Published in final edited form as:
Nat Neurosci. 2021 April ; 24(4): 465–469. doi:10.1038/s41593-021-00807-7.

Rare Rewards Amplify Dopamine Responses

Kathryn M. Rothenhoefer[1,2,3,4], **Tao Hong**[2,3,4,5], **Aydin Alikaya**[1,2,3,4], **William R. Stauffer**[1,2,3,4,*]

[1]Center for Neuroscience, University of Pittsburgh, Pittsburgh, PA, USA.

[2]Center for the Neural Basis of Cognition, University of Pittsburgh, Pittsburgh, PA, USA.

[3]Systems Neuroscience Center, University of Pittsburgh, Pittsburgh, PA, USA.

[4]The Brain Institute, University of Pittsburgh, Pittsburgh, PA, USA.

[5]Program in Neural Computation, Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA, USA.

Abstract

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[1]. These learning signals are approximated by reinforcement learning algorithms, including Temporal Difference (TD) and Rescorla-Wagner learning models[2,3]. According to standard TD learning, ‘reward predictions’ are simply point estimates – formally, the temporally discounted sum of future outcomes[3]. 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[2]. However, point estimate predictions reflect neither predicted uncertainty, nor the shapes of reward distributions, and they are not adequate descriptors of behavior[4–7]. Consider that, learning takes longer when rewards are sampled from broader distributions, compared to when they are sampled from narrower distributions[5,6]. Likewise, decision-makers take longer to choose between options when value differences are small, compared to when differences are large[7]. These results demonstrate that probability distributions over reward values, and not simply

[*]Corresponding Author: wrs@pitt.edu.

Author Contributions: KMR & WRS designed the experiment, KMR, AA & WRS collected data, KMR, TH & WRS analyzed the data and wrote the paper.

Competing Interests Statement: The authors declare no competing interests.

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.

point estimates such as the mean, influence behavior. Dopamine responses adapt to the range or standard deviation of predicted outcomes[8], 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 (RL) has produced remarkable advances in artificial intelligence[9,10] and RL techniques have recently been extended to learning probability distributions[11]. Distributional RL 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 RL were reflected by dopamine neurons, raising the enticing possibility that brains employ a distributional code for value[12]. 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 EVs (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 ($\beta_{Trial} = 0.110$, $p < 0.0001$, $\beta_{Normal} = 0.167$, $p = 0.007$, $N = 6098$ trials*, t*-test). We used a standard RL model[3] 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-predictions 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 Figure 1, Methods). Here, the magnitudes of the S, M, and L rewards were fixed at 0.2, 0.4, and 0.6 ml, respectively (Fig. 1a). Prior choice testing confirmed that Normal and Uniform distributions with these reward size elements had equivalent expected utilities (EUs) (Extended Data Figure 2, Methods). As expected from cues that predict the same EUs, 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 (CS-matched), or the number of times each prediction error was experienced (PE-matched) (Extended Data Figure 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 Uniform distribution trials (Fig. 2c, d, dashed lines). Linear regression revealed that thirty-four neurons were significant for reward size, and that the vast majority of 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 naive 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 ml 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[12], 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 cells' 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, Extended Data Figure 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.

Discussion

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 suggest a novel paradigm 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 (EUs) to both distributions (Extended Data Figure 2). EU is a proxy for the 'predicted reward value' term used to describe dopamine reward prediction error responses[13]. 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 RL algorithms with performance that exceeds expert human performance on Atari games, chess, and Go[9,10]. Recently, a novel machine learning model, distributional TD, was applied to study the activity of dopamine neurons[12]. 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 suggest 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 bi-directional 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 (RL) models that acquire only point estimate predictions are not adequate to describe dopamine activity. Rather, these data suggest that RL algorithms that track uncertainty, such as Kalman TD[14], may provide an appropriate conceptual framework to explain information processing in the reward system. Kalman-like reinforcement signals enables reward prediction and estimation of uncertainty[15], 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.

The amplification of dopamine responses by rare rewards appears to be a distinct phenomenon from novelty driven dopamine responses that we and others have previously observed[16,17]. Stimulus novelty decays with the number of exposures and dopamine responses appear to follow this decay[16]. A recent study has shown that stimulus novelty, specifically, and not rarity, drives the large dopamine responses observed during the first exposures to stimuli, and that novelty-driven CS responses promote learning[17]. None of the rewards used in our study were novel, as only three rewards were used while recording: 0.2

ml, 0.4 ml, and 0.6 ml of blackcurrant juice. The monkeys experienced these three rewards hundreds of times each during every session. Rarity was maintained only during Normal trials, when 0.2 ml and 0.6 ml were rarely given. Thus, the amplification of dopamine responses to rewards drawn from the tails of Normal distributions is likely a function of reward rarity, and distinct from novelty responses.

Pupil diameter was sensitive to prediction errors generated during active learning phases, but the sensitivity sharply decreased after learning. This result is consistent with prior studies showing that pupil diameter is more sensitive to unexpected uncertainty, compared to expected uncertainty[18], and indicates that the monkeys learned the distributions and their associated expected uncertainties. In parallel, we observed that learning was enhanced in the Normal distribution trials. Specifically, we observed that learning became asymptotic after fewer trials in Normal compared to Uniform blocks. Together, these results are consistent with prior studies in humans showing that reward learning is dependent on standard deviation and higher statistical moments of reward distributions[4–6]. Further experiments will be required to disentangle the effects of higher statistical moments, especially standard deviation and kurtosis, on reward learning. Nonetheless, amplified dopamine reward prediction error responses are a candidate neural mechanism to explain how distribution shape affects learning.

One limitation of our study is that the behavioral choice data and neural recordings were collected using different tasks. The behavioral paradigm 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[19]. However, the experimental separation of the behavioral measurement from the neural recordings prevents us from drawing firm conclusions regarding the role of dopamine signal amplifications 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 responses have the ability to modulate dopamine concentrations in the prefrontal cortex (PFC), which are tightly linked to neuronal signaling and working memory performance[20]. 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.

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, 13.9 and 11.2 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 Figure 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 frequency (1/3 probability 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 (CS-matched), and one where there were equal instances of nonzero prediction errors for both Normal and Uniform (PE-matched) (Extended Data Figure 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 inter-trial 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: 10 sessions, ~2800 trials; monkey S: 6 sessions, ~2600 trials).

Choice Tasks for Measuring Distribution Values—For the data presented in Extended Data Figure 2b–d, three cues predicted a Normal distribution (Fig. 1a, right), and three different cues predicted a Uniform distribution (Fig. 1a, left). One small, ‘safe’ cue predicted 0.2 ml of juice and one large, ‘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 towards 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 inter-trial 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 Figure 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 cue had a value range of 0 ml 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 towards 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 inter-trial 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 Figure 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 novel cue had a different pseudo-randomly selected mean that was either 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 towards 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 inter-trial 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 (EUs) of the distributions were the same (Extended Data Figure 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 used six different cues to predict distributions – three cues predicted the Normal distribution and three different cues predicted the Uniform distribution (Extended Data Figure 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[14]. We used classical conditioning to train monkeys on the cue-reward contingencies, then we measured binary choices between the cues (Extended Data Figure 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 Figure 2c) (p = 0.48, N = 9 cue pairs, t-test). Additionally, the monkey chose the Normal distribution over the small reward (Extended Data Figure 2c, $p < 0.0001$, t-test) and the large reward over the distribution (Extended Data Figure 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 Figure 2d, p = 0.001 and 0.005, respectively N = 3 cue pairs, each, 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 EUs were critical to our interpretation of the data, and as such, we replicated this result using a different behavioral paradigm: we independently measured the certainty equivalents (CEs) of Normal and Uniform reward distributions. CEs are the volumes 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 Figure 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 Figure 2f, g). The CEs was the safe values that corresponded to P(Choose Safe) = 0.5 (black arrows in Extended Data Figure 2f, g). Analysis of the session-by-session CEs for the Normal and Uniform blocks found no effect of the distribution type on the CEs (p = 0.2, N = 18. T-test). Therefore, the CEs 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.

$$\log\left(\frac{P(Correct)}{1 - P(Correct)}\right) = \beta_0 + \beta_D * D + \beta_C * C + \beta_T * T$$

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 (RL) model to examine monkeys' choices during learning and to acquire trial-by-trial estimate of chosen and unchosen values[3]. 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:

$$P(pd1)_t = \frac{e^{V_t(pd1)/\beta}}{e^{V_t(pd1)/\beta} + e^{V_t(pd2)/\beta}}$$

where β, the temperature parameter of the softmax rule, determines the level of choice randomness.

In each trial, upon 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:

$$V_{t+1} = V_t + \alpha * \delta_t$$

where α denotes the learning rate, and the prediction error is calculated as the following: $\delta_t = r_t - V_t$, indicates the difference between the predicted and realized reward sizes, V_t and r_t, respectively. The free parameters, α and β, were fit by maximizing the likelihood of the model. After fitting the model, we took the trial-wise mean of the unsigned PE over blocks of the same type (Fig. 1e).

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 faster learning 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 *nideconv*[21,22], 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 3 s post-event; the time window for the presence of rewards started at 0.5 s pre-event and ended at 1.5s post-event. To understand the relationship between pupil diameter and prediction error post-reward, 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 post-reward pupil diameter changes are 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 reward distribution type and calculated 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 (band-pass 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 (a) 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[23], (b) tonically active position coding ocular motor neurons and (c) 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 Figure 1a) and low baseline firing (< 8 impulses/s)[24]. Following standard sample sizes used in studies investigating neuronal responses in non-human 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 Figure 1b, $p < 0.0001$, $N = 40$ neurons; Wilcoxon rank-sum test). Extended Data Figure 1c and d show maps of our recording locations relative to both monkeys' grids, and the number of cells recorded at each location. Extended Data Figure 1e and f show MRI images of monkey S and the location of the recordings.

Analysis of Neural Data

Data Pre-Processing—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 $(1 - e^{-t})e^{-t/T}$, 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 −1000 ms to 0 ms relative to cue onset.

Single Neuron Linear Regression—To determine whether previous rewards influence the current CS response, we fit a linear model to each neurons' CS response, using the rewards from the previous 5 trials as the independent variables. We found that previous outcomes up to 5 trial back did not influence CS response. This result is not particularly surprising in the Normal distribution trials, as the previous 5 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 Figure 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:

$$F_z = \beta_0 + \beta_1 * D + \beta_2 * R + \beta_3 * D \times R$$

where F_z 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. Fig. 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 towards 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[25]. 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 5000 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 distribution types above chance. Therefore, we used a Selectivity Index (SI) to select neurons for decoding. The single-neuron SI 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.

$$\mathrm{SI} = \frac{\overline{F_N} - \overline{F_U}}{\sigma_P}$$

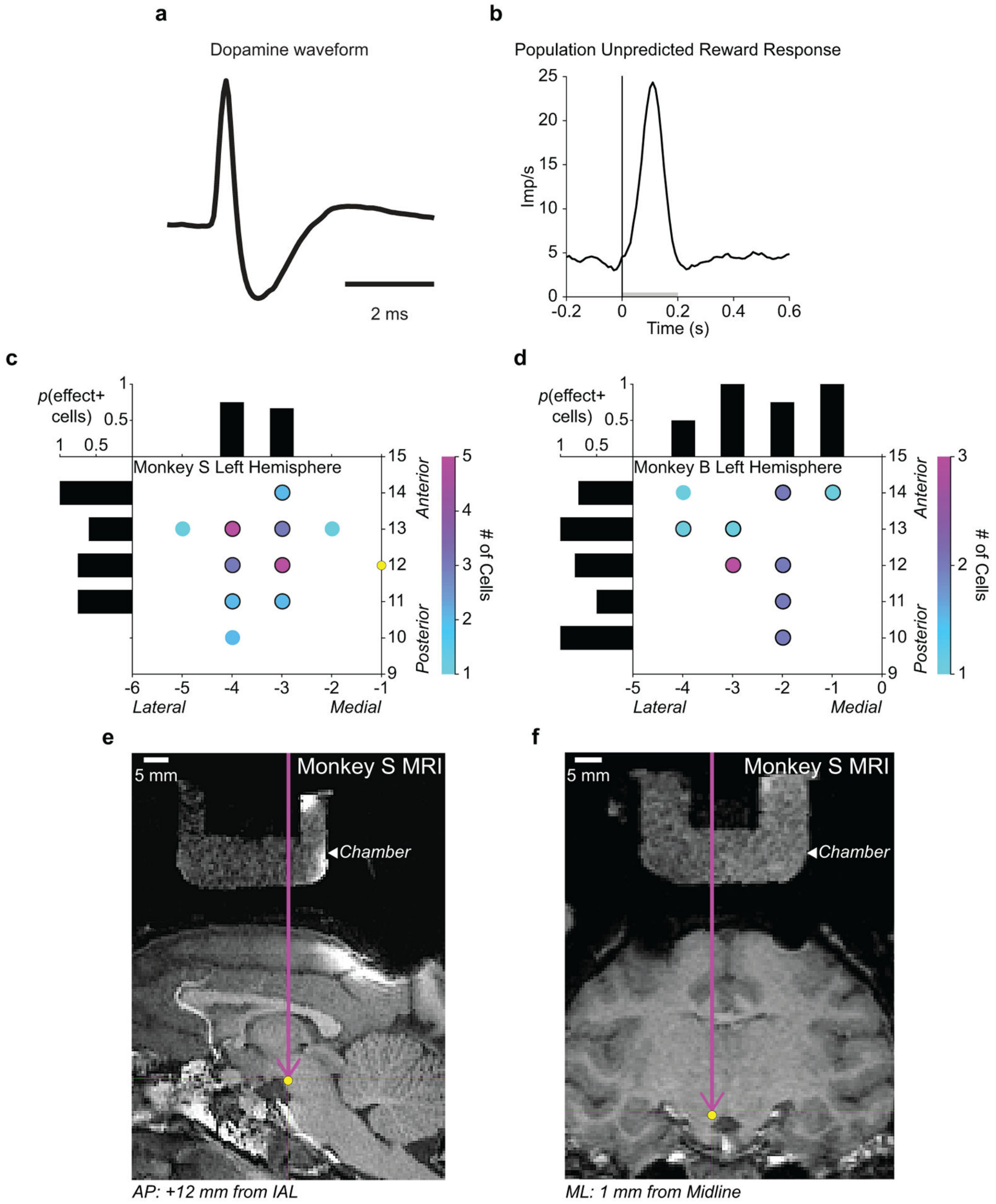
The subset of 11 neurons with the largest SI successfully decoded the predicted distribution from the responses to 0.2 and 0.6 ml (Fig. 2g). To ensure that the rest of the neurons did not encode an opposite effect, we built a classifier with the rest of the neurons (29/40) and did not observe above-chance 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[12], 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 if 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

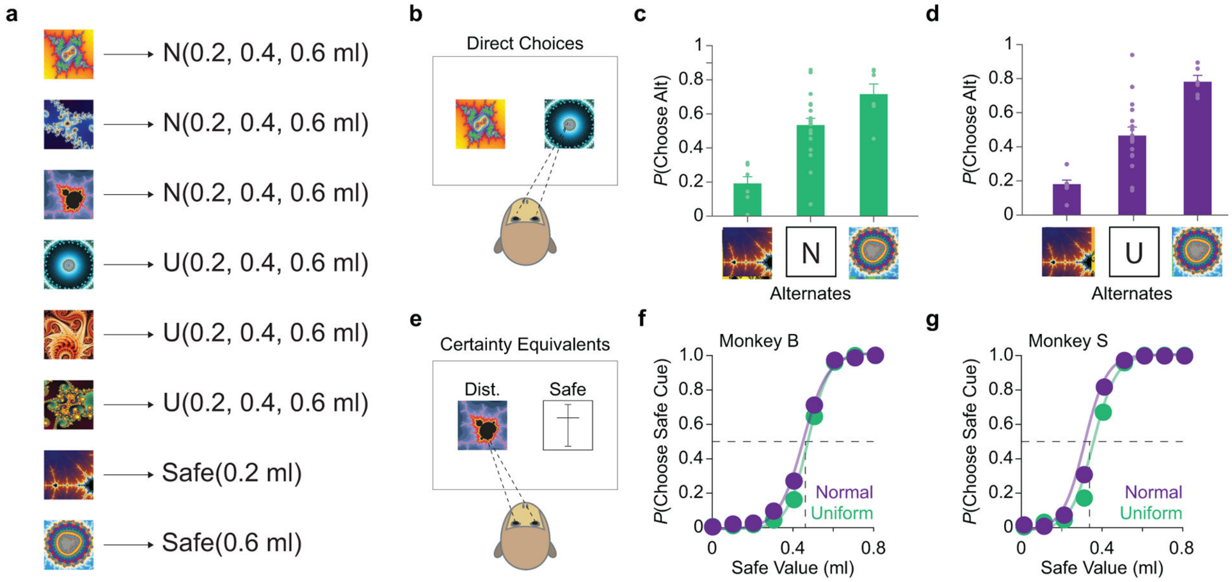
All statistical analyses were performed and all graphs were created in Python 3.7.2, Matlab R2019b and Python package nideconv. No statistical methods were used to predetermine sample sizes, but our sample sizes are similar to those reported in studies investigating neuronal and behavioral responses in non-human primates[8,14,17]. 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 of the 67 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 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.

Extended Data



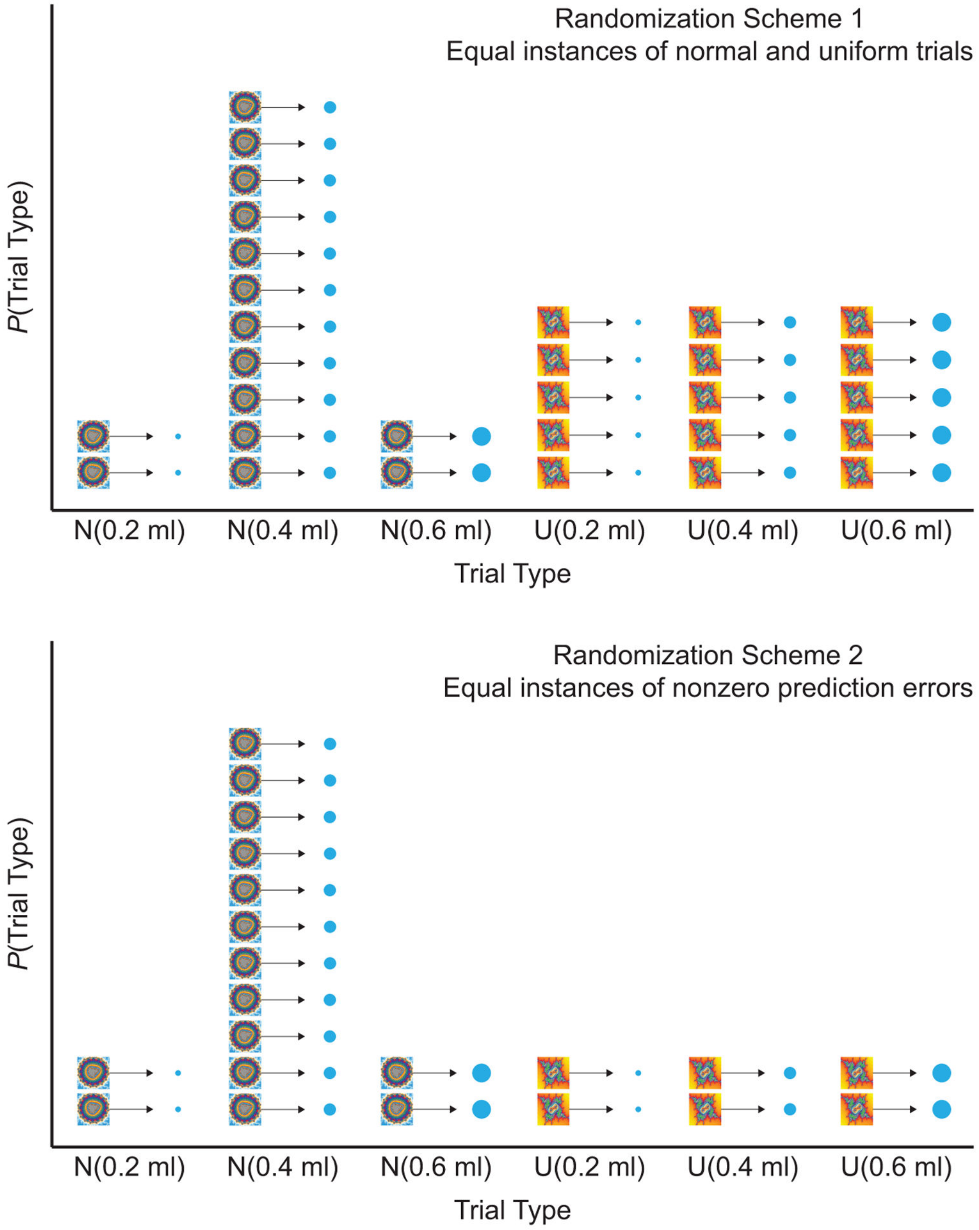
Extended Data Fig. 1. Dopamine neurons and recording sites.

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. Grey 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.



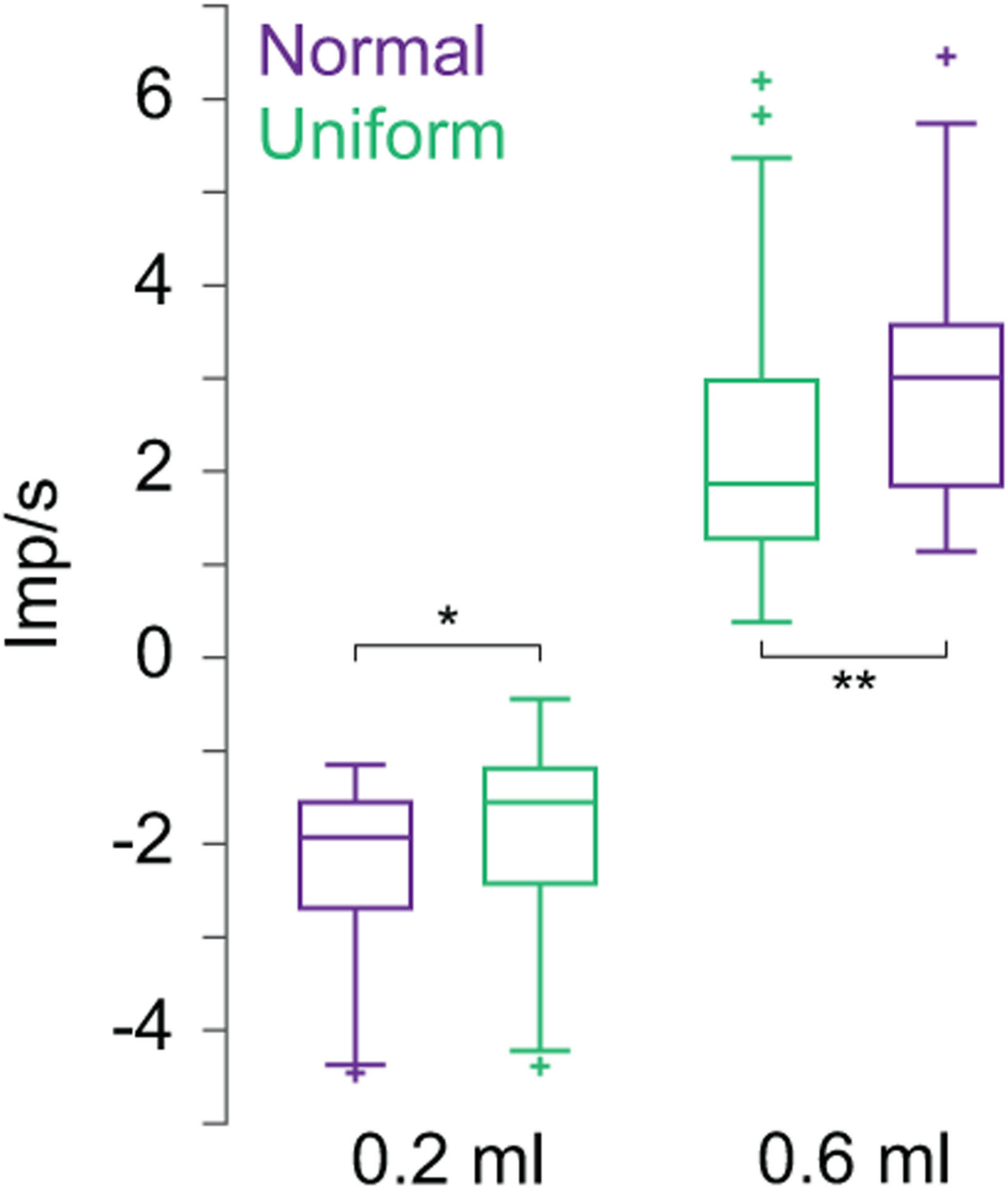
Extended Data Fig. 2. Normal and Uniform reward size distributions have equivalent subjective values.

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 0.4ml. 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, grey 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 (+).

Acknowledgements:

The authors thank Andreea Bostan for comments and discussion, and Jacquelyn Breter for animal care and enrichment.

References

1. Stauffer WR The biological and behavioral computations that influence dopamine responses. Current Opinion in Neurobiology 49, 123–131 (2018). [PubMed: 29505948]
2. Enomoto K et al. Dopamine neurons learn to encode the long-term value of multiple future rewards. Proc Natl Acad Sci U S A 108, 15462–15467 (2011). [PubMed: 21896766]
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. Cerebral Cortex 26, 1818–1830 (2016). [PubMed: 26850528]
5. Diederen KMJ & Schultz W Scaling prediction errors to reward variability benefits error-driven learning in humans. Journal of Neurophysiology 114, 1628–1640 (2015). [PubMed: 26180123]
6. Nassar MR, Wilson RC, Heasly B & Gold JI An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. J Neurosci 30, 12366–12378 (2010). [PubMed: 20844132]
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). [PubMed: 20835253]
8. Tobler PN, Fiorillo CD & Schultz W Adaptive coding of reward value by dopamine neurons. Science 307, 1642–1645 (2005). [PubMed: 15761155]
9. Mnih V et al. Human-level control through deep reinforcement learning. Nature 518, 529–533 (2015). [PubMed: 25719670]
10. Silver D et al. Mastering the game of Go with deep neural networks and tree search. Nature 529, 484–489 (2016). [PubMed: 26819042]
11. Bellemare MG, Dabney W & Munos R in Proceedings of the 34th International Conference on Machine Learning - *Volume* 70 449–458 (JMLR.org, Sydney, NSW, Australia, 2017).
12. Dabney W et al. A distributional code for value in dopamine-based reinforcement learning. Nature 577, 671–675 (2020). [PubMed: 31942076]
13. Stauffer WR, Lak A & Schultz W Dopamine reward prediction error responses reflect marginal utility. Curr Biol 24 (2014).
14. Gershman SJ A Unifying Probabilistic View of Associative Learning. PLoS Comput Biol 11, e1004567 (2015). [PubMed: 26535896]
15. Babayan BM, Uchida N & Gershman SJ Belief state representation in the dopamine system. Nat Commun 9, 1891 (2018). [PubMed: 29760401]
16. Lak A, Stauffer WR & Schultz W Dopamine neurons learn relative chosen value from probabilistic rewards. Elife 5 (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). [PubMed: 32027824]
18. Preuschoff K, Marius't Hart B & Einhauser W Pupil Dilation Signals Surprise: Evidence for Noradrenaline's Role in Decision Making. Frontiers in Neuroscience 5 (2011).
19. Schultz W Neuronal Reward and Decision Signals: From Theories to Data. Physiol Rev 95, 853–951 (2015). [PubMed: 26109341]
20. Vijayraghavan S, Wang M, Birnbaum SG, Williams GV & Arnsten AF Inverted-U dopamine D1 receptor actions on prefrontal neurons engaged in working memory. Nat Neurosci 10, 376–384 (2007). [PubMed: 17277774]

21. Van Slooten JC, Jahfari S, Knapen T & Theeuwes J How pupil responses track value-based decision-making during and after reinforcement learning. PLOS Computational Biology 14, e1006632 (2018). [PubMed: 30500813]

22. de Hollander G & Knapen T nideconv, <https://nideconv.readthedocs.io/en/latest/> (2017).

23. Loe PR, Whitsel BL, Dreyer DA & Metz CB Body representation in ventrobasal thalamus of macaque: a single-unit analysis. J Neurophysiol 40, 1339–1355 (1977). [PubMed: 411896]

24. Guyenet PG & Aghajanian GK Antidromic identification of dopaminergic and other output neurons of the rat substantia nigra. Brain Res 150, 69–84 (1978). [PubMed: 78748]

25. Batista AP et al. Cortical neural prosthesis performance improves when eye position is monitored. IEEE Trans Neural Syst Rehabil Eng 16, 24–31 (2008). [PubMed: 18303802]

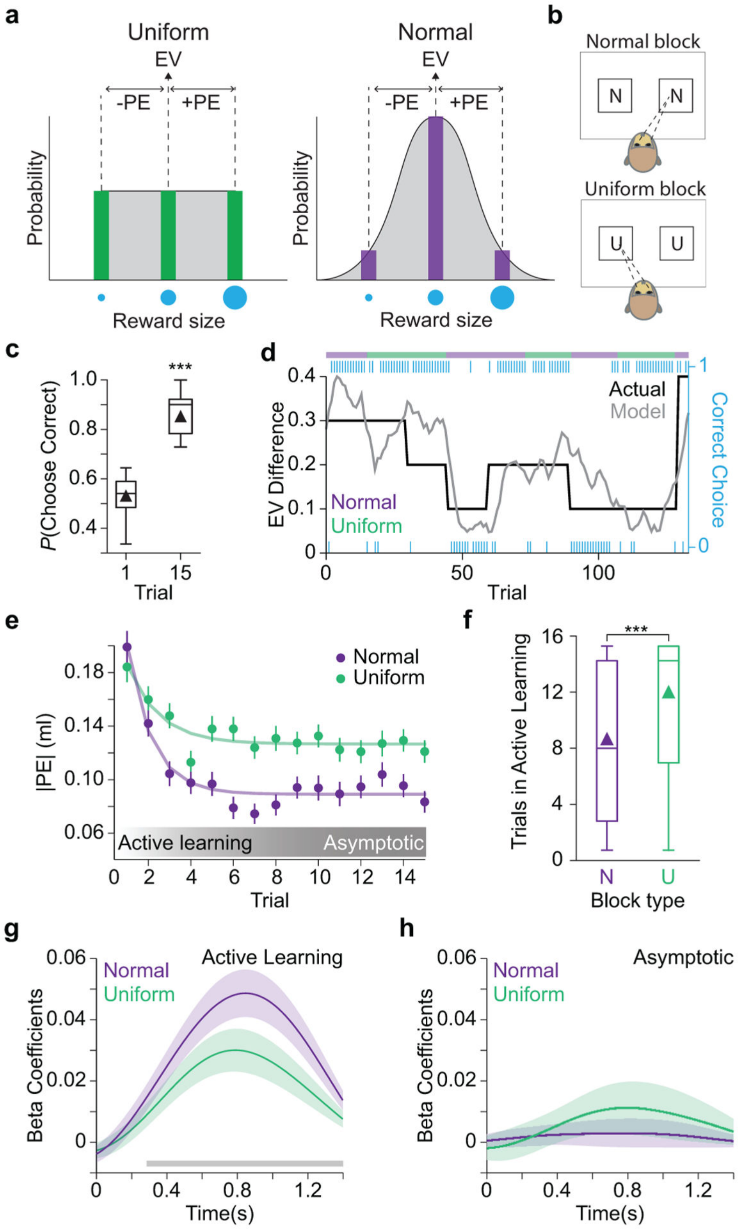


Figure 1. Behavior

a, Schematic of Uniform (green) and Normal (purple) reward size probability distributions. Gray shaded regions illustrate the hypothetical probability density functions. Green and purple bars indicate the discrete probabilities of small (0.2 ml, small blue circle), medium (0.4 ml, medium blue circle), and large (0.6 ml, large blue circle) rewards. Abbreviations: (EV) expected value, (-PE) negative prediction error, (+PE) positive prediction error. A fractal drawing associated with a Uniform distribution predicted that each reward size would be drawn with 1/3 probability. A different fractal drawing associated with a Normal distribution predicted that the small and large reward volumes would be drawn 2/15 times, whereas the medium reward size would be drawn the remaining 11/15 times. **b,** Task setup where each block was either a Normal or a Uniform block. In Normal blocks, two novel fractal images represented Normal distributions with different EVs. Likewise, in Uniform blocks, two novel fractal images represented Uniform distributions with different EVs. **c,** Box and whisker plots show the probability of choosing the higher-valued option on trials 1 (left) and 15 (right), across both distribution types. Triangles represent the averages. *** indicates $p < 0.0001$, $N = 275$ blocks, *t*-test. **d,** RL model performance for a subset of trials. Actual (black) and estimated (grey) value differences for two choice options. Bar at the top indicates either Normal (purple) or Uniform (green) block type. The primary y-axis shows the EV differences between the two choice options, and the x-axis shows trial number. The blue tick marks correspond to correct and incorrect choices, defined by the relative expected values, and indicated by the secondary y-axis. **e,** Absolute prediction errors as a function of trial number within Normal (purple) and Uniform (green) blocks. Error bars are ±SEM across 142 Normal blocks and 133 Uniform blocks, and solid lines are exponential functions fit to the data. Shaded box schematically describes the transitions from “Active Learning” to “Asymptotic” behavior, the actual transition trials were determined on a block-wise basis (Methods). **f,** Box and whisker plot shows the number of trials in the “Active Learning” phase for Normal (N, purple) and Uniform (U, green) distribution blocks. *** indicates $p < 0.0001$, Mann-Whitney U test. **g,** Beta coefficients from a deconvolution analysis on the pupil diameter data, for the trials in the active learning phase of Normal (purple) and Uniform (green) blocks, aligned to reward delivery at time = 0. The grey horizontal bar indicates time points after reward where the Normal beta coefficients were significantly different from the Uniform beta coefficients ($p < 0.05$, $N = 4703$ trials, Cluster-based permutation test). Shaded regions indicate 95% confidence interval over trials. **h,** as in f, for trials in the asymptotic phase. Box plots show, mean (triangles), median (line), quartiles (boxes), and range (whiskers).

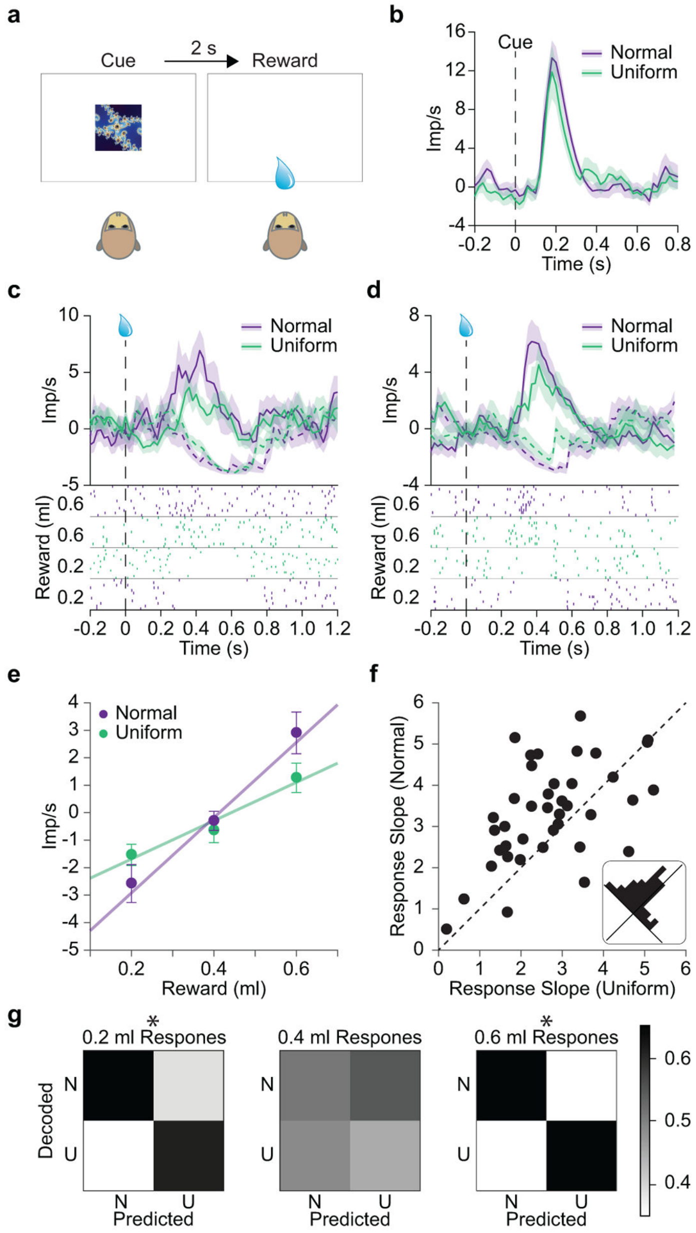


Figure 2. Rare rewards amplified dopamine reward prediction error responses.

a, In the recording task, the monkeys viewed a distribution predicting CS and rewards were delivered two seconds later. **b,** Peri-stimulus time histogram (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 ±SEM across trials. **c,** Single neuron reward responses to rewards during Normal and Uniform trials, recorded using the CS-matched randomization scheme (Extended Data Figure 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 ±SEM across trials. **Bottom:** Raster plots, 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 PE-matched randomization scheme (Extended Data Figure 3). **e,** Single neuron linear regression of a single 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 ± SEM 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: 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 ml, 0.4ml and 0.6ml 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 ml and 0.6 ml ($p = 0.045$ and $p = 0.028$, $N = 11$ neurons, Permutation test, uncorrected p-values). No asterisk above 0.4 responses indicate no significant decoding ($p = 0.642$, $N = 11$ neurons, Permutation test).

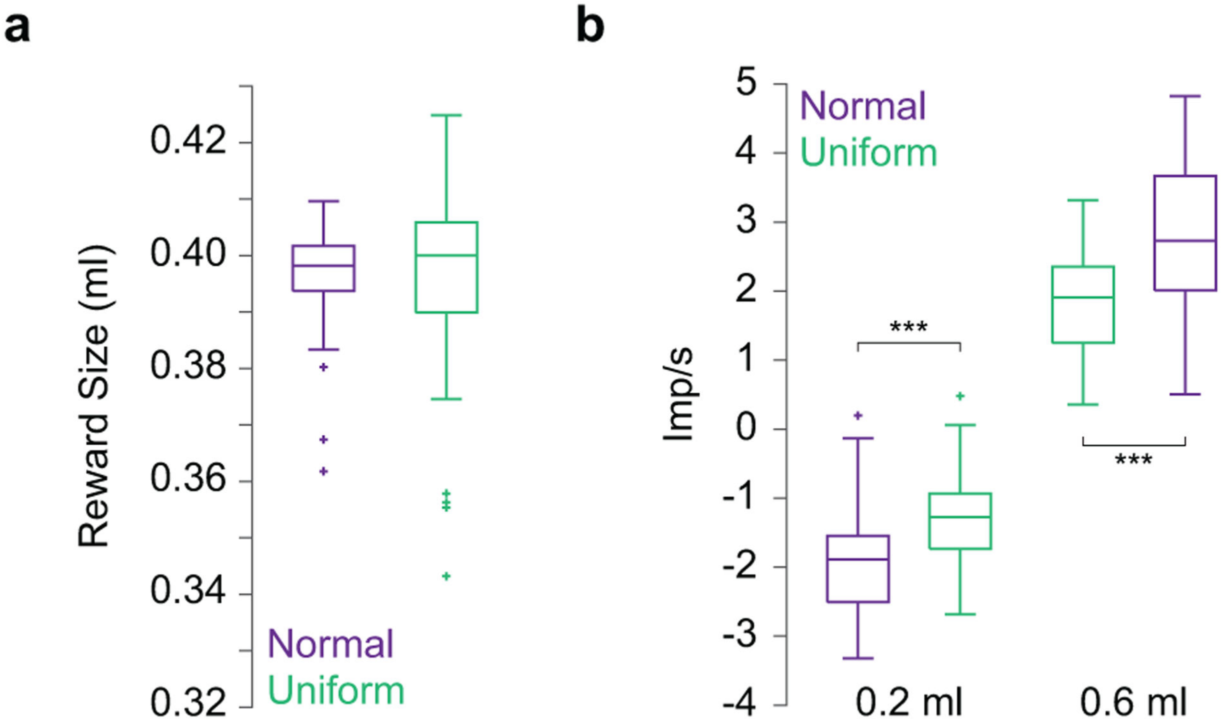


Figure 3. Dopamine pseudo-populations and single neurons simultaneously reflect predicted probability distributions.

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. *** indicates $p < 0.0001$, $N = 29$ 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, median (line), quartiles (boxes), range (whiskers), and outliers (+).