Rare rewards amplify dopamine responses

Kathryn M. Rothenhoefer, Tao Hong, Aydin Alikaya and William R. Stauffer

Dopamine prediction error responses are essential components of universal learning mechanisms. However, it is unknown whether individual dopamine neurons reflect the shape of reward distributions. Here, we used symmetrical distributions with differently weighted tails to investigate how the frequency of rewards and reward prediction errors influence dopamine signals. Rare rewards amplified dopamine responses, even when conventional prediction errors were identical, indicating a mechanism for learning the complexities of real-world incentives.

Dopamine neurons generate reward prediction error responses that guide the direction and magnitude of reward learning. These learning signals are approximated by reinforcement learning algorithms, including temporal difference (TD) and Rescorla–Wagner learning models. According to standard TD learning, ‘reward predictions’ are simply point estimates—formally, the temporally discounted sum of future outcomes. The magnitude of these predictions, often determined by the average value of past outcomes, accurately describe the activity of dopamine neurons in well-controlled laboratory settings. However, point estimate predictions reflect neither predicted uncertainty nor the shapes of reward distributions, and they are not adequate descriptors of behavior. Consider that, learning takes longer when rewards are sampled from broader distributions, compared to when they are sampled from narrower distributions. Likewise, decision-makers take longer to choose between options when value differences are small, compared to when differences are large. These results demonstrate that probability distributions of reward values, and not simply point estimates such as the mean, influence behavior. Dopamine responses adapt to the range or standard deviation of predicted outcomes, but it remains unknown if the weights allocated to the tails of reward distributions—a parameter that determines distribution shape and frequency of prediction errors—affects dopamine responses and neural learning rules.

Reinforcement learning has produced remarkable advances in artificial intelligence and reinforcement learning techniques have recently been extended to learning probability distributions. Distributional reinforcement learning models simultaneously learn many different value predictions that, together, represent probability distributions. It was recently shown that a range of value predictions derived from distributional reinforcement learning were reflected by dopamine neurons, raising the enticing possibility that brains use a distributional code for value. Critically, this distributional code operates at the level of populations rather than individual neurons. Thus, it is unknown how single dopamine neurons may adapt their responses to predicted reward probability distributions.

To investigate whether the distribution shape differentially affected reward learning, we created symmetrical reward size distributions that simulated the shapes of uniform and normal distributions (Fig. 1a). Within each block (15–25 trials), monkeys made choices between two never-before-seen cues that predicted normal or uniform reward size distributions, as in Fig. 1a, with pseudorandomly chosen expected values (Fig. 1b, Methods). As expected, the monkeys performed at chance levels on trial 1, but quickly learned to choose the better option (Fig. 1c). Logistic regression of the choice behavior indicated both trial-by-trial learning and better overall performance in the normal blocks ($\beta_{\text{tune}}=0.110$, $P<0.0001$, $\beta_{\text{normal}}=0.167$, $P=0.007$, $N=6098$ trials, t-test). We used a standard reinforcement learning model to quantify the prediction errors generated during learning (Fig. 1d). This analysis revealed that behavior in both block types was characterized by an active learning phase when the prediction error magnitudes were diminishing, and a later asymptotic phase when the magnitudes were stable (Fig. 1e). However, the number of trials in the active learning phase was significantly fewer in the normal blocks compared to the uniform blocks (Fig. 1f, Methods). Moreover, during the active learning phase, pupil diameter responses were more sensitive to rare reward prediction errors than to common reward prediction errors of the same magnitude (Fig. 1g, Methods). This indicates that greater vigilance or arousal was associated with learning from rare-prediction errors. This effect disappeared during the asymptotic phase (Fig. 1h). Together, these data showed enhanced learning performance in blocks with rewards drawn from normal distributions.

We recorded extracellular dopamine neuron action potentials during a passive viewing task (Fig. 2a, Extended Data Fig. 1 and Methods). Here, the magnitudes of the small, medium and large rewards were fixed at 0.2, 0.4 and 0.6 ml, respectively (Fig. 1a). Previous choice testing confirmed that normal and uniform distributions with these reward size elements had equivalent expected utilities (Extended Data Fig. 2, Methods). As expected from cues that predict the same expected utilities, dopamine neurons were similarly activated by the normal and uniform distribution-predicting cues (Fig. 2b). Thus, the passive viewing task rigorously controlled the magnitudes of conventional prediction errors, defined as received minus predicted reward values.

At the time of reward delivery, dopamine responses were amplified by rare prediction errors. We used two different randomization schemes to control for the number of times each distribution was presented (conditioned stimulus matched), or the number of times each prediction error was experienced (prediction error matched) (Extended Data Fig. 3). Under both randomization schemes, the 0.6 ml reward activated a larger dopamine response in normal distribution trials, compared to dopamine activations following delivery of the same volume reward in uniform distribution trials (Fig. 2c,d, solid lines). Likewise, dopamine responses were more strongly suppressed by delivery of 0.2 ml reward during normal distribution trials, compared to delivery of the same reward during the active learning phase when the prediction error magnitudes were diminishing, and a later asymptotic phase when the magnitudes were stable. Together, these data showed enhanced learning performance in blocks with rewards drawn from normal distributions.
uniform distribution trials (Fig. 2c,d, dashed lines). Linear regression revealed that 34 neurons were significant for reward size, and that most of the neurons (29/40) had steeper slopes for the normal condition, compared to the uniform condition (Fig. 2e,f). Thus, rare prediction errors resulted in bidirectional amplification of the responses, compared to common prediction errors of the same magnitude. We applied a naïve Bayes classifier to 11 neurons with the greatest selectivity for rare rewards (Methods). The classifier was able to decode distribution identity from the responses to 0.2 and 0.6 ml, but failed to decode the distribution from the responses to 0.4 ml (Fig. 2g). Together, these results demonstrate that phasic dopamine responses reflect predicted probability distributions.

Finally, we investigated whether reversal-point variability reflected the predicted distributions. We categorized responses as activations or suppressions and calculated the reversal points for each neuron in each distribution (Methods). As predicted by the distributional TD model, the uniform distribution evoked a larger spread of reversal points compared to the normal distribution (Fig. 3a). We subtracted cell- and distribution-specific reversal points from each cell’s average responses to the three different rewards and tested whether the differential reversal points accounted for the bidirectional response amplification. Following reversal-point correction, we still observed significantly amplified responses, in both the negative and positive domain, to identical rewards drawn from the normal compared to the uniform distribution (Fig. 3b and Extended Data Fig. 4), but no significant difference in the reversal-point-corrected responses to 0.4 ml. These results demonstrate that the bidirectional amplification of responses is not accounted for by the reversal points. Moreover, these results hint that the single cell-level amplification of responses and the population level distributional TD model could be complementary schemes for learning the shapes of probability distributions.

Here we show that dopamine reward prediction error responses are amplified by rare rewards. Amplified dopamine responses were evident even when identically sized rare and common rewards generated identical TD prediction errors. This result demonstrates that dopamine responses are sensitive to the shapes of predicted probability distribution, rather than just the predicted mean. These findings indicate a new model for phasic dopamine responses and reward learning that is distinct from, but complementary to, conventional reward prediction error updating.
Several lines of evidence indicate that the amplifications of dopamine responses were not explained by differences in conventional prediction errors. Behavioral assays showed that the monkeys assigned similar expected utility values to both distributions (Extended Data Fig. 2). Expected utility is a proxy for the ‘predicted reward value’ term used to describe dopamine reward prediction error responses\(^1\). Accordingly, dopamine responses to normal and uniform distribution-predicting cues were indistinguishable (Fig. 2b). Therefore, the amplified dopamine responses we observed here were not explained by differences in conventionally defined prediction errors. Rather, the dynamic ranges of the neurons adapted to the shapes of the predicted probability distributions (Fig. 2c–f).

Biological learning signals have inspired deep reinforcement learning algorithms with performance that exceeds expert human performance on Atari games, chess and Go\(^3,4\). Recently, a new machine learning model, distributional TD, was applied to study the activity of dopamine neurons\(^5\). A fundamental distinction between distributional TD and the results we present here is the scale at which outcome distributions are represented. In distributional TD, the probability distribution is represented at the level of dopamine neuron populations. In contrast, our results show that single dopamine neurons are sensitive to the shape of the probability distribution (Fig. 2c–f). Our data indicate that two mechanisms, one operating at the level of populations and the other at the level of single neurons, are complementary schemes for learning probability distributions. Indeed, our data confirmed one prediction from the distributional TD model: for the same population of dopamine neurons, the spread of the measured reversal points is larger for uniform, when compared to normal predicted reward distributions (Fig. 3a). Nevertheless, even after accounting for the distribution-sensitive reversal points, we still observe bidirectional amplification of dopamine responses to rare rewards (Fig. 3b). These results reveal complementary learning schemes within the same population of dopamine neurons.

At the level of single neurons, the amplified dopamine responses to rare rewards indicate that reinforcement learning models that acquire only point estimate predictions are not adequate to describe dopamine activity. Rather, these data indicate that reinforcement learning algorithms that track uncertainty, such as Kalman TD\(^4\), may provide an appropriate conceptual framework to explain information processing in the reward system. Kalman-like reinforcement signals enable reward prediction and estimation of uncertainty\(^5\), and therefore may be critical for implementing Bayesian inference. In this sense, the observed amplification of dopamine responses by rare rewards is consistent with a signal that could guide Bayesian inference of the most likely outcomes. Nevertheless, future studies will be required to understand whether phasic dopamine responses can support explicit Bayesian inference for optimal economic choices.

---

**Fig. 2 | Rare rewards amplified dopamine reward prediction error responses.** a. In the recording task, the monkeys viewed a distribution-predicting conditioned stimulus (CS) and rewards were delivered 2s later. b. PSTH of CS-evoked responses to the normal and uniform distribution-predicting cues in a single neuron. There was no significant difference between the response magnitudes (\(P=0.69, N=40\) neurons, Wilcoxon rank-sum). Shaded regions represent \(\pm\)s.e.m. across trials. c. Single-neuron reward responses to rewards during normal and uniform trials, recorded using the CS-matched randomization scheme (Extended Data Fig. 3). Top: PSTHs show impulse rate as a function of time. Solid lines show responses to 0.6 ml, whereas dashed lines show responses to 0.2 ml of juice. Shaded regions represent \(\pm\)s.e.m. across trials. Bottom: raster plots that are separated by normal and uniform trials and by reward sizes. Every tick mark represents the time of an action potential and every row represents a trial. Black vertical dashed line indicates the time of reward. d. As in c, for a neuron recorded using the prediction error-matched randomization scheme (Extended Data Fig. 3). e. Single-neuron linear regression of an individual neuron (c) showed steeper response slopes to rare rewards drawn from the normal distributions. Solid lines indicate the fitted slopes in normal and uniform distribution trials. Dots represent the average neural response rewards in normal and uniform distribution trials. Error bars represent \(\pm\)s.e.m. across trials (all data points had between 13 and 76 trials). f. Scatter plot of normal and uniform distribution response slopes from every neuron (\(P=0.003, N=40\) neurons, \(t=3.19, t\)-test). Inset: the histogram shows the density of the dots relative to the diagonal unity line. g. Confusion matrices of distribution identity decoding from neuronal responses to 0.2, 0.4 and 0.6 ml rewards in the normal and uniform distributions. The matrix sectors are shaded according to the proportion of trials decoded as normal (N) and uniform (U). The scale bar on the right shows that darker shades indicate higher proportions. Black asterisks indicate decoding performance above chance level for the responses to 0.2 and 0.6 ml (\(P=0.045\) and 0.028, \(N=11\) neurons, permutation test, uncorrected \(P\) values). No asterisk above the 0.4 responses indicates no significant decoding (\(P=0.642, N=11\) neurons, permutation test).
The behavioral model enabled us to directly measure learning differences; however, it required models to do post hoc estimation of the underlying reward prediction errors. This dependency on model-derived estimates constrained our ability to control the magnitudes of reward prediction errors. Therefore, we used a passive viewing task to control prediction errors during neuronal recordings (Fig. 2a). This strategy of measuring behavior in one version of the task and doing neural recordings in a simplified version of the task has been used many times previously by ourselves and others. However, the experimental separation of the behavioral measurement from the neural recordings prevents us from drawing firm conclusions regarding the role of dopamine signal amplification in learning. Future studies that combine complex behavior and neural recording in the same task will be critical for determining the trial-by-trial relationship between dopamine response amplification and behavior.

It is tantalizing to speculate about the possibility that the neural circuits responsible for value processing evolved in a world where the normal distribution makes frequent appearances—and that this evolutionary history makes it easier for individuals (and their dopamine neurons) to learn normal statistics. Regardless, the amplified dopamine responses coupled with the faster learning dynamics observed here suggest that the magnitude of dopamine release may affect cellular learning mechanisms in the striatum. Moreover, dopamine release in the prefrontal cortex is tightly linked to working memory signals and performance. These findings raise the possibility that amplified dopamine responses could contribute to the exaggerated salience of rare events and postulate a neural mechanism to explain aberrant learning behaviors associated with debilitating mental health disorders such as psychosis, schizophrenia and depression.

Online content
Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41593-021-00807-7.

Received: 19 December 2019; Accepted: 20 January 2021; Published online: 8 March 2021

References
1. Stauffer, W. R. The biological and behavioral computations that influence dopamine responses. *Curr. Opin. Neurobiol.* 49, 123–131 (2018).
2. Enomoto, K. et al. Dopamine neurons learn to encode the long-term value of multiple future rewards. *Proc. Natl Acad. Sci. USA* 108, 15462–15467 (2011).
3. Sutton, R. & Barto, A. Reinforcement Learning: An Introduction (MIT Press, 1998).
4. d’Acremont, M. & Bossaerts, P. Neural mechanisms behind identification of leptokurtic noise and adaptive behavioral response. *Cereb. Cortex* 26, 1818–1830 (2016).
5. Diederien, K. M. J. & Schultz, W. Scaling prediction errors to reward variability benefits error-driven learning in humans. *J. Neurophysiol.* 114, 1628–1640 (2015).
6. Nassar, M. R., Wilson, R. C., Heaps, B. & Gold, J. I. An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *J. Neurosci.* 30, 12366–12378 (2010).
7. Krajbich, I., Armel, C. & Rangel, A. Visual fixations and the computation and comparison of value in simple choice. *Nat. Neurosci.* 13, 1292–1298 (2010).
8. Tobler, P. N., Fiorillo, C. D. & Schultz, W. Adaptive coding of reward value by dopamine neurons. *Science* 307, 1642–1645 (2005).
9. Minn, V. et al. Human-level control through deep reinforcement learning. *Nature* 518, 529–533 (2015).
10. Silver, D. et al. Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 484–489 (2016).
11. Bellemare, M. G., Dabney, W. & Munos, R. A distributional perspective on reinforcement learning. *Proc. Mach. Learn. Res.* 70, 449–458 (2017).
12. Dabney, W. et al. A distributional code for value in dopamine-based reinforcement learning. *Nature* 577, 671–675 (2020).

Fig. 3 | Dopamine pseudo-populations and single neurons simultaneously reflect predicted probability distributions. a, Box and whisker plots show the spread of reversal points for the population of neurons in normal (purple) and uniform (green) trials ((0.0065, 0.0129) and (0.0133, 0.0221), N = 40 neurons, bootstrap 90% confidence interval for standard deviation). b, Box and whisker plots show the baseline subtracted responses to 0.2 and 0.6 ml of juice. ** P < 0.0001, N = 29 neurons, Wilcoxon signed-rank test, Bonferroni corrected. Responses to 0.4 ml were not significantly different and so not shown (P = 0.226, N = 29 neurons, Wilcoxon signed-rank test). Box and whisker plots show the median (line), quartiles (boxes), range (whiskers) and outliers (+).
13. Stauffer, W. R., Lak, A. & Schultz, W. Dopamine reward prediction error responses reflect marginal utility. *Curr. Biol.* 24, 2491–2500 (2014).

14. Gershman, S. J. A unifying probabilistic view of associative learning. *PLoS Comput. Biol.* 11, e1004567 (2015).

15. Babayan, B. M., Uchida, N. & Gershman, S. J. Belief state representation in the dopamine system. *Nat. Commun.* 9, 1891 (2018).

16. Lak, A., Stauffer, W. R. & Schultz, W. Dopamine neurons learn relative chosen value from probabilistic rewards. *eLife* 5, e18044 (2016).

17. Morrens, J., Aydin, Ç., Janse van Rensburg, A., Esquivelzeta Rabell, J. & Haesler, S. Cue-evoked dopamine promotes conditioned responding during learning. *Neuron* 106, 142–153.e147 (2020).

18. Preuschoff, K., Marius't Hart, B. & Einhauser, W. Pupil dilation signals surprise: evidence for noradrenaline’s role in decision making. *Front. Neurosci.* 5, 115 (2011).

19. Schultz, W. Neuronal reward and decision signals: from theories to data. *Physiol. Rev.* 95, 853–951 (2015).

20. Vijayraghavan, S., Wang, M., Birnbaum, S. G., Williams, G. V. & Arnsten, A. F. Inverted-U dopamine D1 receptor actions on prefrontal neurons engaged in working memory. *Nat. Neurosci.* 10, 376–384 (2007).

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
© The Author(s), under exclusive licence to Springer Nature America, Inc. 2021
Methods

Animals, surgery and setup. All animal procedures were approved by Institutional Animal Care and Use Committee of the University of Pittsburgh. We used two male Rhesus macaque monkeys (Macaca mulatta) for these studies (both 6 years of age, 9 and 11 kg). A titanium head holder (Gray Matter Research) and a recording chamber (Crist Instruments, custom made) were aseptically implanted under general anesthesia before the experiment (Extended Data Fig. 1e,f).

The recording chamber for vertical electrode entry was centered 8 mm anterior to the interaural line. During experiments, monkeys sat in a primate chair (Crist Instruments) positioned 30 cm from a computer monitor. During behavioral training, testing and neuronal recording, eye position was monitored noninvasively using infrared eye tracking (Eyelink Plus 1000). Licking was monitored with an infrared optical sensor positioned in front of the juice spout (Balluff). Eye, lick and digital task event signals were sampled at 2 kHz. Custom-made software (MATLAB, Mathworks Inc.) running on a Microsoft Windows 7 computer controlled the behavioral tasks.

Behavioral tasks. Pavlovian task for neural recordings. Two visually distinct cues (fractal images) were used to predict reward. One cue predicted a uniform distribution, where 0.2, 0.4 and 0.6 ml were delivered with equal probability (1/3) for each reward. A second cue predicted a normal reward distribution, where 0.2 and 0.6 ml were delivered with low frequency (2/15 probability for each of the two rewards) and the middle reward (0.4 ml) was delivered with a much higher frequency (11/15 probability). Finally, there was an unpredicted reward condition, where 0.4 ml of juice was delivered with no preceding cue.

We used two different randomization schemes, one where there were equal instances of normal and uniform trials (conditioned stimulus matched), and one where there were equal instances of nonzero prediction errors for both normal and uniform (prediction error matched) (Extended Data Fig. 3). In each trial, the situation was pseudorandomly chosen with replacement, according to the randomization scheme. The cue-reward interval was always 2 s. Trials were separated with intertrial intervals of 2–5 s, chosen from a truncated exponential distribution. Before recording, all cues were well learned after experiencing them repeatedly over multiple sessions (monkey B, ten sessions, roughly 2,800 trials; monkey S, six sessions, roughly 2,600 trials).

Choice tasks for measuring distribution values. For the data presented in Extended Data Fig. 2b–d, three cues predicted a normal distribution (Fig. 1a, right), and three different cues predicted a uniform distribution (Fig. 1a, left). One ‘safe’ cue predicted 0.2 ml of juice and a different safe cue predicted 0.6 ml of juice. Monkey S was offered binary choices between normal and uniform distribution-predicting cues, and between distribution-predicting cues and safe cues. Following successful central fixation for 0.5 s, two choice options appeared on the monitor and the monkey indicated its choice by a saccade toward one of the cues. The monkey was allowed to saccade as soon as it wanted. The monkey had to keep its gaze on the chosen cue for 0.5 s to confirm its choice. Reward was delivered 1.5 s later. Trials were separated with intertrial interval of 1.5–6.5 s, drawn from a truncated exponential distribution. Failure to maintain the central fixation or early break of the fixation on the chosen option resulted in a 4 s time-out, and a repeat of the failed trial.

For the data presented in Extended Data Fig. 2e–g, monkeys made choices between well-learned distribution-predicting fractal cues and ‘safe’ value bar cues that indicated the magnitude of the alternative option. The value bar cues had a value range of 0 to 0.8 ml, in 0.1 ml increments. Wherever the horizontal bar intersected the vertical scale indicated with 100% certainty the size of juice the monkeys would receive if they chose it. The mean of the distribution-predicting cues was 0.4 ml. In each choice trial, after successful central fixation for 0.5 s, the two choice options appeared on the monitor and the monkey indicated its choice by a saccade toward one of the cues. The monkey was allowed to saccade as soon as it wanted. The monkey had to keep its gaze on the chosen cue for 0.5 s to confirm its choice. Reward was delivered 1.5 s later. Trials were separated with intertrial interval of 1.5–6.5 s, drawn from a truncated exponential distribution. Failure to maintain the central fixation or early break of the fixation on the chosen option resulted in a 4 s time-out, and a repeat of the failed trial.

Choice task to measure learning. For the data presented in Fig. 1b–g, monkeys were offered two never-before-seen cues on the first trial of every block. The block length was selected from a truncated exponential distribution between 15 to 25. Within each block both the cues predicted rewards drawn from the same type of distribution, normal or uniform. Further, each new cue had a different pseudorandomly selected mean that was 0.2, 0.3, 0.4, 0.5 or 0.6 ml. For example, if it were a uniform block, and the means selected for the two cues were 0.3 and 0.6 ml, the rewards for one cue would be 0.2, 0.3 and 0.4 ml (drawn with equal frequency) and 0.5, 0.6 and 0.7 ml (also drawn with equal frequency). In each choice trial, after successful central fixation for 0.5 s, the two choice options appeared on the monitor and the monkey indicated its choice by a saccade toward one of the cues. The monkey was allowed to saccade as soon as it wanted. The monkey had to keep its gaze on the chosen cue for 0.5 s to confirm its choice. Reward was delivered 1.5 s later. Trials were separated with intertrial interval of 1.5–6.5 s, drawn from a truncated exponential distribution. Failure to maintain the central fixation or early break of the fixation on the chosen option resulted in a 4 s time-out, and a repeat of the failed trial.

Choice task for measuring the subjective value of reward size distributions. The overall goal of this study was to investigate how predicted distribution shape influenced dopamine responses. To fairly investigate this, we required that the predicted distribution values be the same. Accordingly, we created the uniform and normal reward size distributions such that they were composed of the same three elements and had the same expected values (Fig. 1a). However, dopamine neurons reflect subjective values, so we used two choice tasks to verify that the expected utilities of the distributions were the same (Extended Data Fig. 2).

We first used a direct choice task to measure the relative subjective values of the distributions. Visual cues (fractal images) were used to predict rewards. To avoid preferences between cues, we predicted different values for different cues. Three different cues predicted the uniform distribution (Extended Data Fig. 2a). To ensure that the monkey was making valid economic choices rather than choosing randomly, we also created two safe cues that predicted a small (0.2 ml) and large (0.6 ml) reward. We reasoned that subjects making valid economic choices should choose the large reward option over both distributions, and both distributions over the small reward option. We used classical conditioning to train monkeys on the cue-reward contingencies, then we measured binary choices between the cues (Extended Data Fig. 2b). The monkey selected the normal cue over the uniform cue with a probability of 0.53 ± 0.19; this was not significantly different from chance (Extended Data Fig. 2c) (P = 0.48, N = 9 cue pairs, t-test). Additionally, the monkey chose the normal distribution over the small reward (Extended Data Fig. 2c, P < 0.0001, t-test) and the large reward over the distribution (Extended Data Fig. 2c, P = 0.0004, t-test). Similarly, the monkey chose the uniform distribution over the small reward, and the large reward over the distribution (Extended Data Fig. 2d, P = 0.001 and 0.005, respectively N = 3 cue pairs, t-test) Thus, while making valid economic choices, the monkey was choice indifferent between the distributions. These results provide strong evidence that the predicted values of the two distributions were the same.

The expected utilities were critical to our interpretation of the data, and as such, we replicated this result using a different behavioral model: we independently measured the certainty equivalents of normal and uniform reward distributions. Certainty equivalents are the values of rewards the subject would exchange for a gamble; in these experiments, the distributions were the gambles. Monkeys made choices between cues that predicted a distribution and cues that explicitly indicated safe options (Extended Data Fig. 2e, Methods). We plotted the probability of choosing the safe option as a function of the safe option volume and generated psychometric functions (Extended Data Fig. 2f,g). The certainty equivalents were the safe values that corresponded to P(choose safe) = 0.5 (black arrows in Extended Data Fig. 2g). Analysis of the session-by-session certainty equivalents for the normal and uniform blocks found no effect of the distribution type on the certainty equivalents (P = 0.2, N = 18; t-test). Therefore, the certainty equivalents strongly agree with the direct choice data indicate that the normal and uniform reward size distributions had similar subjective values. These results indicated that the prediction errors generated from the distributions could be readily compared and ensured that disparities between prediction error responses were not driven by differences in the predicted subjective values.

Analysis of behavioral data. Logistic regression. We used logistic regression to quantify the influence of reward distribution on monkeys’ behaviors, controlling for trial numbers since a new block starts and the difference between the values of two cues.

\[
\log \left( \frac{P(\text{Correct})}{1 - P(\text{Correct})} \right) = \beta_0 + \beta_2 D + \beta_3 C + \beta_4 T
\]

where \( D \) is a binary variable for reward distribution type (0 for uniform and 1 for normal), \( C \) is a continuous variable for the difference between the values of two cues and \( T \) is a categorical variable for the trial number since the start of a new block.

Reinforcement learning model. We used a fixed learning-rate reinforcement learning model to examine monkeys’ choices during learning and to acquire trial-by-trial estimate of chosen and unchosen values. The model had two value functions representing the learned values of probability distribution 1 (pd1) and probability distribution 2 (pd2), respectively. In each trial \( t \), the probability that the model chooses pd1 over pd2 was estimated by the softmax rule as follows:

\[
P(\text{pd1}) = \frac{e^{V_{\text{pd1}}(t)} / T}{e^{V_{\text{pd1}}(t)} / T + e^{V_{\text{pd2}}(t)} / T}
\]

where \( \beta_t \) is the temperature parameter of the softmax rule, determines the level of choice randomness.

In each trial, on making a choice and receiving an outcome, the value of the chosen option on that trial, \( V_t \), was updated according the reward prediction error, as follows:
where $\alpha$ denotes the learning rate, and the prediction error is calculated as the following: $\delta t = r - V$, indicates the difference between the predicted and realized reward sizes, $V$, and $r$, respectively. The free parameters, $\alpha$ and $\beta$, were fit by maximizing the likelihood of the model. After fitting the model, we took the trial-wise mean of the unsigned prediction error over blocks of the same type (Fig. 1a).

To characterize the transition from active learning to asymptotic behavior, we fit logarithmic functions to each block, and the collected the block by block transition trials that marked the crossing of a predetermined threshold that separated active learning from asymptotic behavior. When the first derivative of the fitted prediction errors decreased below a predetermined threshold, we considered that the animal had stopped actively learning. When the magnitude of the prediction errors stayed below 0.1 for more than two trials, we considered that the animal successfully estimated the true value, since the true difference between the lowest/highest values from the mean was 0.1 ml. We designated the boundary between active learning and asymptotic phases as the trial when both conditions were met. The fast transitions exhibited in the normal distribution block was robust under a wide range of prescribed thresholds.

Deconvolution. Event-related pupil responses were analyzed trial-by-trial using nidecove (23), a Python package that specializes in fMRI and pupil signal deconvolution. The design matrix for a trial consisted of a total of four event types: the onset of central dot for fixation, the onset of cue presentation, the monkeys’ saccades to indicate choice, offset of cue presentation (in temporal order), and the onset of reward. The pupil diameter changes related to fixation and the offset of cue presentation were analyzed 0.5 s pre-event until 2 s post-event; the time windows for the onset of cue presentation and monkeys’ saccades started 0.5 s pre-event and ended 0.5 s post-event and the time window for the presence of reward started at 0.5 s pre-event and ended at 1.5 s postevent. To understand the relationship between pupil diameter and prediction error postward, reward prediction errors and value estimates derived from the model were used as covariates in the deconvolution algorithm. Consequently, we obtained a measure of how sensitive the postreward pupil diameter changes to the prediction errors in each reward distribution, by looking at the beta coefficients in the prescribed time window. Finally, we grouped the deconvolved signal based on the active/asymptotic learning period distinction and created the ensemble average across trials.

Neural data acquisition. Custom-made, movable, glass-insulated, platinum-plated tungsten microelectrodes were positioned inside a stainless-steel guide cannula and advanced by an oil-driven micromanipulator (Narishige). Action potentials from single neurons were amplified, filtered (bandpass 100 Hz to 3 kHz), and converted into digital pulses when passing an adjustable time-amplitude threshold (Bak Electronics). We stored both analog and digitized data on a computer using custom-made data collection software (MATLAB).

Dopamine neurons were functionally localized with respect to (1) the trigeminal somatosensory thalamus explored in awake monkeys (very small perioral and intraoral receptive fields, high proportion of tonic responses, 2–3 mm dorsoventral extent), (2) tonically active position coding ocular motor neurons and (3) phasically direction coding ocular premotor neurons in awake monkeys. Individual dopamine neurons were identified using established criteria of long waveform (>2.5 ms, Extended Data Fig. 1a) and low baseline firing (<8 impulses per s). Following standard sample sizes used in studies investigating neuronal responses in nonhuman primates, we recorded extracellular activity from 67 dopamine neurons. Forty neurons had a sufficient number of trials and we used these neurons for further analysis.

The neurons that met these criteria showed the typical phasic activation after unexpected reward (Extended Data Fig. 1b, P < 0.0001, N = 40 neurons; Wilcoxon rank-sum test). Extended Data Fig. 1c,d show maps of our recording locations relative to both monkeys’ grids, and the number of cells recorded at each location. Extended Data Fig. 1e,f show MRI images of monkey 5 and the location of the recordings.

Analysis of neural data. Data preprocessing. We constructed peri-stimulus time histograms (PSTHs) by aligning the neuronal impulses to task events and then averaging across multiple trials. We smoothed the PSTHs by convolving with $(1 - e^{-t/T})$, where $T$ is set to be 20 ms. The analysis of neuronal data used defined time windows, individual to each neuron, that included the major positive and negative response components following cue onset and juice delivery, as detailed for each analysis and each figure caption. The neural activity within time window following juice delivery was baseline-corrected by subtracting the average activity from $-1,000$ to 0 ms relative to cue onset.

Single-neuron linear regression. To determine whether previous rewards influence the current conditioned stimulus response, we fit a linear model to each neuron’s conditioned stimulus response, using the rewards from the previous five trials as the independent variables. We found that previous outcomes up to five trials back did not influence the conditioned stimulus response. This result is not particularly surprising in the normal distribution trials, as the previous five outcomes were most often 0.4 ml. This reward magnitude evoked no reward prediction errors. The uniform distribution, on the other hand, did generate more prediction errors. The lack of a clear learning effect in the uniform distribution has two main causes, we think. First, trial types were determined at random (Extended Data Fig. 3). Thus, the previous uniform trial could be several trials back. Second, the monkeys had experienced the cues so often that the learning rate was likely very low.

To assess if reward responses for an individual neuron were enhanced bidirectionally by rare prediction errors, we fit the following linear model to each neuron:

$$F_i = \beta_0 + \beta_1 \times D + \beta_2 \times R + \beta_3 \times (D + R)$$

where $F_i$ is the normalized firing rate in the time window following juice delivery, $D$ is a binary variable for reward distribution type (normal distribution as reference group), $R$ is a continuous variable for reward magnitude and $D \times R$ represents the interaction effect between reward distribution and reward magnitude. Figure 2f was obtained by scatter plotting each neuron’s slope for the normal distribution against its slope for the uniform distribution. A paired t-test was used to see if the slopes were significantly biased toward normal distribution.

Decoding distribution type. For each of the three reward magnitudes, we used a Gaussian naïve Bayes classifier to decode the normal and uniform reward distributions from the average firing rate in the time window following juice delivery (24). We then used leave-one-out cross-validation to assess the performance of the decoder. The resulting confusion matrix was normalized by the number of trials. After cross-validation, permutation tests with 5,000 iterations were performed to see if the accuracy of the decoder is significantly different from chance for each reward magnitude. A decoder including all 40 neurons was not able to correctly classify the reward types above chance; therefore, we used a selectivity index (SI) to select neurons for decoding. The single-neuron selectivity index for a particular reward magnitude was defined as the difference between mean reward responses in two reward distribution, divided by the pooled variance of two conditions.

$$SI = \frac{\bar{V}_{\text{reward}} - \bar{V}_{\text{neutral}}}{\sigma_{V}}$$

The subset of 11 neurons with the largest selectivity index successfully decoded the predicted distribution from the responses to 0.2 and 0.6 ml (Fig. 2g). To ensure that the rest of the neurons did not encode an opposite effect, we built a classifier with the rest of the neurons (29/40) and did not observe above-chance performance ($P = 0.515, P = 0.329, P = 0.549$, for 0.2, 0.4 and 0.6 ml, respectively, permutation test).

Reversal-point correction. To account for variability of reversal point reported in the literature (25), we corrected the reward response of each neuron by subtracting the estimated reward response of its reversal point. We estimated neuron- and distribution-specific reversal point by splitting the distribution of responses for each neuron in each distribution, into two groups. One group contained the trials with activations, and the other group contained the trials with suppressions. We then averaged the reward sizes that were associated with the responses in the two groups, and the reversal points were obtained by taking the mean of the two averages (Fig. 3a). The neural activity corresponding to the reversal point was estimated by plugging the reversal point into the single-neuron linear regression described above. For each neuron in each distribution, we subtracted this estimated activity from the responses to 0.2, 0.4 and 0.6 ml. We used a two-tailed Wilcoxon signed-rank test to test whether neurons with steeper response slopes to rewards from normal distributions show bidirectional stretch in their reward responses, after reversal-point correction (Fig. 3b).

Statistics and reproducibility. All statistical analyses were performed and all graphs were created in Python v.3.7.2, MATLAB R2019b and Python package nideco. No statistical methods were used to predetermined sample sizes, but our sample sizes are similar to those reported in studies investigating neuronal and behavioral responses in nonhuman primates (26). For data collected in the choice experiment, based on the metrics we adopted, only choices in blocks with length equal to 15 were included in the statistical tests to avoid skewing the results (Fig. 1c,e,f). For neural data analysis, we recorded extracellular activity from 67 dopamine neurons in two monkeys, and 40 neurons had a sufficient number of trials for further analyses. 27/60 of the dopamine neurons were excluded due to an insufficient number of trials in all of the trial types used ($<7$). Effects were considered significant at $P < 0.05$. Statistical details for each analysis (for example, $N$ and $P$) are specified in the respective part of the text. Data distribution was assumed to be normal but was not formally tested in parametric tests (for example, $t$-test). Data collection and analysis were not performed blind to the conditions of the experiments.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data that support the findings of this study are available from the corresponding author upon request.

Code availability

The code used to analyze these data are available from the corresponding author upon request.
References
21. Van Slooten, J. C., Jahfari, S., Knapen, T. & Theeuwes, J. How pupil responses track value-based decision-making during and after reinforcement learning. *PLoS Comput. Biol.* **14**, e1006632 (2018).
22. de Hollander, G. & Knapen, T. nideconv v0.1.0 (2017); https://nideconv.readthedocs.io/en/latest/
23. Loe, P. R., Whitsel, B. L., Dreyer, D. A. & Metz, C. B. Body representation in ventrobasal thalamus of macaque: a single-unit analysis. *J. Neurophysiol.* **40**, 1339–1355 (1977).
24. Guyenet, P. G. & Aghajanian, G. K. Antidromic identification of dopaminergic and other output neurons of the rat substantia nigra. *Brain Res.* **150**, 69–84 (1978).
25. Batista, A. P. et al. Cortical neural prosthesis performance improves when eye position is monitored. *IEEE Trans. Neural Syst. Rehabil. Eng.* **16**, 24–31 (2008).

Acknowledgements
We thank A. Bostan for comments and discussion, and J. Breter for animal care and enrichment. This work was supported by the National Institutes of Health (NIMH) grant DP2MH113095 to W.R.S.

Author contributions
K.M.R. and W.R.S. designed the experiment. K.M.R., A.A. and W.R.S. collected data. K.M.R., T.H. and W.R.S. analyzed the data and wrote the paper.

Competing interests
The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41593-021-00807-7.
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41593-021-00807-7.
Correspondence and requests for materials should be addressed to W.R.S.

Peer review information *Nature Neuroscience* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.
Extended Data Fig. 1 | Dopamine neurons and recording sites. a, Example dopamine waveform from one of the neurons in our population. b, The population of 40 neurons used for our analyses in the Pavlovian and choice task had significant activations following unpredicted rewards – a characteristic feature of dopamine neurons. Gray bar along the x axis indicate the response window used for analysis. c, Recording locations for the left hemisphere of monkey S. X axis indicates lateral to medial location in the grid in millimeters, relative to midline (0). Right y axis indicates posterior to anterior location in the grid in millimeters, relative to interaural line (IAL). Each locations' color indicates the number of neurons recorded for that location. Black circles surrounding the individual locations indicated that neurons recorded there were part of the population of 29 neurons that had a steeper response slopes in normal compared to uniform condition. Bar graphs on the left and top axes indicate the proportion of cells in that AP (left) or ML (top) location that were effect positive. Yellow dot corresponds to location indicated in MRI scan shown in d and e. d, Recording locations for the left hemisphere of monkey B. Same as panel c. e, Sagittal view MRI of the recording chamber of monkey S. Purple arrow indicates the AP location in the grid (+12 mm from IAL). f, Coronal view MRI of the recording chamber of monkey S. Purple arrow indicates the ML location in the grid (1 mm from Midline). Yellow dot in e and f correspond to approximate recording grid location in c.
Extended Data Fig. 2 | Normal and Uniform reward size distributions have equivalent subjective values. a, Schematic of the distribution-predicting fractal cues used to represent Normal (N) and Uniform (U) distributions, and safe values for the choice task in b. Three unique cues were used to predict a Normal distribution of rewards, and three unique cues were used to predict a Uniform distribution of rewards. All the distribution predicting cues were comprised of the same three reward volumes (0.2, 0.4, and 0.6 ml), and thus the same expected value (EV) of 0.4 ml. Additionally, one fractal cue predicted a sure reward of 0.2 ml, and another fractal cue predicted a sure reward of 0.6 ml. b, Monkeys made saccade-guided choices between Normal distribution-predicting cues, Uniform distribution-predicting cues, and safe rewards. c, Bar graphs are the probability of choosing the alternate cue over a Uniform distribution-predicting cue with an EV of 0.4 ml. The alternates from left to right on the x axis are a safe cue predicting 0.2 ml, a Normal distribution-predicting cue with a mean of 0.4 ml, and a safe cue predicting 0.6 ml. Data points are from individual blocks, and error bars represent ±SEM across blocks (between 6 and 18 blocks per condition). d, Same as in c, but the probability of choosing an alternate cue over a Normal distribution-predicting cue with an EV of 0.4 ml, and the middle alternate option represents Uniform distribution-predicting cues with an EV of 0.4 ml. e, The choice task used to measure subjective value. Animals made saccade-directed choices between a distribution predicting cue and a safe alternative option. The safe alternative option was a value bar with a minimum and maximum of 0 and 0.8 ml at the bottom and top, respectively. The intersection between the horizontal bar and the scale indicated the volume of juice that would be received if monkeys selected the safe cue. f, Probability of choosing the safe cue as a function of the value of the safe option, when the distribution predicting cue had an expected value (EV) of 0.4 ml. Dots show average choice probability for 9 safe value options for monkey B. Solid lines are a logistic fit to the data. Red indicates data from normal distribution blocks, gray indicates data from uniform distribution blocks. The dashed horizontal lines indicate subjective equivalence, and the CE for each distribution type is indicated with the dashed vertical lines. g, Same as in f, for monkey S.
Extended Data Fig. 3 | Reward randomization schemes used to determine trial types. **Top**, “CS matched” randomization with equal frequencies of Normal and Uniform trials. **Bottom**, “PE matched” randomization with equal frequencies of 0.2 ml and 0.6 ml reward trials in each distribution. In both graphs, the y axis represents the probability of drawing the trial type (trial types drawn with replacement). The 6 trial types divided according to distribution type (N and U) and reward size (0.2, 0.4 and 0.6 ml). The number of instances in each trial type “stack” indicates the probability of drawing the trial type.
Extended Data Fig. 4 | Amplification effect was robust. Box and whisker plots show the baseline subtracted responses to 0.2 and 0.6 ml of juice, as in Fig. 3b, but applied to all 34 neurons that were significantly modulated by value. * indicates $p < 0.05$, ** indicates $p < 0.01$, $N = 34$ neurons, Wilcoxon signed-rank test, Bonferroni corrected for multiple comparisons. Box and whisker plots show, median (line), quartiles (boxes), range (whiskers), and outliers (+).
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
- Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted. Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection | Custom MATLAB (R2019b) data collection software

Data analysis | MATLAB R2019b, Python 3.7.2, Python package nideconv

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data are available upon reasonable request from the corresponding author, William R. Stauffer wrs@pitt.edu
# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- [ ] Life sciences
- [ ] Behavioural & social sciences
- [ ] Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](http://nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| Sample size | Following standard sample sizes used in studies investigating neuronal responses in non-human primates, we recorded extracellular activity from 67 dopamine neurons in two monkeys, and 40 neurons had a sufficient number of trials for further analyses. Citations in text. |
|-------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Data exclusions | Per predetermined criterion, 27 of the 67 neurons were excluded due to an insufficient number of trials in all of the trial types used. |
| Replication | The amplification of dopamine responses was confirmed in a second experimental animal. |
| Randomization | Not applicable, there were no experimental or control groups. |
| Blinding | Not applicable, there were no experimental or control groups. |

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

| n/a | Involved in the study |
|-----|-----------------------|
| [ ] | Antibodies |
| [ ] | Eukaryotic cell lines |
| [ ] | Palaeontology and archaeology |
| [ ] | Animals and other organisms |
| [ ] | Human research participants |
| [ ] | Clinical data |
| [ ] | Dual use research of concern |

### Methods

| n/a | Involved in the study |
|-----|-----------------------|
| [ ] | ChIP-seq |
| [ ] | Flow cytometry |
| [ ] | MRI-based neuroimaging |

## Animals and other organisms

Policy information about [studies involving animals; ARRIVE guidelines](http://ARRIVEguidelines.org) recommended for reporting animal research

| Laboratory animals | Two male rhesus macaques, age 6 |
|--------------------|---------------------------------|
| Wild animals       | No wild animals were used in this study. |
| Field-collected samples | No field-collected samples were used in this study. |
| Ethics oversight   | All animal procedures were approved by Institutional Animal Care and Use Committee of the University of Pittsburgh. |

Note that full information on the approval of the study protocol must also be provided in the manuscript.