Supplemental Material: A distributional code for value in dopamine-based reinforcement learning

Will Dabney, Zeb Kurth-Nelson, Naoshige Uchida, Clara Kwon Starkweather, Demis Hassabis, Rémi Munos, Matthew Botvinick

October 31, 2019

In this supplement, we present additional information concerning computational and experimental methods, as well as additional data and analysis. This material is divided into six sections: Section 1 covers the basic mechanisms underlying distributional RL, including a discussion of why distributional RL accelerates learning; Section 2 considers a range of alternative models, each implementing a different explanations of the experimental results; Section 3 tests the robustness of our results to modeling assumptions involved in our analyses; Section 4 presents supplementary experimental results; Section 5 discusses relations to previous work; and Section 6 gives further predictions of the theory.

1 Distributional Reinforcement Learning: Basic Principles

1.1 Distributional Temporal Difference Learning

In this section we provide a more complete and accessible introduction to distributional TD learning. The problem of RL is finding the best behavior policy. A policy, π, is a specification for every possible state of which action(s) to take in that state. Thus, starting from a state x, running the policy forward determines the future rewards R_t that will be received at all future times t. The sum of these future rewards, discounted by γ, is called the *return*, Z:

$$Z^\pi(x) = \sum_{t=0}^{\infty} \gamma^t R_t$$

However, Z is not a single number but a random variable whose value depends on the possible rewarding outcomes. Given a known starting state x and a policy π, there are two sources of randomness: the environment may have stochastic transitions and rewards, and the policy itself could be probabilistic. Most existing RL algorithms learn about $V = \mathbb{E}[Z]$, the expectation of the return. For example, the simple TD rule with learning rate α:

$$\delta = r + \gamma V(x') - V(x), \quad V(x) \leftarrow V(x) + \alpha\delta \tag{1}$$

computed at the transition from state x to state x', moves V toward the expectation of the return[1].

However, it may be useful to learn about the full distribution. There is a simple way to learn the distribution, using a very small modification to standard TD learning[2]. In this method, instead of a single value function, a set of value functions is learned. For each value function V_i, a distinct reward prediction error δ_i is computed:

$$\delta_i = r + \gamma V_j(x') - V_i(x),$$

where $V_j(x')$ is a sample from the distribution $V(x')$.

The value update rule of Equation 1 is modified (see also Bowling and Veloso[3]) such that different learning rates α^+ and α^- apply to positive and negative RPEs:

$$V_i(x) \leftarrow V_i(x) + \alpha_i^+ \delta_i \qquad \text{for } \delta_i > 0 \tag{2}$$

$$V_i(x) \leftarrow V_i(x) + \alpha_i^- \delta_i \qquad \text{for } \delta_i < 0 \tag{3}$$

At convergence, the learned V_i together comprise a set of sufficient statistics for the distribution of returns from state x. Although the population of statistics, when averaged, mimics the classic mean-value function, the individual statistics vary significantly from this mean. Thus, distribution learning arises automatically from a very simple modification to standard TD learning.

Next, we discuss the intuition for why the learned V_i comprise a set of sufficient statistics for the return distribution. The learning rule described in equations 2 and 3 can be generalized as follows:

$$V_i(x) \leftarrow V_i(x) + \alpha_i^+ f(\delta_i) \qquad \text{for } \delta_i > 0 \tag{4}$$

$$V_i(x) \leftarrow V_i(x) + \alpha_i^- f(\delta_i) \qquad \text{for } \delta_i < 0 \tag{5}$$

where f is a function that transforms the prediction error. For any non-decreasing f, the learned V_i's comprise a set of statistics for the return distribution.

This is particularly easy to see for the special case of $f(x) = \text{sign}(x)$. Suppose $\alpha_i^+ = 3$ and $\alpha_i^- = 1$. In other words, every positive prediction error (PE) is $+3$ and every negative PE is -1. In this case, V_i will converge to predict exactly the upper quartile of the return distribution. At the upper quartile, there will be three times as many negative PEs as positive PEs, so after being weighted by α_i^+ and α_i^-, positive and negative PEs will exactly balance. Thus, this is the fixed point of the learning dynamics. Under mild assumptions, each V_i converges exactly on a *quantile* of the return distribution. The particular quantile is $\tau_i = \frac{\alpha_i^+}{\alpha_i^+ + \alpha_i^-}$. In other words, V_i is the return that has cumulative probability τ_i. As long as τ_i spans the range from 0 to 1, the set of V_i together form the complete quantile function (aka inverse cumulative distribution function) for the distribution [2,4].

Quantile regression, where $f(\delta) = \text{sign}(\delta)$, is useful to visualize distribution learning. However, it is physiologically implausible that dopaminergic firing follows a sign function. Maximum likelihood estimation (MLE) is widely used to estimate the parameters of a statistical model and provides many appealing properties and guarantees. In our work we rely heavily upon a class of *maximum likelihood-like* estimators, known as M-estimators, which provide many of the same guarantees while significantly expanding the class of functions [5]. For a prediction error δ, the quantile regression loss,

$$\rho_\tau(\delta) = (\tau - \mathbb{I}_{\delta \leq 0})\delta, \quad \rho'_\tau(\delta) = |\tau - \mathbb{I}_{\delta \leq 0}| \times \text{sign}(\delta),$$

is an asymmetric absolute loss whose minimum is the quantile function at τ. But quantile regression is just one of many asymmetric regression methods that learn statistics that fully describe a scalar distribution. Given a response function $f(\delta)$ and incremental update following the gradient

$$\rho'_\tau(\delta) = |\tau - \mathbb{I}_{\delta \leq 0}| \times f(\delta)$$

various M-estimators can be found depending on the choice of response function f. In the main text, we used a linear response function, $f(\delta) = \delta$. This response function causes a different statistic, called *expectiles*, to be learned [6]. We chose this function because neural responses in estimated utility space were empirically close to linear in each domain (negative and positive). The qualitative pattern of results was not sensitive to the choice of response function.

1.2 Benefits of distributional RL

In the main text we showed that distributional RL significantly improves performance in simulated agents; this finding has also been replicated multiple times [7–10]. To contextualize the scale of the performance gains, here we compare the benefit of distributional RL against other architectural advancements (Extended Data Figure 2). The techniques of prioritized replay and double DQN were both considered to induce step-changes in performance relative to standard DQN [11], and distributional RL yields a performance advantage greater than either.

Why does distributional RL help? One hypothesis for why distributional RL improves performance of artificial agents is that it drives learning of richer representations. States associated with the same expected return but different return distributions may be represented as different under distributional RL, but not under ordinary RL. (A similar proposal has been made about the performance benefits that arise from requiring an agent to learn multiple value functions at multiple discount rates [12]).

To illustrate this idea, we analyzed the representations learned by DQN and distributional TD in the Atari 2600 game Ms. Pacman. For both fully trained agents we generated trajectories by allowing

the agents to play the game, but with a higher than usual (0.1) probability of taking random actions. For each trajectory we recorded the full-resolution frame and the activations of the final hidden layer of the neural network (a 512 dimension real-valued representation vector). We then trained, for each agent, a linear decoder from the representation vector to the game frame. How well the agent is able to reconstruct the full game state can be taken as an indication of how rich a representation has been learned.[13] We split each agent's trajectories into a training and testing set. In Extended Data Figure 2b,c we show example reconstructions as well as the mean-squared reconstruction error (MSE) on the test set.

If representation learning is the reason for the improved performance of distributional TD, we should be able to control the degree to which distributional TD boosts performance by manipulating the representational demand of a task. We therefore designed a simple, laboratory-style task with seven states arranged in a tree. The task required no actions; the agent's only job was to predict reward. In each episode, the agent first transitioned with 0.5 probability from the starting state to either a "dog" state or a "bird" state. The agent would then observe an image (drawn at random from CIFAR-10)[14] of the corresponding image class. Next, the agent would transition to a second-stage state where it received a positive or negative reward. Following the dog state, the agent could receive -1 or +3 reward, and following the bird state, the agent would always receive +1 reward. For the first 1000 episodes (Task 1), the agent received -1 or +3 reward with 0.5 probability following the dog. For the second 1000 episodes (Task 2), the agent received a guaranteed reward of +3 following the dog. We observed that distributional TD was much faster than classic TD to learn the correct reward prediction after the change in transition probabilities (Extended Data Figure 3a-c).

This speedup is attributable to the fact that the distributional loss should encourage the network to learn features that discriminate between dogs and birds, because they lead to different reward distributions. As a direct measurement of this, we looked at the representations learned by classic TD and distributional TD in each task (Extended Data Figure 3e). In Task 1, distributional TD but not classic TD learned features that discriminated between birds and dogs.

Finally, we performed a control experiment, identical to the first bird-dog experiment, but replacing birds and dogs with airplanes and ships after the first 1000 episodes. We found that distributional TD's advantage was greatly reduced in this experiment, in line with our predictions (Extended Data Figure 3d). For all experiments using CIFAR-10 data, we used an 11-layer convolutional neural network suitable for classification in CIFAR-10, but with the final classification layer replaced for regression. We trained with mini-batch sizes of 256 and using the Adam optimizer with learning rate 0.0005[15].

Does distributional TD increase network capacity? We have argued that distributional RL drives representation learning, and therefore a natural question is whether such effects are due to the learning process itself or to adding representation capacity to the neural network itself by way of additional parameters. In Extended Data Figure 2a we clarify that it is entirely the former, with $|\phi| = 512$ and convolutional network number of filters, filter size, and stride as $|A| = (32, 8, 4)$, $|B| = (64, 4, 2)$ and $|C| = (64, 3, 1)$. Both algorithms represent outputs as linear functions on top of the same non-linear function, ϕ. That is, although there are more linear functions, w_a^i, for distributional TD than for DQN, w_a, there is no extra capacity in the network.

Extension to risk-sensitive behavior In the neural network simulations of Section 1.2, the agent selected *actions* according to the mean (so the policy effectively discarded information about the distribution) – the benefits of distributional RL arose from facilitating *learning*. However, it may also be the case that an organism does not always act according to the mean return. For example, if the rat needs some amount of cheese not to starve, then it should risk everything for the long odds of reaching the largest outcome. In general, distributional RL may be useful if an agent wants to rapidly shift its risk preferences without requiring new (slow) learning. For example, humans and animals shift their risk preferences according to hunger[16,17].

2 Evaluation of Alternative Models

In this section we introduce a set of potential null models, which attempt to explain our experimental results without invoking distributional RL. We consider whether several of our primary results can be mimicked by any of these null models. In this section, we will evaluate these null models in the context of the variable-magnitude task (described in the main text).

One important assumption underlying most of these null models is that reward magnitudes are *range normalized* before entering into TD computations. Range normalization is a very well established feature of reward learning[18–21], and therefore would appear to constitute an unobjectionable modeling assumption. Importantly, range normalization alone cannot account for our observation that, across neurons, the set of reversal points is distributed to match the range of reward magnitudes (Figure 2a,b in main text). This is because, in classic TD, even if the rewards are range-normalized, the learned reward prediction will converge on the mean; there is no mechanism for any form of noise, superimposed on this learned mean, to generate between-cell firing variability that matches the range of possible rewards.

2.1 Model 1: RPE+noise

One important alternative model to rule out is one that simply adds noise to the standard TD RPE:

$$DA = r + \gamma V(x') - V(x) + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma),$$

This model can be excluded based on its failure to explain our findings of (1) reliably diverse reversal points and (2) correlation between reversal point and response asymmetry.

In the analyses reported in the main text, we measured the reliability of reversal-point diversity by measuring reversal points separately in two disjoint halves of the data and confirming a correlation across halves. Another way to assess the reliability of the differences between cells is by model comparison. We performed a leave-one-trial-out cross-validation procedure on the dopamine data. For each held-out trial, we fit two models to the remaining data. One model estimated reversal points for each cell using the same procedure described in the main text. The other model proceeded as if all trials came from a single cell, and fit a single reversal point. Each model then predicted whether the left-out trial should have a positive or negative response. The former model correctly predicted significantly more trials (Extended Data Figure 4a,b; p = 3.7e-8 by Fisher exact test).

2.2 Model 2: RPE+bias+noise

One might wonder if the diversity in optimism that we report in the main text could be due not to distributional RL, but simply to stable differences across neurons in the form of a "bias" term, leading to a null model with the following form:

$$DA = r + \gamma V(x') - V(x) + b_i + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma)$$

In the standard TD framework, dopamine firing at each moment reflects the instantaneous prediction error. During the inter-trial interval, in each moment there is a prediction error but it is close to zero because there is essentially zero new information. Therefore, adding a consistent bias term to the RPEs cannot produce the diversity that we reported in the main text. This bias term would be added to both the pre-stimulus baseline and the reward response, and so it would be eliminated when performing baseline-subtraction.

2.3 Model 3: σ(RPE+bias)+noise

The "RPE+bias+noise" model becomes more interesting if a non-linearity is applied to the dopamine response after the cell-specific bias is added, but only if the noise occurs outside the non-linearity,

$$DA = \sigma(r + \gamma V(x') - V(x) + b_i) + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma).$$

We analyzed synthetic data from this model using the same analysis used to generate Figure 2c in the main text. For each simulated DA cell we estimated the reward magnitude at which that cell's response was equal to its baseline firing rate. As with the actual experimental data, we carried out this procedure in two independent subsets of the data. The bias is still removed by baseline-subtraction, leading to no actual change in reversal point, but the effect of noise in the saturating region can produce artifacts. Thus, to our initial surprise, this analysis yielded reliable (though artifactual) between-cell diversity in apparent reversal points (Extended Data Figure 4f). However, this null model is excluded by other aspects of the experimental data. In particular, in frank contradiction to the experimental data, the null model yields a negative correlation between reversal point and response asymmetry; and again unlike the experimental data there is no reward magnitude that gives rise in the model to a situation in which

more than 5% of cells generate positive responses and more than 5% of cells generate negative responses (Extended Data Figure 4g).

Finally, we also considered a variant of this model where the bias and the sigmoid's slope are positively correlated with each other across cells. This variant similarly produced a negative correlation between asymmetry and reversal point (data not shown). The reason is the same as in the plotted version with uncorrelated bias and slope. Although all cells have the same actual reversal point, cells with steep slopes in the positive domain tend to have their reversal points estimated lower because of noise. But this same effect necessarily produces a negative correlation between measured reversal point and measured asymmetric scaling.

2.4 Model 4: RPE+bias$_T$+noise

Another possibility is that a cell-specific bias term is added to the RPE *only* at the time of a reward outcome, and not during the inter-trial interval:

$$DA = r + \gamma V(x') - V(x) + b_i(t) + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma)$$

This model appears physiologically implausible: we don't know of any mechanism for a bias to be added to dopamine firing selectively at particular moments. If this model were true, it could explain the systematic between-cell differences in measured optimism. However, it cannot account for the correlation we observe between asymmetric scaling and reversal point.

As an aside, it is worth noting that this null model shares some properties with distributional RL. In particular, it could be viewed as learning a set of diverse value predictions. If the distribution of returns in the environment were close to uniform, the learned "set of predictions" would be a reasonable approximation of the true distribution. However, in this null model it would not be possible to perform distribution decoding as shown in main text Figure 5.

2.5 Model 5: "One-sided" distributional RL

Distributional RL requires some degree of anatomical coupling between reward prediction (V) cells, and reward prediction error (δ) cells (see Section 2.7 below). Here we examine two null models in which this requirement is relaxed. First, what if there are stable variations in scaling of positive versus negative RPEs, but these are averaged and fed back to the value predictors non-specifically? This null model has the form:

$$\delta_i = r + \gamma V^*(x') - V_i(x)$$

$$DA = \alpha_i^+ \cdot \delta_i \qquad \text{for } \delta_i > 0 \tag{6}$$

$$DA = \alpha_i^- \cdot \delta_i \qquad \text{for } \delta_i \leq 0 \tag{7}$$

$$V_i(x) \leftarrow V_i(x) + \mathbb{E}_i[DA]$$

Under this model, there is no mechanism to produce consistent diversity in value baselines. The averaging of RPEs causes all V_i to converge to the same prediction. Interestingly, if we extend the model to have a saturating non-linear response function, it is possible to produce *apparent* diversity in reversal points (Extended Data Figure 4m). This artifact arises due to a combination of noise and non-linearity in the cell's response function. For example, if a set of cells all listen to the same reward prediction V_i, but have different relative scalings of positive versus negative RPEs, and also have noise in their responses, then they can appear to have consistently different reversal points. Importantly, however, this kind of model cannot explain the positive correlation we observe between reversal point and asymmetric scaling (Figure 4), nor the precise decoding of the shape of the distribution shown in Figure 5.

This model is a special case of the more general null model in which there are stable variations between cells in any parameters controlling the shape of the dopamine response function. We have also explored parameterizing these stable variations using a Hill function where each cell has its own randomly selected parameters controlling the shape of the function. Similarly to before, this model produces apparent diversity in reversal points but not positive correlations with asymmetry (data not shown).

Second, what if there are there are stable variations in scaling of positive versus negative RPEs, *and* there are stable variations in the value baselines that set the zero point for each dopamine neuron, but

nevertheless the resulting RPEs are averaged and fed back to the value predictors non-specifically? Note that this model is, in fact, a form of distributional RL, albeit a different one than we have proposed. This form of the theory would explain the diversity of reversal points spanning the range of reward magnitudes, but it would not explain the correlation between asymmetric scaling and reversal point shown in Figure 5. However, such a mechanism could co-exist with asymmetric scaling of firing rates; we discuss that possibility in the next section.

2.6 Model 6: Distributional RL implemented by synaptic asymmetries

In this paper, we have explored a form of distributional RL in which asymmetries between scaling of above-baseline vs below-baseline dopamine firing rates drive learning of optimistic or pessimistic value predictions. However, the math of distributional RL is equally compatible with a neurobiological realization where *firing rates* are scaled equally in the positive versus the negative domain, but *synaptic mechanisms* are asymmetrically scaled downstream of dopamine firing. For example, it could be that some RPE channels have a higher ratio of D1 to D2 receptors, and therefore learn faster from positive versus negative RPEs. This model does not predict a correlation between reversal points and asymmetric scaling of firing rates (Extended Data Figure 4p,q). Therefore, this type of synaptic asymmetry mechanism doesn't seem to be the sole mechanism in operation – however, it is possible that both mechanisms operate, and this is an interesting line for future research. It is notable that to do distributional TD learning in multi-step transition models, one must sample from the value distribution at the new state [6]. The brain could approximate such samples using sparse non-topographic connections.

2.7 Model 7: Coupling between prediction and prediction error

In our main implementation of distributional TD learning, for simplicity we assumed a perfect anatomical mapping such that each dopamine cell projects to one value predictor cell, which in turn projects back to exactly the same dopamine cell. Although there is topographic organization in the basal ganglia, this level of precision seems unlikely to prevail in the brain. Therefore we here examine the consequences of relaxing this 'connectomic' assumption. We find that, as long as value predictor cells project more strongly back to approximately the same region of VTA that they receive RPEs from, the distribution-learning mechanism remains intact (Extended Data Figure 4i-k). In these simulations we computed the weights for each connection between a value predicting cell i and an RPE cell j as

$$w_{ij} \propto \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{|i-j|^2}{2\sigma^2}}$$

where σ controls the degree of precision in the mapping (with $\sigma = 0$ being the case explored in the main text). Even at relatively large values of σ, the characteristics of the distribution are preserved. When $\sigma \to \infty$, the algorithm becomes standard (non-distributional) TD learning.

3 Robustness of Results to Modeling Assumptions

Given the novelty of the data analysis techniques employed in the main text, we conducted a wide range of complementary analyses to establish that qualitative results were not dependent on specific modeling choices, and also to confirm that artifacts could not arise for example from interacting non-linearities in data and model. Results from these analyses – which were in all cases reassuring – are presented in the following sections.

3.1 Mapping from reward volume to utility

In the work presented in the main text, the subjective utility of each reward volume in the variable-magnitude task was modeled following a procedure introduced and validated by Stauffer et al. [22]. Specifically, utility was quantified as the mean (baseline-subtracted) dopaminergic firing rate in response to the relevant reward volume. We repeated the analyses in main text Figure 4 using linear, logarithmic, or power law utility functions in place of the empirically-defined function just described, and obtained the same pattern of results in all cases (results for linear utility shown in Extended Data Figure 5g).

3.2 Measuring asymmetric scaling

As a corroborating measure of reliability in the diversity of asymmetric scaling, we performed model comparison between two models (Extended Data Figure 5b). The 'single asymmetry' model fit an overall scale factor to each cell (applied to both positive and negative domains). It also fit, to all cells together, a single scale factor that applied only to the positive domain, and a single scale factor that applied only to the positive domain. The 'diverse asymmetry' model similarly fit a reversal point and overall scale factor to each cell, but it allowed each cell to have its own positive domain scale factor. Thus, the diverse asymmetry model had an extra N-2 parameters compared to the single asymmetry model, where N is the number of cells. To avoid a potential confound whereby noisy estimation of reversal points could create the appearance of diverse asymmetry, we fixed the reversal points for all cells to the mean reversal point across cells.

We performed cross-validation by fitting the model on M-1 trials, where M is the total number of trials across all cells. We used the fit model to predict dopamine cell firing on the held-out trial, and we recorded the mean squared error (MSE) across all such predictions. We found that the MSE was 3.4% lower for the diverse asymmetry model ($p = 1.4 \times 10^{-11}$), indicating that the additional parameters – allowing each cell to have its own unique asymmetry – were justified by the data.

In order to establish a frame of reference for interpreting this 3% difference in MSE, we simulated 100 TD learners on the variable-magnitude task. The learners had different degrees of 'distributionalness,' ranging from non-distributional models with little variation in positive-to-positive asymmetry, to highly distributional models with a high degree of diversity (Extended Data Figure 5c). Each simulation had 40 RPE channels, and across simulations we varied the degree to which there was *diversity* in asymmetric scaling between channels. Gaussian noise was added to simulations, with a scale estimated from the neural data (although the results are potentially sensitive to the covariance between cells of the noise, we simulate independent noise). For each model, we performed the same model-comparison analysis as above and computed percent difference in MSE for each simulation. We also performed distribution decoding on each simulation and measured the percent difference in Wasserstein distance between the decoded distribution and (1) the true task reward distribution and (2) a Gaussian with the same mean and variance. Plotting these two differences against each other, it was observed that 3.4% difference in firing-rate space translated to a roughly 50% difference in distribution space. The simulations are sensitive to the noise scale, and for these simulations we use a noise scale estimated from the real dopamine data. Finally, we note that when the simulation uses a non-distributional TD learner, it is possible for the simpler model to achieve a better MSE than the more complex model, because the model comparison is performed in cross-validation.

3.3 Measuring 'optimism' in variable-probability task

In Figure 3 of the main text, we showed simultaneous optimistic and pessimistic probability coding. In panels a, b, and e of that figure, responses (or simulated responses), were sorted into three groups, based upon two single-tailed Mann-Whitney tests comparing the per-trial responses of 10% to 50% and 50% to 90% cues. Extended Data Figure 7a provides an illustration of the outcomes for these two tests and how they determine the groups.

The individual line for each (simulated) neuron was colored based upon the associated t-statistic, which was also shown in the histograms next to each. Notice that distributional RL does predict that some neurons should be neutral, that is, close to the mean. This can be understood as a consequence of the asymmetric regression depicted in Extended Data Figure 1a. Classic TD predicts that no neurons should learn a reward prediction systematically different from the mean. This means that under classic TD we would expect to see only 5% of neurons significantly different from the mean at a $p = 0.05$ threshold. Equivalently, we would expect the distribution of t-statistics to be t-distributed, which is not observed in the neural data.

3.4 Response functions and the asymmetry/reversal-point correlation

In Figure 5e, we report a positive correlation between $\alpha^+/(\alpha^+ + \alpha^-)$ and reversal point. One important choice in this analysis involved the function used to fit dopaminergic tuning curves. In this section, we consider the sensitivity of the correlation result to this choice.

There is an interesting effect whereby if the dopamine response function (i.e., the function that converts a reward prediction error into a firing rate) is more concave than the function used for fitting, then the analysis can produce false positives (Extended Data Figure 5d). This is because at higher

reversal points, the positive data are on average closer to the reversal point, and therefore lie in the steeper part of the concave function. (Likewise the negative data are on average further from the reversal point and so lie toward the shallower part of the function.) On the other hand, note that if a real positive correlation exists in the data, then our simulations show that it is detectable under any combination of generating and fitting functions (Extended Data Figure 5e).

In Figure 4, we used linear functions to estimate slopes in the positive and negative domains. So is it possible that this form of artifact is responsible for the correlation that we measured in the neural data? This appears highly unlikely for the following reasons. First, the neural data themselves are close to piecewise linear (i.e., the positive domain is linear and the negative domain is linear). For example, when we performed nested model comparison (by likelihood ratio test) between linear and quadratic models, separately for the negative and positive domains of each cell, we found only 6/79 tests favored concavity (the positive domain in one cell could not be estimated as there were no positive responses).

Second, in Extended Data Figure 5f we re-fit the neural data using a very convex function, the Hill function (also used in Eshel et al. [23]). Reassuringly, we found that the correlation between $\alpha^+/(\alpha^+ + \alpha^-)$ and reversal point remained significant ($r = 0.41, p = 0.006$). Importantly, this function is more convex than the neural data themselves, which strongly guards against the possibility of the aforementioned artifact. Together, these observations provide reassurance that the results presented in the main text are not artifactually arising from non-linearities in the data.

3.5 Negative RPEs and pausing

Negative reward prediction errors may be coded by the length of pausing in dopamine cell firing [24]. In our analyses, we averaged firing rates over a relatively long time window in order to account for both firing rate changes and durations of pauses. For completeness, we also performed the same analyses in the variable-magnitude task using a longer window of 200 to 1000 ms, which might capture the longest pauses (with the tradeoff of including more noise). The key results of the paper still held: (1) Reversal point estimated in half of the data was correlated with reversal point estimated in the other half of data ($p = 1.3 \times 10^{-4}$ by linear regression, geometric mean across 1000 random partitions). (2) Different cells had different asymmetric scalings ($p = 3.9 \times 10^{-16}$ by ANOVA). (3) Asymmetric scaling was positive correlated with reversal point ($p = 2.6 \times 10^{-6}$ by linear regression).

3.6 Distribution decoding

Distribution decoding is the result of solving a minimization problem defined by taking the distributional TD model as a loss function between (1) reversal points, (2) asymmetries, and (3) samples from the reward distribution. Thus, the minimization problem yields a distribution decoding by fixing the paired reversal point estimates Z_{τ_i} and asymmetries $\tau_i = \alpha_i^+/(\alpha_i^+ + \alpha_i^-)$, and solving for the sampling distribution that minimizes the loss.

For N the number of estimated reversal points, and $M = 100$ the number of samples z_m to decode,

$$\begin{aligned} &\underset{z_1,\ldots,z_m}{\arg\min}\, \mathcal{L}(z, V, \tau), \\ \mathcal{L}(z, V, \tau) &= \frac{1}{M} \sum_{m=1}^{M} \sum_{n=1}^{N} |\tau_n - \mathbb{I}_{z_m \le V_n}| (z_m - V_n)^2. \end{aligned}$$

We solved this minimization problem using the SciPy 1.2.1 minimize function using the truncated Newton method, with solutions constrained to fall in the total range of observed utilities. This decoding method was recently suggested, with the decoding denoted as the *imputed distribution*, as a general method for deriving distributional RL algorithms [6].

To overcome noise in our estimates of reversal points and asymmetries we generated 20000 random samples for $(z_1, \ldots, z_m)$, uniform between the range of rewards, and initialized the minimization problem at the sample points with the smallest loss.

Extended Data Figure 8 shows our investigation into the robustness of this decoding procedure. Figure 5c in the main text shows the value distribution decoded from dopaminergic responses in the variable-magnitude experiment. The RPE asymmetry for each neuron was interpreted, in line with the distributional RL model (2), as defining the expectile of the value distribution encoded by that neuron, and the above procedure was used to back out the distribution that best fit the data, under the distributional RL model. The best-fit distribution closely resembles a kernel-density estimate of the true distribution, based on the rewards actually received by experimental animals.

Critically, the decoding depended on the specific pattern of reversal points and RPE asymmetries observed. In Extended Data Figure 8e-h we show that when asymmetries were non-linearly transformed, the fidelity of decoding was seriously compromised. In order to verify that the decoding objective was sensitive to the modes of the empirical distribution we evaluate the loss against reference distributions (Extended Data Figure 8d). We consider the uniform and Gaussian distributions with ground truth mean and variances, "mirroring" of the true return distribution about its mean, and perturbations of the mean and variance of the ground truth reward distribution. This shows that the decoding loss is highly sensitive not only to the first two moments of the distribution, but also to structure such as multimodality. Note, as this is a loss function, lower values indicate better fits. Finally, in Extended Data Figure 8a-c, we apply our decoding analysis without constraining the resulting reward values, which produces qualitatively similar results.

For the variable-magnitude task we use the reversal points and asymmetries estimated on each dopamine neuron by first averaging responses over trials for each reward magnitude, as opposed to the split-halves testing used elsewhere. For the variable-probability task we use the normalized responses reported in the Figure 3. Because we cannot estimate asymmetries for this task we used the observation of correlation in Figure 4 to approximate the asymmetries here as,

$$\tau_n \approx \frac{V_n(x_{50}) - \min_i V_i(x_{50})}{\max_i V_i(x_{50}) - \min_i V_i(x_{50})},$$

where $V_n(x_{50})$ denotes the normalized response to the 50% cue for neuron n. This sacrifices some decoding accuracy, and is unlikely to work well for more complex reward distributions.

When evaluating the loss function $\mathcal{L}$ against a reference distribution, we generate $M = 2000$ samples from the distribution and compute the loss function for the estimated reversal points and asymmetries. When testing the sensitivity to the pairing of reversal points and asymmetries, we computed monotonic transformations of the values τ_n, for a shift of $s \in [-2.5, 2.5]$,

$$\hat{\tau}_n = \phi(\phi^{-1}(\tau_n) + s), \tag{8}$$

where ϕ is the standard normal cumulative distribution functions (see Extended Data Figure 8e).

4 Supplementary Results

4.1 Both positive and negative responses to 5μL reward

A prediction of the distributional model, but not of some null models, is that in response to a single reward magnitude in the variable-magnitude task, there should be some cells that reliably fire positive RPEs and others that reliably fire negative RPEs. This prediction is borne out in the dopamine data (Extended Data Figure 9g).

4.2 Optimism and pessimism in variable-probability task

In Extended Data Figure 7, we show the firing rate data used to calculate the t-statistics displayed in the histograms in Figure 3b of the main text.

4.3 Simultaneous recordings support diversity

Most of the cells in the data sets were not simultaneously recorded. This raises the question of whether apparent diversity in reward predictions across cells could actually reflect session-to-session differences. However, in one instance in the variable-probability task there were four simultaneously recorded cells. These are shown in main text Figure 3c, and in Extended Data Figure 9e.

For statistical purposes, we evaluated the quantity c_{50}: the response to the 50% cue as a proportion of the distance between the mean response to the 10% cue and the mean response to the 90% cue. This quantity could be computed for each trial when the 50% cue was delivered. One-way ANOVA rejected the null hypothesis that all four cells had the same mean c_{50} (p = 1.4e-7).

4.4 Relationship between cue-time and outcome-time responses

In Extended Data Figure 6a-b we use the variable-probability task to illustrate the relationship between cue- and outcome-response in the distributional TD model, and show how this compares with normalized firing rates in the neural data,

$$c_{cue}^{norm} = (c_{cue} - mean(c_{10}, c_{50}, c_{90}))/std(c_{10}, c_{50}, c_{90}).$$

The cause of the complicated relationship, which is hard to see in noisy data, is that raw prediction errors at delivery time are smallest for *optimistic* cells, where the scaling is largest, and largest for *pessimistic* cells, where scaling is smallest. This produces a natural canceling-out effect that removes much of the correlation between cue and outcome prediction errors we would otherwise expect to see. Despite this rather strange relationship, noisy simulations closely resemble actual (normalized) neural firing rates.

4.5 Further examples of diversity in single-cell RPE signalling

In main text Figure 2d we showed rasters of two cells that had particularly dramatic (positive and negative) responses. We noticed that the example cell with the positive response also appeared to have a small dip afterwards. This was not a general characteristic of the data. For example, here are two more cells from the same animal, with positive responses and without dips (Extended Data Figure 9h).

4.6 Asymmetric scaling as a function of baseline firing rate

A priori it seems possible that cells with a lower baseline firing rate could have a shallower slope in the negative domain because they cannot fire below 0 Hz. Thus, cells with a lower baseline firing rate might be expected to have a larger value of $\alpha^+/(\alpha^+ + \alpha^-)$. If this were true, then the variability in baseline firing rates between cells could be a source of the variability in asymmetric scaling that we observed in the variable-magnitude task. This would not conflict with the distributional RL theory, but would suggest a mechanism for the origin of variability. To assess this possibility, we looked at the relationship between baseline firing rate and asymmetric scaling (Extended Data Figure 9f). We found no relationship.

4.7 Licking data

In the datasets analyzed, most recording sessions consisted of a single cell. This raises the question of whether apparent diversity between cells (in terms of optimism) could actually arise from between-session (or between-animal) differences in preferences at the whole-organism level. As mentioned in the main text, anticipatory licking data argues against this interpretation, because between-cell differences in optimism (measured in the variable-probability task) were not predicted by between-session differences in anticipatory licking (Extended Data Figure 9a-d).

5 Relationship to past work

Previous work has modeled how temporal difference learning in humans and animals can incorporate epistemic uncertainty[25,26]. It is also well-established experimentally that neural circuits for economic decision making have access to higher-order statistics of the outcome distribution, including variance and skewness[27–34]. Finally, methods have been proposed to learn value functions with specific risk preferences[35]. Our proposal goes beyond these to explain how the full distribution of outcomes is learned through a remarkably simple mechanism implemented in dopamine cells. We also consider the relationship to the findings of specific papers below.

Eshel et al., 2016 Eshel et al.[23] observed that the response functions of individual dopamine neurons are well-predicted as uniformly scaled copies of the mean (across neurons) response function. In the present work, we have used the same data to argue for *diversity* in response functions, specifically, diversity in the relationship of scaling of the positive domain to scaling of the negative domain. How can these two perspectives be compatible?

In Extended Data Figure 10a-h, we reproduce the analysis of Eshel et al.[23]. Note that these analyses are performed using the time window from 0 to 600 ms after reward onset (illustrated by dotted lines in panel a). In Extended Data Figure 10d, we plot the same data as in panel c, but zoomed in to show

that the predictability of the dopamine response breaks down in the negative domain. (The predictions are based on the entire range, and the positive domain dominates.) This is also reflected in the fact that only 80% of the total variance in responses is explained by the common response function.

In Extended Data Figure 10e-h, we show exactly the same analyses as in panels a-d, but using data from the time window 200 to 600 ms after reward onset. This is the time window we have used for analyses in the main text, because it allows us to establish "reversal points" by excluding the large initial positive response[36]. In this time window, the common response function predicts only 40% of the variance in the responses.

In conclusion, there are two factors explaining the relationship between our results and those of Eshel et al. First, even using the 0 to 600 ms time window used by Eshel et al, there is variance unexplained by a common response function, which we argue corresponds to different cells having different relative scalings from positive and negative RPEs. Second, this diversity is exaggerated when using a time window that permits measurement of negative responses to below-expected reward magnitudes.

Fiorillo et al. (2013), Matsumoto et al. (2016) One surprising finding of this work was the existence of within-animal diversity in the asymmetric response to positive versus negative outcomes. To further support these findings, we consider a simple free-reward/airpuff experiment. Past work using similar tasks has shown remarkable diversity in the response asymmetry to ostensibly positive versus negative prediction errors. In particular, Fiorillo et al. investigated diversity in response to a variety of rewarding, and aversive, stimuli. Although unremarked upon at the time, their findings show substantial differences between neurons in the asymmetry of response (not just in their overall magnitude of response). Similarly, Matsumoto et al., while studying the effect of reward-context on response to aversive stimuli, found precisely the type of diverse asymmetry we have been discussing.

However, in these previous studies results are presented aggregated over animals, leaving uncertainty as to whether this diverse asymmetry is (as we have already found in this work) present within single animals, or only between animals. In Extended Data Figure 6c we show that, indeed, even within individual animals the results seen in these past works are reproduced. Finally, observe in Extended Data Figure 6d that distributional TD, but not classical TD, predicts similar types of asymmetric diversity.

6 Predictions

Distributional RL makes many other interesting predictions, a few of which we briefly highlight:

- The degree of optimism for a dopamine cell should be persistent between different tasks, even while the corresponding reversal point changes.
- During learning, the degree of optimism measured for a dopamine cell should predict how fast a cell changes its reward prediction in response to positive versus negative prediction errors.
- In particular, optimistic cells should be slower, relative to pessimistic cells, to devalue.
- The *blocking* phenomenon should be affected by the distribution of rewards, due to distributional TD errors persisting even if the mean is well-predicted.
- Inputs to dopaminergic neurons should should show preferential responses that relate to the degree of optimism in their downstream dopaminergic neurons.
- If risk-sensitive behavior is driven by the dopamine-based distribution of returns, then risk-sensitivity could be induced due to a changing task. That is, a changing task should induce behavior that mimicks risk-sensitivity despite being deterministic.
- Generalization benefits, of the type illustrated in Extended Data Figure 3, should be seen when the change in probabilities are due to changing behavior as well as changing task. This effect could show up by differences in how quickly an animal adapts behavior, where previous training on related stochastic tasks enhances adaptation to new tasks.

References

[1] Richard S Sutton and Andrew G Barto. *Reinforcement learning: an introduction*, volume 1. MIT press Cambridge, 1998.

[2] Will Dabney, Mark Rowland, Marc G Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. *AAAI Conference on Artificial Intelligence*, 2018.

[3] Michael Bowling and Manuela Veloso. Multiagent learning using a variable learning rate. *Artificial Intelligence*, 136(2):215–250, 2002.

[4] Roger Koenker and Kevin F Hallock. Quantile regression. *Journal of economic perspectives*, 15(4): 143–156, 2001.

[5] Peter J Huber. Robust statistics. *Wiley Series in Probability and Mathematical Statistics*, 1981.

[6] Mark Rowland, Robert Dadashi, Saurabh Kumar, Rémi Munos, Marc G Bellemare, and Will Dabney. Statistics and samples in distributional reinforcement learning. In *International Conference on Machine Learning*, 2019.

[7] Felipe Petroski Such, Vashisht Madhavan, Rosanne Liu, Rui Wang, Pablo Samuel Castro, Yulun Li, Ludwig Schubert, Marc Bellemare, Jeff Clune, and Joel Lehman. An atari model zoo for analyzing, visualizing, and comparing deep reinforcement learning agents. *arXiv preprint arXiv:1812.07069*, 2018.

[8] Pablo Samuel Castro, Subhodeep Moitra, Carles Gelada, Saurabh Kumar, and Marc G Bellemare. Dopamine: a research framework for deep reinforcement learning. *arXiv preprint arXiv:1812.06110*, 2018.

[9] Shangtong Zhang, Borislav Mavrin, Linglong Kong, Bo Liu, and Hengshuai Yao. QUOTA: The quantile option architecture for reinforcement learning. In *AAAI Conference on Artificial Intelligence*, 2019.

[10] Nikolay Nikolov, Johannes Kirschner, Felix Berkenkamp, and Andreas Krause. Information-directed exploration for deep reinforcement learning. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=Byx83s09Km`.

[11] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

[12] William Fedus, Carles Gelada, Yoshua Bengio, Marc G Bellemare, and Hugo Larochelle. Hyperbolic discounting and learning over multiple horizons. *arXiv preprint arXiv:1902.06865*, 2019.

[13] Mohammad Gheshlaghi Azar, Bilal Piot, Bernardo Avila Pires, Jean-Bastian Grill, Florent Altché, and Rémi Munos. World discovery models. *arXiv preprint arXiv:1902.07685*, 2019.

[14] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[16] Dino J Levy, Amalie C Thavikulwat, and Paul W Glimcher. State dependent valuation: the effect of deprivation on risk preferences. *PloS one*, 8(1):e53978, 2013.

[17] Alex Kacelnik and Melissa Bateson. Risk-sensitivity: crossroads for theories of decision-making. *Trends in cognitive sciences*, 1(8):304–309, 1997.

[18] Kenway Louie, Mel W. Khaw, and Paul W. Glimcher. Normalization is a general neural mechanism for context-dependent decision making. *Proceedings of the National Academy of Sciences*, 110(15):6139–6144, 2013. doi: 10.1073/pnas.1217854110. URL `https://www.pnas.org/content/110/15/6139`.

[19] Philippe N Tobler, Christopher D Fiorillo, and Wolfram Schultz. Adaptive coding of reward value by dopamine neurons. *Science*, 307(5715):1642–1645, 2005.

[20] Neil Stewart, Nick Chater, Henry P Stott, and Stian Reimers. Prospect relativity: how choice options influence decision under risk. *Journal of Experimental Psychology: General*, 132(1):23, 2003.

[21] Camillo Padoa-Schioppa. Range-adapting representation of economic value in the orbitofrontal cortex. *Journal of Neuroscience*, 29(44):14004–14014, 2009.

[22] William R Stauffer, Armin Lak, and Wolfram Schultz. Dopamine reward prediction error responses reflect marginal utility. *Current biology*, 24(21):2491–2500, 2014.

[23] Neir Eshel, Ju Tian, Michael Bukwich, and Naoshige Uchida. Dopamine neurons share common response function for reward prediction error. *Nature neuroscience*, 19(3):479–486, 2016.

[24] Hannah M Bayer, Brian Lau, and Paul W Glimcher. Statistics of midbrain dopamine neuron spike trains in the awake primate. *Journal of Neurophysiology*, 98(3):1428–1439, 2007.

[25] Nathaniel D Daw, Yael Niv, and Peter Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, 8(12):1704, 2005.

[26] Samuel J Gershman. A unifying probabilistic view of associative learning. *PLoS computational biology*, 11(11):e1004567, 2015.

[27] Christopher D Fiorillo, Philippe N Tobler, and Wolfram Schultz. Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, 299(5614):1898–1902, 2003.

[28] Wolfram Schultz. Dopamine signals for reward value and risk: basic and recent data. *Behavioral and brain functions*, 6(1):24, 2010.

[29] Kerstin Preuschoff, Peter Bossaerts, and Steven R Quartz. Neural differentiation of expected reward and risk in human subcortical structures. *Neuron*, 51(3):381–390, 2006.

[30] Christopher Trepel, Craig R Fox, and Russell A Poldrack. Prospect theory on the brain? toward a cognitive neuroscience of decision under risk. *Cognitive brain research*, 23(1):34–50, 2005.

[31] Mkael Symmonds, Nicholas D Wright, Dominik R Bach, and Raymond J Dolan. Deconstructing risk: separable encoding of variance and skewness in the brain. *Neuroimage*, 58(4):1139–1149, 2011.

[32] Christopher J Burke and Philippe N Tobler. Reward skewness coding in the insula independent of probability and loss. *Journal of neurophysiology*, 106(5):2415, 2011.

[33] Yael Niv, Jeffrey A Edlund, Peter Dayan, and John P O'Doherty. Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *Journal of Neuroscience*, 32(2): 551–562, 2012.

[34] Ilya E Monosov and Okihide Hikosaka. Selective and graded coding of reward uncertainty by neurons in the primate anterodorsal septal region. *Nature neuroscience*, 16(6):756, 2013.

[35] Oliver Mihatsch and Ralph Neuneier. Risk-sensitive reinforcement learning. *Machine learning*, 49 (2-3):267–290, 2002.

[36] Wolfram Schultz. Dopamine reward prediction-error signalling: a two-component response. *Nature Reviews Neuroscience*, 17(3):183, 2016.

[37] Christopher D Fiorillo, Sora R Yun, and Minryung R Song. Diversity and homogeneity in responses of midbrain dopamine neurons. *Journal of Neuroscience*, 33(11):4693–4709, 2013.

[38] Hideyuki Matsumoto, Ju Tian, Naoshige Uchida, and Mitsuko Watabe-Uchida. Midbrain dopamine neurons signal aversion in a reward-context-dependent manner. *Elife*, 5, 2016.