Action Initiation Shapes Mesolimbic Dopamine Encoding of Future Rewards

Emilie C.J. Syed[1,2,3], Laura L. Grima[3], Peter J. Magill[2], Rafal Bogacz[1,2], Peter Brown[1,2,4], Mark E. Walton[3,4]

[1] Nuffield Department of Clinical Neurosciences; [2] Medical Research Council Brain Network Dynamics Unit, Department of Pharmacology; [3] Department of Experimental Psychology; University of Oxford

[4] These authors jointly directed the work.

Correspondence to:

mark.walton@psy.ox.ac.uk

It is widely held that dopamine signaling encodes predictions of future rewards and such predictions are regularly used to drive behavior, but the relationship between these two is poorly defined. Here, we demonstrate in rats that nucleus accumbens dopamine following a reward-predicting cue is attenuated unless movement is correctly initiated. These results demonstrate that dopamine release in this region is contingent upon correct action initiation and not just reward prediction.

The phasic activity of many dopamine-containing neurons codes a quantitative reward prediction error (RPE)[1-4]. Consistent with this, the firing of midbrain dopamine neurons, and dopamine release in their targets like the nucleus accumbens core (NAcc), in response to predictive cues both scale with anticipated future rewards[5-9]. Moreover, manipulations of mesolimbic dopamine indicate dopaminergic transmission is not only required to drive behavioral responses to incentive cues but can also facilitate action initiation[10-13]. However, to date, the precise relationship between cue-elicited dopamine release, reward prediction and movement remains ambiguous, partly because little is known about dopamine release dynamics in the context of withholding an action to gain reward.

To address this issue directly, we trained rats to perform symmetrically-rewarded Go/No-Go tasks, and then monitored subsond dopamine concentration in the NAcc using fast-scan cyclic voltammetry during task performance (**Fig. 1, Supplementary Fig. 1** and Online Methods). In the first task we used, a trial was initiated when the rat voluntarily entered and stayed in a central nose-poke for 0.5 s, after which they were presented with one of three auditory cues instructing the animal to Go Left (GoL), Go Right (GoR) or remain in the nose-poke (No-Go) (**Fig. 1**). Go trials required animals to exit the nose-poke and make 2 responses on the correct lever within 5 s of cue onset to receive reward. During No-Go trials, reward delivery was contingent on the animal staying in the nose-poke for at least a further 1.7–1.9 s. The task design ensured that the successful outcome – and therefore reward prediction at cue presentation – was matched for all three trial types, while action requirements of each trial type were distinct.

Animals performed with a similar success rate on all three trial types ($\chi^2(2) = 4.88$, $p = 0.17$), and rapidly initiated a response on correct Go trials (head exit (mean ± S.E.M.): 0.58 ± 0.06 s after cue onset) while refraining from responding on No-Go trials (head exit: 2.33 ± 0.05 s after cue onset; **Supplementary Fig. 2a,b**). Importantly, the time from instructive cue to reward delivery was comparable on both Go and No-Go trials (Fig 1d, $\chi^2(2) = 1.66$, $p = 0.4$; **Supplementary Fig. 2c**). Therefore, any disparity in dopamine release in the different trial conditions cannot be explained by differences in the temporally-discounted expected value associated with instructive cues. Indeed, when we simulated performance on this task using a simple reinforcement learning model (standard actor-critic model), there were comparable positive RPEs at cue presentation in all conditions in spite of the model displaying an equivalent range of correct trials in the different conditions (**Supplementary Fig. 2d,e**).

As previously observed[5, 14], NAcc dopamine concentration rapidly increased following presentation of either the GoL or GoR cue on correctly performed trials, peaking just prior to reward collection (**Fig. 2a,c**, **Supplementary Fig. 3**). In striking contrast, there was no equivalent increase in dopamine during presentation of the No-Go cue when the rats successfully stayed in the nose-poke, even though this cue conveyed quantitatively similar information about future rewards as either of the Go cues (**Fig. 2b,c**). Accordingly, direct comparison of the post-cue dopamine signals using an auROC analysis (see Supplementary Methods) demonstrated significant discrimination between Go and No-Go trials during the cue period (**Fig. 2c**). These data suggest that NAcc dopamine release is modulated by action initiation and not just reward prediction.

To further explore the relationship between movement and NAcc dopamine release, we re-aligned the dopamine signal in all trials to the moment of action initiation, defined as the time when the animal exited the nose-poke (**Fig. 2d**). Here, after head exit, we observed a rapid increase in dopamine on both Go and No-Go trials; dopamine concentrations could no longer be used to distinguish between trial types within a 2.5 s window after movement initiation, confirming that the difference in dopamine signals on the Go and No-Go trial types is related to differences in response requirements. Indeed, modulation of dopamine release by action initiation was not only observed when comparing "valid" Go trials to No-Go trials; it was also possible to discriminate a subset of "delayed" Go trials (excluded from the analysis above), where rats held the nose-poke for ≥1.7s after a Go cue prior to pressing the correct lever, from valid Go trials when the signals were time-locked to cue onset (**Supplementary Fig. 4a**). Again, this trial type distinction disappeared when dopamine signals were re-aligned to movement initiation (**Supplementary Fig. 4b**). Intriguingly, it was not the case that the initiation of *any* action resulted in dopamine release. Comparison of dopamine signals on correct and wrong Go trials (the latter being defined where a response was correctly initiated but ended in selection of the incorrect lever) not only showed that dopamine did not increase during wrong Go trials, but also that it was possible to discriminate the two trial types after head exit, but before the error was explicitly signaled at the time of the first lever press (**Supplementary Fig. 5**). This influence of action accuracy on dopamine release could not be accounted for by differences in response latencies (**Supplementary Fig. 5**).

Although these additional data demonstrate that NAcc dopamine release is clearly modulated by movement initiation and not just changes in reward prediction following

instructive cues, they do not relegate the overall importance of the relationship between reward prediction and dopamine levels. For instance, during failed No-Go trials in which the animals did not sustain the nose poke for >1.7 s after cue, there was a significant negative deflection in dopamine immediately after the animals exited the nose poke; this change coincided with the houselight turning on to signal that there would be no reward in the current trial (**Supplementary Fig. 6**). Moreover, it was possible to discriminate dopamine signals after cue onset on Go trials when they were arranged in each rat by preference (high / low success) rather than side (left / right) (**Supplementary Fig. 7**). To explore this formally, we performed a linear regression on all cue-aligned dopamine release on correctly performed trials, with explanatory variables of interest of (1) success rate and (2) holding time acting as proxies for reward prediction and action initiation respectively. Crucially, *both* factors were significant predictors of NAcc dopamine release during the cue period (**Fig. 2e**).

Experiment 1 demonstrates that action initiation has a marked influence on dopamine release when all trials are equally rewarded. To address whether this observation would hold true when cues not only instruct the response but also signal the potential magnitude of reward to be gained, we conducted a sond Go/No-Go task study. Experiment 2 was performed on another animal group and was similar to Experiment 1, except that there were now 4 possible auditory instructing cues, which signaled both the action requirement (Go/No-Go) and future reward (Large/Small) in a factorial design (**Fig. 3a**). In this task, animals performed significantly better on Go Large trials ($\chi^2(3) = 13.62$, $p = 0.004$; Go Large trials versus other trial types, (all $p = 0.03$, $W_6 = 0$, Wilcoxon Signed Ranks Test); however, success rates of Go Small and both No-Go types were comparable (Go Small v No-Go Small:

p = 0.44, W_6 = 15; Go Small v No-Go Big: p = 0.69, W_6 = 13; No-Go Small v No-Go Big: p = 0.31, W_6 = 16) (**Supplementary Fig. 8a**). Nose poke exit latencies were significantly faster on Go than No-Go trials and, importantly, on Large than Small Reward trials for both conditions (both p = 0.03, W_6 = 0, Wilcoxon Signed Ranks Test) (**Supplementary Fig. 8b**), showing that the rats understood the cue-response-reward associations.

Dopamine levels rapidly increased after Go Large cues, similar to observations in Experiment 1 (**Fig. 3b**). However, dopamine release initially decreased following Go Small cues (**Fig. 3b**). This is consistent with a positive and negative RPE, respectively. There was also a small transient increase in dopamine following the No-Go Large cue and a decrease following the No-Go Small cue. Crucially, however, the former increase was markedly attenuated compared to the Go Large trials, and dopamine release in both No-Go conditions remained suppressed below the level of either Go condition while the animals correctly stayed in the nose-poke and delayed their actions (**Fig. 3b**, **Supplementary Fig. 9**). This resulted in there being a significant interaction between RPE and action initiation, as well as main effects of RPE and action initiation, during the 2.5s post-cue onset (**Fig. 3c**). Once again, when all the data were re-aligned to the point of movement initiation, dopamine release increased in all correctly performed conditions (**Fig. 3d**, **Supplementary Fig. 10**). Therefore, dopamine release is not only modulated by the learned outcome of a movement sequence but also by initiating that response.

Taken together, the present data demonstrate that, after cue associations have been acquired, NAcc dopamine release in response to instructive reward-predicting cues is shaped by correct movement initiation and not only RPE coding. While the increase in

dopamine levels is not straightforwardly related to either the latency to initiate movement (**Supplementary Fig. 5d**) or the speed to reach the lever (**Supplementary Fig. 7b**), it may nonetheless partly reflect the vigor of action initiation or confidence in the accuracy of the action. Several studies have already hinted at a connection between striatal dopamine and the promotion of reward seeking through movement[10, 12-16]. Here, we unequivocally confirm this link for mesolimbic dopamine by demonstrating that instructive reward-predicting cues elicit increased dopamine release if and when a correct goal-directed action is initiated. Although our data do not allow us to determine whether dopamine release is causally responsible for appropriate action initiation, they nevertheless support the possibility that the rapidly-evolving transient increases in NAcc dopamine may act to promote the correct and prompt selection and execution of actions to enable reward to be efficiently realized.

Acknowledgements

This work was funded by a Wellcome Trust Research Career Development Fellowship (WT090051MA to MEW), the Medical Research Council UK (awards MC_UU_12020/5 and MC_UU_12024/2 to PJM, MC_UU_12024/5 to RB and MC_UU_12024/1 to PB), and a studentship from the Economic and Social Research Council and St John's College Oxford (to LLG). We would like to thank Mathieu Baudonnat and Lev Tankelevitch for assistance with data collection, Nick Hollon and Nils Kolling for analysis advice, and Scott Ng-Evans for technical support.

Author contributions

PB conceived the core study, and PB, PJM and MEW developed this and, with EJCS, planned the experiments. For Experiments 1 and 2, EJCS performed surgeries, collected and analyzed the data; for Experiment 2, LLG also performed surgery and collected the data. RB performed the simulations. MEW supervised the study. MEW and EJCS prepared the manuscript, with input from the other authors.

References

1. Montague, P.R., Dayan, P. & Sejnowski, T.J. *J Neurosci* **16**, 1936-1947 (1996).
2. Schultz, W., Dayan, P. & Montague, P.R. *Science* **275**, 1593-1599 (1997).
3. Bayer, H.M. & Glimcher, P.W. *Neuron* **47**, 129-141 (2005).
4. Cohen, J.Y., Haesler, S., Vong, L., Lowell, B.B. & Uchida, N. *Nature* **482**, 85-88 (2012).
5. Gan, J.O., Walton, M.E. & Phillips, P.E. *Nat Neurosci* **13**, 25-27 (2010).
6. Tobler, P.N., Fiorillo, C.D. & Schultz, W. *Science* **307**, 1642-1645 (2005).
7. Roesch, M.R., Calu, D.J. & Schoenbaum, G. *Nat Neurosci* **10**, 1615-1624 (2007).
8. Day, J.J., Roitman, M.F., Wightman, R.M. & Carelli, R.M. *Nat Neurosci* **10**, 1020-1028 (2007).
9. Howe, M.W., Tierney, P.L., Sandberg, S.G., Phillips, P.E. & Graybiel, A.M. *Nature* **500**, 575-579 (2013).
10. Nicola, S.M. *J Neurosci* **30**, 16585-16600 (2010).
11. Robbins, T.W. & Everitt, B.J. *Psychopharmacology* **191**, 433-437 (2007).
12. Flagel, S.B.*, et al. Nature* **469**, 53-57 (2011).
13. Phillips, P.E., Stuber, G.D., Heien, M.L., Wightman, R.M. & Carelli, R.M. *Nature* **422**, 614-618 (2003).
14. Roitman, M.F., Stuber, G.D., Phillips, P.E., Wightman, R.M. & Carelli, R.M. *J Neurosci* **24**, 1265-1271 (2004).
15. Guitart-Masip, M.*, et al. Proc Natl Acad Sci U S A* **109**, 7511-7516 (2012).
16. Jin, X. & Costa, R.M. *Nature* **466**, 457-462 (2010).

Figure Legends

Figure 1. Go/No-Go task and behavioral performance in Experiment 1. Schematic of the behavioral task (left) and trial types (right). Grey shading marks the period when the auditory cues remain on.

Figure 2. NAcc dopamine signals on Go and No-Go trials in Experiment 1. (a-b) Example single trial FCV recordings during a Go **(a)** and a No-Go **(b)** trial. Upper plots depict the dopamine level over time for the two trials types (L1 = first lever press). Inset, an example cyclic voltammogram identifying the detected current as dopamine. Color plots in the lower panels show the background-subtracted cyclic voltammograms as a function of the applied voltage over time. **(c-d)** Average changes in dopamine levels (mean ± S.E.M., n=7 rats / electrodes) recorded during different trial types aligned to cue onset **(c)** or to time of head exit from nose-poke **(d)**, along with average discriminability between the Go and No-Go trial types for each timepoint (gray lines = max / min of 1,000 permuted sessions; line: *, $p < 0.05$ permutation tests, corrected for multiple comparisons). Gray shading marks a 2.5 s window used to focus the analyses. As can be observed, the sustained increase in dopamine levels observed on Go trials after cue presentation is delayed on No-Go trials until after movement initiation. **(e)** Average absolute effect sizes (mean ± S.E.M.) from a general linear model with regressors for (i) success rate and (ii) action initiation time. (line: *, $p < 0.05$ permutation tests, corrected for multiple comparisons). Both factors were significant predictors of NAcc dopamine release during the cue period. a.u., arbitrary units

Figure 3. Experiment 2: behavior and dopamine signals. (a) Schematic of the trial types and **(b,d)** Average changes in dopamine levels (mean ± S.E.M. n = 9 electrodes from 6 rats)

recorded during different trial types aligned to cue onset **(b)** or to time of head exit from nose-poke **(d)**. **(c)** Average absolute effect sizes (mean ± S.E.M.) from a general linear model with regressors for RPE (success rate x reward size), action initiation time, and their interaction. (line: *, $p < 0.05$ permutation tests, corrected for multiple comparisons). a.u., arbitrary units.

ONLINE METHODS

Action Initiation Shapes Mesolimbic Dopamine Encoding of Future Rewards

Emilie C.J. Syed, Laura L Grima, Peter J. Magill, Rafal Bogacz, Peter Brown, Mark E. Walton

MATERIALS AND METHODS

All procedures were carried out in accordance with the UK Animals (Scientific Procedures) Act (1986) and its associated guidelines. For Experiment 1, a total of 17 naïve male Sprague-Dawley rats were used (Harlan, UK), aged ~ 2 months at the start of training. 2 rats were excluded for being unable to maintain a No-Go response for the required time, 5 rats were excluded for misplaced electrodes outside the nucleus accumbens core (either in the medial or ventral shell), 1 rat was unable to be connected to the recording device due to a misaligned implant, and 2 rats had broken/noisy electrodes. This left a total of 7 rats / electrodes that provided the data included for Experiment 1. For Experiment 2, a total of 12 naïve male Sprague-Dawley rats were used (Harlan, UK), aged ~5 months at the start of training. 1 rat was culled due to post-surgical complication; and out of the 22 remaining electrodes, 6 were broken/noisy, and 7 were misplaced. This left a total of 6 rats with 9 working electrodes that provided the data included for Experiment 2. Animals were maintained on a twelve-hour light/dark cycle (lights on 07.00) and were group housed during initial habituation and training but individually housed following surgery. All testing was carried out during the light phase. During the training and testing periods, access to food was restricted such that rats' weights were kept ~85–90% of their free-feeding weight. Water was available *ad libitum* while animals were in their home cages.

Behavioral Task

Apparatus

Testing was carried out in operant chambers (30.5 x 24.1 x 29.2 cm; Med Associates, VT, USA). Each chamber was housed within a custom-built sound-attenuating cabinet ventilated with a fan, which provided constant background noise of ~64 dB. Each chamber contained two retractable levers 9.5cm on either side of a central nose-poke, which was fitted with an infrared beam signaling when animals entered the receptacle. The wall opposite was fitted with an extra-tall food magazine into which sucrose pellets (Sandown Scientific, UK) could be dispensed. Each chamber was also fitted with a house-light and a speaker for delivering auditory stimuli.

Training

Animals were first habituated to the conditioning box and learned to retrieve pellet rewards from the food magazine tray. Rats then commenced training on the No-Go trial type. Testing was carried out with the house light turned off. To initiate a trial, the animal voluntarily made and sustained a head entry into the nose-poke.

For training animals for use in Experiment 1, on session 1, 20 ms after nose-poke entry, a single auditory cue was presented (either a tone, buzz or white noise, counterbalanced across animals, each ~70 dB), to signal a “No Go” trial. If the animals stayed in the nose poke for another 230 ms, a single 20 mg sucrose pellet was delivered to the food magazine on the opposite wall. A 5s inter-trial interval (ITI) then commenced during which the nose-poke remained inactive; the end of the ITI was not signaled by any external cue. A premature head exit caused the house light to be illuminated for the

duration of a 5s time-out immediately as the animal exited the nose poke, after which the house light turned off and a standard 5 s ITI commenced.

On reaching behavioral criterion (success rate >= 60%), the time of cue and of pellet delivery was gradually extended across sessions using the same criterion until the rats were able to perform the pre-cue nose-poke period of 0.5s and maximum post-cue hold period of 1.9s.

After the No-Go trial type was learned, the animals were then trained on the Go trials. They first learned to press each lever on a fixed ratio 1 schedule until they had made at least 20 presses on each lever. In the subsequent sessions, they trained on a simplified version of the task without No-Go trials. The two levers were extended during the entire session. As with No-Go training, a nose-poke entry sustained for 0.5s would elicit one of two different auditory cues, one for “Go Left” trials and the other for “Go Right” trials. This cue would stay on for 60 s or until the rat pressed a lever. Pressing the correct lever during cue presentation would elicit a single 20 mg sucrose pellet to be delivered to the food magazine on the opposite wall. Pressing the incorrect lever or failing to press a lever during the duration of the cue would result in the house light illuminating for a 5 s time-out period, before the house light turned off and the ITI commenced. During training, an error-correction procedure was used so that the next trial after an error would always be of the same cue/trial-type with the wrong lever withdrawn. Once a criterion of ≥ 60% successful Go responses was reached, the cue duration (and therefore maximum reaction time) was lowered to 5s, and then interleaved No-Go trials were re-introduced. Once an average ≥ 60% success rate on all three trial types within a session was achieved, the number of necessary lever presses on Go trials was increased to two, error correction trials were removed and the full task commenced.

Training protocols were largely similar for Experiment 2, except that four cues were used adding a clicker sound to the tone, white noise and buzz cues. As before, cue associations were pseudo-counterbalanced, restricted by the constraints of the initial group size (12 rats) and number of possible cue combinations (24 potential combinations). Rats were again initially trained on a single No-Go trial type with a single pellet as reward ("small" reward); only once this had been acquired was the second No-Go cue introduced, associated with two food pellets ("large" reward). Once they had achieved criterion (success rate >= 60% for both No-Go trial types), the animals then commenced training for the Go trials. Training for the Go trials was as described previously except that one lever was rewarded with a small reward and the other a large reward (side counterbalanced across animals). Note that the requirements for successful completion of both large and small reward No-Go trials and both large and small Go trials were identical. The rest of the training protocol was identical to Experiment 1, including the error-correction procedure.

Behavior: Recording Sessions

Experiment 1. Each session commenced with the house light turning off and the two levers extending into the chamber. A trial started after the rat had voluntarily entered the nose-poke and remained there for a pre-cue period of 0.5s. Exiting the nose-poke before this period would result in an aborted trial, with no consequences other than that the nose-poke timer reset to zero. Note that such aborted trials were common: on average, only 54.41% ± 2.88 SD of initiated nose-pokes after the ITI period had ended lasted >500ms and resulted in cue presentation (61.31% ± 2.01 SD in Experiment 2). Moreover, on ~20% of trials, animals first initiated a nose-poke during the ITI, which could not trigger cue presentation. This meant that the cues did not deterministically follow nose poke entry after the ITI and so did

carry meaningful information about a state change (see "Actor-critic model" below and **Supplementary Fig. 2d,e** for more details). Following a successfully sustained pre-cue period, one of three auditory cues would sound: a tone, buzz or white noise indicating either a Go Right, Go Left or No-Go trial; each with a 33% probability. This distribution maintained that a majority of trials were Go trials, thereby making the Go lever-press a pre-potent response[17]. On Go trials, the auditory cue sounded until animals pressed the correct lever twice (Go Correct) or until they pressed the wrong lever (Go Wrong) or for a maximum of 5 s if they failed to press any lever (Go Fail). On No-Go trials, rats had to maintain their position in the nose-poke for a hold period of 1.7–1.9s, jittered over trials. The No-Go cue sounded until the end of the hold period (No-Go Correct) or, if the rats exited the nose-poke prematurely, until the time of nose poke exit (No-Go Fail). In the case of any successful trial – either at the time of the sond lever press on a Go trial or at the end of the hold period on a No-Go trial – a 20 mg sucrose pellet was immediately delivered to the food magazine, after which a 5s ITI commenced during which animals were unable to initiate a new trial. No cue indicated the end of the ITI and animals were free to initiate the subsequent trial whenever they chose. In the case of an error or failed trial, the house-light would immediately illuminate for a 5 s time-out period as the animal exited the nose-poke, after which the house light turned off and the usual ITI commenced. Each session ended after animals had either gained 100 rewards or had worked for 60 minutes, whichever came first. This resulted in at least 30 trials of each trial type in each recording session.

Experiment 2. Each session was run in a very similar manner to Experiment 1. However, there were several key changes. First, the pre-cue period was randomly jittered between 0.3–0.7s such that the rat would not be able to fully predict cue onset timing.

Sond, following a successful sustained nose-poke during the pre-cue period, one of four auditory cues sounded – tone, buzz, white noise or click – each with a 25% probability. These cues indicated: (1) a No Go trial associated with a large reward (two 20 mg sucrose pellets), 2) a No Go trial associated with a small reward (one pellet), 3) a Go trial to the lever associated with a large reward (two pellets, side counterbalanced across animals but fixed over sessions for each individual), or (4) a Go trial to the other lever associated with a small reward (one pellet). The requirements for successfully completing a Go trial or a No-Go trial were identical to Experiment 1. However, in Experiment 2, reward delivery was delayed for 1s after successful completion of a trial (pressing the correct lever twice on Go trials, or remaining in the nose-poke for at least 1.7–1.9s on No-Go trials). Similarly, the error signal (the house-light being illuminated) was also delayed for 1 s following an erroneous response. A session ended after the animals had either gained 100 rewards or had worked for 60 minutes, whichever came first. Each rat performed this task at least twice while voltammetric recordings were made.

Surgical Procedures

Animals were anaesthetized using isoflurane (4% v/v in O_2 induction and 1.5% for maintenance) and given buprenorphine (Vetergesic, 0.1 ml/kg) at the start of the surgical procedure. Body temperature was maintained at 37 ± 0.5°C with the use of a homeothermic heating blanket. Corneal dehydration was prevented with application of ophthalmic ointment (Lacri-Lube®, Allergan, UK). After induction, the rat was sured in a stereotaxic frame, the scalp was shaved and cleaned with dilute hibiscrub and 70% alcohol, and a local anesthetic, bupivacaine, was applied to the area. The skull was then exposed and holes were drilled for the Ag/AgCl reference electrode (AP: –3.7, ML: –1.4), 4 anchoring screws

(Precision Technology Supplies Ltd, UK) and a voltammetric recording electrode in each hemisphere. After the screws were sured and the reference electrode inserted, custom-made carbon fiber microelectrodes were then either lowered into the NAcc (AP: +1.4, ML: ±1.3, DV: –7.0) and the dorsomedial striatum (AP: +1.2, ML: ±1.9, DV: –4.4: data not presented here) (Experiment 1) or bilaterally into NAcc (Experiment 2). Implanted hemispheres were counterbalanced across animals. The carbon fiber microelectrodes and reference electrode were attached to a 6-pin headstage connector, which was sured in place along with a head post with dental cement. Following surgery, animals were again administered buprenorphine (0.1 ml/kg) and meloxicam (Metacam, 0.2/ml/kg), and given palatable food for consumption. Meloxicam was also administered for at least 3 days following surgery. Animals had on average two weeks of post-surgery recovery with food and water *ad libitum*, prior to food restriction and further behavioral training.

Fast-scan cyclic voltammetry

Fast-scan cyclic voltammetry recordings were performed as described previously[5,18,19]. In brief, voltammetric scans were performed at a frequency of 10 Hz throughout the session. Prior to a scan, the carbon fiber was held at a potential of –0.4 V (vs Ag/AgCl) and then, during the scan, ramped up to +1.3 V and back to –0.4 V at 400 V/s. The application of this waveform causes redox reactions in electrochemically active species, such as dopamine, at the surface of the carbon fiber, which can be recorded as changes in current over time. Based on previously established criteria[5], the recorded current in response to un-cued sucrose pellet delivery, obtained at the start and end of each recording session, was used to determine the chemical sensitivity of the recording electrode to dopamine on that particular session. An extracted cyclic voltammogram was linearly regressed against a dopamine

standard, with $R^2 \geq 0.75$ set as the criterion based on the discriminability of dopamine from other common neurochemicals in a flow cell[20]. Only sessions where sufficient discriminability was confirmed were included in the analysis presented in this article.

For Experiment 2, where animals had electrodes bilaterally targeted to the NAcc, we took data from any NAcc electrode that passed this discriminability test (n=9 electrodes, made up from 3 rats with only 1 electrode that passed the criterion and 3 rats with 2 electrodes, one in each hemisphere, that passed the criterion). However, to enhance the independence of the signals recorded from different hemispheres in the same animal, we recorded two sessions of data for each rat and only analyzed the data from one electrode *per recording session* for each individual. For rats with only one electrode, we used the session with, first, the best behavioral performance, and sond, dopamine that correlated most highly with our dopamine standard.

Data Analysis

Success rate was measured as percentage of correct trials for each trial type. Holding time or response latency was measured as the time from cue onset to nose poke exit and was used as a proxy for movement initiation. Reward time was measured as the time from cue onset to delivery of a sucrose pellet. This corresponds to the end of the holding time (1.7 or 1.9 s) for No-Go trials or the time of sond lever press for Go trials.

Voltammetric analysis was initially carried out using software written in LabVIEW (National Instruments). Data were low-pass filtered at 2 kHz. As has been described previously[21,12], principal component analysis using a standard training set of stimulated dopamine release detected by chronically implanted electrodes was used to isolate changes in dopamine concentration from other electrochemical signals as the first principal

component among other unrelated electrochemical fluctuations such as changes in pH[22, 23]. Trials where the PCA failed to successfully extract dopamine current on >50% of data points in a trial were excluded. Once dopamine-related current changes were extracted all further analysis was undertaken using Matlab® (Mathworks, MA, USA). Unless stated, all data were smoothed using a 0.5 s moving window and baselined by subtracting the average data in a 0.5 s window prior to cue onset from all data points.

The discriminability of dopamine signals, smoothed using a 0.5 s moving window, in the different trial types was analyzed in each individual animal at each time point in a 5 s period after an event of interest (i.e., time-locked to cue onset, head exit or first lever press) using the area under the receiver operating characteristic curve (auROC), an approach from signal detection theory[24]. For the data from Experiment 1, the Go trial type or, in a sond series of analyses, successful trials were considered as positive cases. For the data in Experiment 2, the large reward trial types were always considered as positive cases, except when comparing Go Large against No-Go Large conditions, where the Go trial type was positive.

To quantify which factors affected dopamine levels, regression coefficients were estimated for each animal at each time point in a 6s window spanning from 1s previous to an event of interest (either cue onset or head exit) to 5s following the event, here termed as a "trial". A linear model was used with a constant term, representing an ordinary least-squares fit of the given regressors to the data over trials. For analysis of the cue-evoked dopamine on all correct and valid Go and No-Go trials in Experiment 1 (**Fig. 2e**), the regressors were: (1) action initiation (trial-by-trial holding time between cue and head exit); and (2) success rate (for each trial type); for analysis of the cue-elicited dopamine on Correct v Wrong Go trials

(**Supplementary Fig. 5e**), the regressors were: (1) action initiation and (2) accuracy (correct trials were assigned 1 and incorrect trials were assigned –1); and for analysis of cue-evoked dopamine on all correct and valid Go and No-Go trials in Experiment 2 (**Fig. 3c**), the regressors were: (1) action initiation (trial-by-trial holding time between cue and head exit); and (2) "RPE" (calculated for each given trial type as: [EV – average(EV)], where EV = trial type success rate * reward size). Each trial in each regressor was modeled with a single value. All regressors, whether continuous or categorical, were mean-centered. Regression coefficients in each animal were averaged and then transformed into absolute values (i.e., any negative number was made positive).

Statistics

Behavioral data from the included recording sessions was analyzed using non-parametric statistics: Kruskal-Wallis ANOVA and Wilcoxon Signed Rank Test (though note that all effects remained the same when analyzed using equivalent parametric tests). For Experiment 2, we analyzed the average performance from each rat (n = 6) across their two potential recording sessions, irrespective of whether their electrochemistry data from just one or both sessions were included in the analysis of dopamine signals. Again, the effects remained essentially unchanged if only the analyzed recording sessions were included or if repeated-measures parametric statistics were used.

To analyze the discriminability of dopamine signals recorded in pairs of conditions, the auROC from each animal was averaged and significant discriminability at each time point was determined using 1,000 random permutations of the trial types and re-computing the auROC to generate a null distribution. Permutation tests were considered significant at any time point when $p < 0.05$, corrected for multiple comparisons (i.e. $p < 0.001$). To determine

whether there was a significant negative deflection following a failed No-Go trial, the linear fit in each animal of the dopamine in the 2.5 s post-cue period was also calculated for these trials and compared to correct No-Go trials using a non-parametric Wilcoxon rank sum test.

The significance of the regression coefficients was tested against a population of 1,000 coefficients obtained by randomly permuting the pairings between the regressors and the data. Permutation tests were considered significant at any time point when the regression coefficient from the real data exceeded the maximum or minimum of the permuted population of coefficients ($p < 0.05$, corrected for multiple comparisons over the 5 s after event onset; i.e., $p < 0.001$ uncorrected).

Randomization or blinding was not used during the analysis. No statistical methods were used to pre-determine sample sizes, but our sample sizes are similar to those reported in previous publications[5,8,12].

Actor-critic model

In order to qualitatively compare the observed fluctuations in dopamine release in Experiment 1 with the predictions of the classical reinforcement learning theory, we simulated the standard actor-critic model in this task. The actor-critic model[25], as other reinforcement learning models, describes how animals learn which actions are worth taking in different states. We assumed that the state of the animal is determined by its sensory stimuli (for simplicity, we did not model the animal's ability to estimate time from past events). In particular, in our simulation the state was a combination of: position, auditory cue being presented, and the presence of reward in environment.

For simplicity we considered just 4 positions (relevant to the task): by the left lever (L), by the right lever (R), in the nose-poke (N), and in the food magazine (F). The auditory cues could take 4 values: a cue indicating that pressing the left lever gives reward (L), pressing the right lever gives reward (R), "No-Go" gives reward (N), or no sound (–). Finally, we assumed that the animal recognized 3 types of environments: with pellet in the magazine (Pellet), with house light turned on (Light) (which occurred after an error), and the standard task environment with house light off (–) (present otherwise). This meant, in total, there were 4x4x3=48 possible states, but not all of them occurred in the simulations (e.g. no sound was played unless the environment was neutral).

For simplicity, we assumed that the animal could take 6 actions: move to the left lever (L), move to the right lever (R), move to the nose-poke (N), move to the food magazine (F), press the lever (P) or stay in the current position (–). Some actions were allowed only in certain states, i.e. the animal could only press a lever when it was by a lever, and the animal could not move to its current position (such behavior was described by the stay action (–)).

For simplicity, time was divided into discrete steps of 0.5 s. At each step, the simulated animal performed one of the actions and the new position of the animal was determined by its action. The other components of the state, as well as the reward, were dependent on contingencies of state and action, and additional conditions as shown in **Supplementary Table 1**.

The actor-critic model assumed that animals learned the tendency $Q_{a,s}$ to choose action a in state s (these values are learned by "actor") and the overall values V_s of being in state s (these values are learned by the "critic"). In each time step, after taking each action a at state s, $Q_{a,s}$ and V_s were updated:

$$Q_{a,s} \leftarrow Q_{a,s} + \alpha\delta$$

$$V_s \leftarrow V_s + \alpha\delta$$

In the above equations α was a learning rate constant (which we set to α=0.1) and δ was the reward prediction error that was computed as:

$$\delta = r + \gamma V_{s'} - V_s$$

The reward prediction error (RPE) was equal to the difference between the obtained reward and the expected reward, which was equal to the value V_s of the state in which the action was taken. The obtained reward included two components: an immediate reward *r* (determined according to Table S1) and an expected future reward $V_{s'}$ of the state *s'* to which the simulated animal transitioned, scaled by a discount factor γ (which we set to γ =0.9; setting γ to a value lower than 1 corresponded to an assumption that a reward in the future was worth less than an immediate reward).

At the start of each time step, an action was chosen among the ones available in a given position stochastically, such that the probability of selecting action *a* in state *s* was equal to:

$$P_a = \frac{\exp\left(\beta Q_{a,s}\right)}{\sum_{a' \in Available\,in\,s} \exp\left(\beta Q_{a',s}\right)}$$

According to the above equation the probability of selecting action *a* depends on $Q_{a,s}$, because the larger $Q_{a,s}$, the larger the numerator. The denominator is simply a

normalization term that ensures that all probabilities add up to 1. The parameter β controls how deterministic the choice is (we set β=1).

At the start of each simulation $Q_{a,s}$ and V_s were initialized to 0. The model was "pre-trained" by simulating it for 5,000 s in a simplified version of the task in which pellet delivery was triggered by a single lever press after Go cue or staying in the nose poke for a single time step after No-Go cue. Then the model was trained by simulating the main task for 1,000 s (these numbers were chosen so that the model produced similar accuracy as experimental animals). Finally, the model was simulated for 2,500 s, and the behavioral results, as well as reward prediction errors, were recorded. The whole simulation was repeated 7 times (which corresponded to the 7 animals performing Experiment 1) and the results of the simulations are visualized in **Supplementary Fig. 2**.

Histology

After recordings were completed, animals were deeply anaesthetized with sodium pentobarbitone (200 mg/kg; i.p.), microlesions were made at the electrode locations via current stimulation, and animals were transcardially perfused with saline followed by a 10% formaline solution. Brains were cut into 50 μm-thick coronal stions using a vibrating-blade microtome (Leica, UK). Stions were then mounted on glass slides and stained with cresyl violet to confirm the electrode locations (**Supplementary Fig. 1**).

A **Supplementary Methods Checklist** is available.

References

17. Bari, A. & Robbins, T.W. Inhibition and impulsivity: behavioral and neural basis of response control. *Progress in neurobiology* **108**, 44-79 (2013).
18. Phillips, P.E., Robinson, D.L., Stuber, G.D., Carelli, R.M. & Wightman, R.M. Real-time measurements of phasic changes in extracellular dopamine concentration in freely moving rats by fast-scan cyclic voltammetry. *Methods Mol Med* **79**, 443-464 (2003).
19. Clark, J.J*., et al.* Chronic microsensors for longitudinal, subsond dopamine detection in behaving animals. *Nat Methods* **7**, 126-129 (2010).
20. Heien, M.L., Phillips, P.E., Stuber, G.D., Seipel, A.T. & Wightman, R.M. Overoxidation of carbon-fiber microelectrodes enhances dopamine adsorption and increases sensitivity. *The Analyst* **128**, 1413-1419 (2003).
21. Wanat, M.J., Kuhnen, C.M. & Phillips, P.E. Delays conferred by escalating costs modulate dopamine release to rewards but not their predictors. *J Neurosci* **30**, 12020-12027 (2010).
22. Heien, M.L., Johnson, M.A. & Wightman, R.M. Resolving neurotransmitters detected by fast-scan cyclic voltammetry. *Anal Chem* **76**, 5697-5704 (2004).
23. Heien, M.L*., et al.* Real-time measurement of dopamine fluctuations after cocaine in the brain of behaving rats. *Proc Natl Acad Sci U S A* **102**, 10023-10028 (2005).
24. Green, D.M. & Swets, J.A. *Signal Detection Theory and Psychophysics* (Wiley, New York, 1966).
25. Sutton, R.S. & Barto, A.C. *Reinforcement learning: An introduction* (MIT Press, London, 1998).