward cue, or omission, on state value. At trial start the discounted value of a future reward will be less if that reward is less likely. Lower value provides less motivational drive to start work - producing e.g. longer latencies. If a cue signals that upcoming reward is certain, the value function jumps up to the (discounted) value of that reward. For simplicity, the value of subsequent rewards is not included. (b) The reward prediction error δ reflects abrupt changes in state value. If the discounted value of work reflects an unlikely reward (e.g. probability = 0.25) a reward cue prompts a larger δ than if the reward was likely (e.g. probability = 0.75). Note that in this idealized example, δ would be zero at all other times. (c) *Top,* Task events signal updated times-to-reward. Data is from the same example session as Fig.3c. Bright red indicates times to the very next reward, dark red indicates subsequent rewards. Green arrowheads indicate average times to next reward (harmonic mean, only including rewards in the next 60s). As the trial progresses, average times-to-reward get shorter. If the reward cue is received, rewards are reliably obtained ~2s later. Task events are considered to prompt transitions between different internal states (Supplementary Fig.5) whose learned values reflect these different experienced times-to-reward. (d) Average state value of the RL model for rewarded (red) and

unrewarded (blue) trials, aligned on the Side-In event. The exponentially-discounting model received the same sequence of events as in Fig.3c, and model parameters (γ=0.68, γ=0.98) were chosen for the strongest correlation to behavior (comparing state values at Center-In to latencies in this session, Spearman r=-0.34). Model values were binned at 100ms, and only bins with at least 3 events (state transitions) were plotted. (e) Example of the [DA] signal during a subset of trials from the same session, compared to model variables. Black arrows indicate Center-In events, red arrows Side-In with Reward Cue, blue arrows Side-In alone (Omission). Scale bars are: [DA], 20nM; V, 0.2; δ, 0.2. Dashed grey lines mark the passage of time in 10s intervals. (f) Within-trial [DA] fluctuations are more strongly correlated with model state value (V) than with RPE (δ). For every rat the [DA] : V correlation was significant (number of trials for each rat: 312, 229, 345, 252, 200, 204; $p<10^{-14}$ in each case; Wilcoxon signed-rank test of null hypothesis that median correlation within trials is zero) and significantly greater than the [DA] : δ correlation ($p<10^{-24}$ in each case, Wilcoxon signed-rank test). Groupwise, both [DA] : V and [DA] : δ correlations were significantly non-zero, and the difference between them was also significant (n=6 sessions, all comparisons p=0.031, Wilcoxon signed-rank test). Model parameters (γ=0.4, γ =0.95) were chosen to maximize the average behavioral correlation across all 6 rats (Spearman r = −0.28), but the stronger [DA] correlation to V than to δ was seen for all parameter combinations (Supplementary Fig.5). (g) Model variables were maximally correlated with [DA] signals ~0.5s later, consistent with a slight delay caused by the time taken by the brain to process cues, and by the FSCV technique.

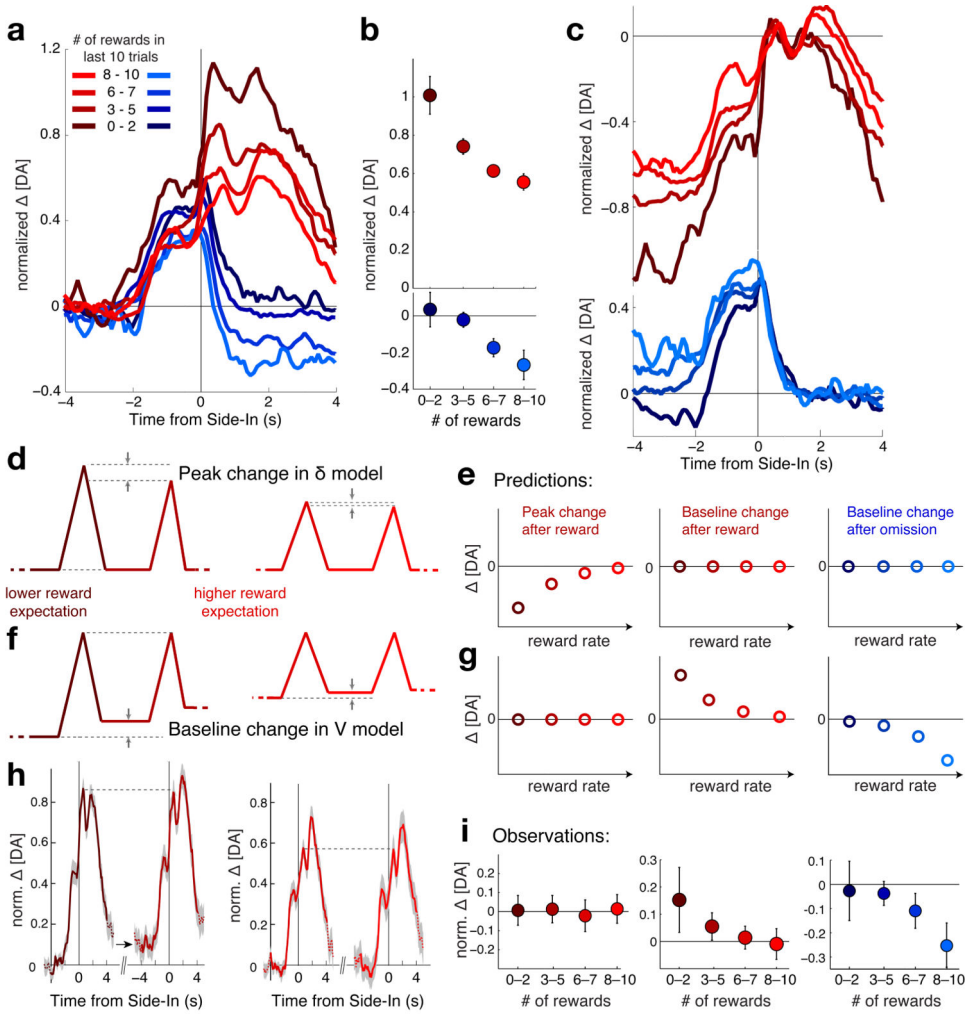

Figure 5. Between-trial dopamine shifts reflect updated state values

(a) Less-expected outcomes provoke larger changes in [DA]. [DA] data from all FSCV sessions together (as in Fig.3d), broken down by recent reward history and shown relative to pre-trial “baseline" (−3 to −1s relative to Center-In). Note that the [DA] changes after reward omission last at least several seconds (shift in level), rather than showing a highly transient dip followed by return to baseline as might be expected for encoding RPEs alone. (b) Quantification of [DA] changes, between baseline and reward feedback (0.5-1.0s after Side-In for rewarded trials, 1s-3s after Side-In for unrewarded trials). Error bars show SEM. (c) Same data as (a), but plotted relative to [DA] levels *after* reward feedback. These [DA] observations are consistent with a variable “baseline” whose level depends on recent reward history (as in Fig.4b model). (d) Alternative accounts of [DA] make different predictions for between-trial [DA] changes. When reward expectation is low, rewarded trials provoke large RPEs, but across repeated consecutive rewards RPEs should decline. Therefore if absolute [DA] levels encode RPE, the peak [DA] evoked by the reward-cue should decline between consecutive rewarded trials (and baseline levels should not change). For simplicity this cartoon omits detailed within-trial dynamics. (e) Predicted pattern of [DA] change under this account, which also does not predict any baseline shift after reward omissions (right). (f) If

instead [DA] encodes state values, then peak [DA] should not decline from one reward to the next, but the baseline level should increase (and decrease following unrewarded trials). (g) Predicted pattern of [DA] change for this alternative account. (h) Unexpected rewards cause a shift in baseline, not in peak [DA]. Average FSCV data from consecutive pairs of rewarded trials (all FSCV sessions combined, as in a), shown relative to the pre-trial baseline of the first trial in each pair. Data were grouped into lower reward expectation (left pair of plots, 165 total trials; average time between Side-In events = 11.35s +/− 0.22s SEM) and higher reward expectation (right pair of plots, 152 total trials; time between Side-In events = 11.65s +/− 0.23s) by a median split of each individual session (using # rewards in last 10 trials). Dashed lines indicate that reward cues evoked a similar absolute level of [DA] in the second rewarded trial, compared to the first. Black arrow indicates the elevated pre-trial [DA] level for the second trial in the pair (mean change in baseline [DA] = 0.108, p=0.013, one-tailed Wilcoxon signed rank test). No comparable change was observed if the first reward was more expected (right pair of plots; mean change in baseline [DA] = 0.0013, p=0.108, one-tailed Wilcoxon signed rank test). (i) [DA] changes between consecutive trials follow the pattern expected for value coding, rather than RPE coding alone.

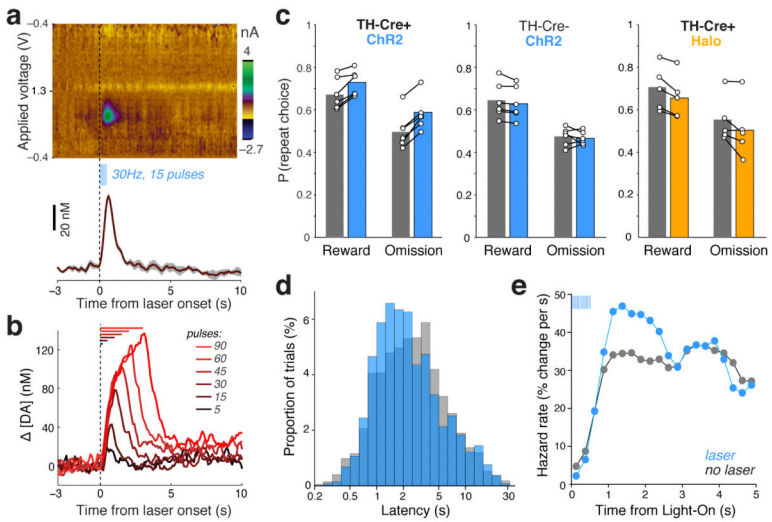

Figure 6. Phasic dopamine manipulations affect both learning and motivation

(a) FSCV measurement of optogenetically-evoked [DA] increases. Optic fibers were placed above VTA, and [DA] change examined in nucleus accumbens core. Example shows dopamine release evoked by a 0.5s stimulation train (average of 6 stimulation events, shaded area indicates +/−SEM). (b) Effect of varying the number of laser pulses on evoked dopamine release, for the same 30Hz stimulation frequency. (c) Dopaminergic stimulation at Side-In reinforces the chosen left or right action. *Left,* in *TH-Cre*$^{+}$ rats stimulation of ChR2 increased the probability that the same action would be repeated on the next trial. Circles indicate average data for each of 6 rats (3 sessions each, 384 trials/session ± 9.5 SEM). *Middle,* this effect did not occur in *TH-Cre*$^{-}$ littermate controls (6 rats, 3 sessions each, 342±7 trials/session). *Right,* in *TH-Cre*$^{+}$ rats expressing Halorhodopsin, orange laser stimulation at Side-In reduced the chance that the chosen action was repeated on the next trial (5 rats, 3 sessions each, 336±10 trials/session). See Supplementary Fig.8 for additional analyses. (d) Laser stimulation at Light-On causes a shift towards sooner engagement, if the rats were not already engaged. Latency distribution (on log scale, 10 bins per log unit) for non-engaged, completed trials in *TH-Cre*$^{+}$ rats with ChR2 (n=4 rats with video analysis; see Supplementary Fig.9 for additional analyses). (e) Same latency data as d, but presented as hazard rates. Laser stimulation (blue ticks at top left) increases the chance that rats will decide to initiate an approach, resulting in more Center-In events 1-2s later (for these n=4 rats, one-way ANOVA on hazard rate $F(1,3) = 18.1$, $p=0.024$). See Supplementary Fig.10 for hazard rate time courses from the individual rats.