Article

A distributional code for value in dopamine-based reinforcement learning

https://doi.org/10.1038/s41586-019-1924-6

Received: 3 January 2019

Accepted: 19 November 2019

Published online: 15 January 2020

Will Dabney[1,5*], Zeb Kurth-Nelson[1,2,5], Naoshige Uchida[3], Clara Kwon Starkweather[3], Demis Hassabis[1], Rémi Munos[1] & Matthew Botvinick[1,4,5]

Since its introduction, the reward prediction error theory of dopamine has explained a wealth of empirical phenomena, providing a unifying framework for understanding the representation of reward and value in the brain[1–3]. According to the now canonical theory, reward predictions are represented as a single scalar quantity, which supports learning about the expectation, or mean, of stochastic outcomes. Here we propose an account of dopamine-based reinforcement learning inspired by recent artificial intelligence research on distributional reinforcement learning[4–6]. We hypothesized that the brain represents possible future rewards not as a single mean, but instead as a probability distribution, effectively representing multiple future outcomes simultaneously and in parallel. This idea implies a set of empirical predictions, which we tested using single-unit recordings from mouse ventral tegmental area. Our findings provide strong evidence for a neural realization of distributional reinforcement learning.

The reward prediction error (RPE) theory of dopamine derives from work in the artificial intelligence (AI) field of reinforcement learning (RL)[7]. Since the link to neuroscience was first made, however, RL has made substantial advances[8,9], revealing factors that greatly enhance the effectiveness of RL algorithms[10]. In some cases, the relevant mechanisms invite comparison with neural function, suggesting hypotheses concerning reward-based learning in the brain[11–13]. Here we examine a promising recent development in AI research and investigate its potential neural correlates. Specifically, we consider a computational framework referred to as distributional reinforcement learning[4–6] (Fig. 1a, b).

Similar to the traditional form of temporal-difference RL—on which the dopamine theory was based—distributional RL assumes that reward-based learning is driven by a RPE, which signals the difference between received and anticipated reward. (For simplicity, we introduce the theory in terms of a single-step transition model, but the same principles hold for the general multi-step (discounted return) case; see Supplementary Information.) The key difference in distributional RL lies in how 'anticipated reward' is defined. In traditional RL, the reward prediction is represented as a single quantity: the average over all potential reward outcomes, weighted by their respective probabilities. By contrast, distributional RL uses a multiplicity of predictions. These predictions vary in their degree of optimism about upcoming reward. More optimistic predictions anticipate obtaining greater future rewards; less optimistic predictions anticipate more meager outcomes. Together, the entire range of predictions captures the full probability distribution over future rewards (more details in Supplementary Information).

Compared with traditional RL procedures, distributional RL can increase performance in deep learning systems by a factor of two or more[5,14,15], an effect that stems in part from an enhancement of representation learning (see Extended Data Figs. 2, 3 and Supplementary Information). This prompts the question of whether RL in the brain might leverage the benefits of distributional coding. This question is encouraged both by the fact that the brain utilizes distributional codes in numerous other domains[16], and by the fact that the mechanism of distributional RL is biologically plausible[6,17]. Here we tested several predictions of distributional RL using single-unit recordings in the ventral tegmental area (VTA) of mice performing tasks with probabilistic rewards.

Value predictions vary among dopamine neurons

In contrast to classical temporal-difference (TD) learning, distributional RL posits a diverse set of RPE channels, each of which carries a different value prediction, with varying degrees of optimism across channels. (Value is formally defined in RL as the mean of future outcomes, but here we relax this definition to include predictions about future outcomes that are not necessarily the mean.) These value predictions in turn provide the reference points for different RPE signals, causing the latter to also differ in terms of optimism. As a surprising consequence, a single reward outcome can simultaneously elicit positive RPEs (within relatively pessimistic channels) and negative RPEs (within more optimistic ones).

This translates immediately into a neuroscientific prediction, which is that dopamine neurons should display such diversity in 'optimism'. Suppose an agent has learned that a cue predicts a reward whose magnitude will be drawn from a probability distribution. In the standard RL theory, receiving a reward with magnitude below the mean of this distribution will elicit a negative RPE, whereas larger magnitudes will elicit positive RPEs. The reversal point—the magnitude at which prediction errors transition from negative to positive—in standard RL is the expectation of the magnitude's distribution. By contrast, in

[1]DeepMind, London, UK. [2]Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London, London, UK. [3]Center for Brain Science, Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA, USA. [4]Gatsby Computational Neuroscience Unit, University College London, London, UK. [5]These authors contributed equally: Will Dabney, Zeb Kurth-Nelson, Matthew Botvinick. *e-mail: wdabney@google.com

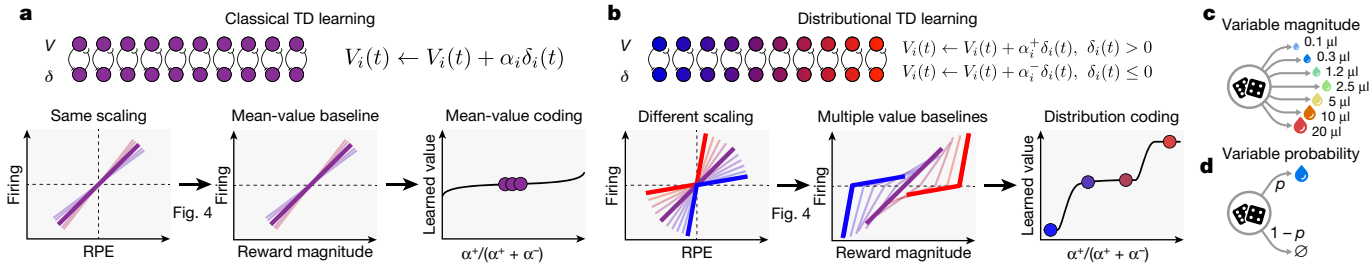


Fig. 1 | Distributional value coding arises from a diversity of relative scaling of positive and negative prediction errors. a, In the standard temporal-difference (TD) theory of the dopamine system, all value predictors learn the same value *V*. Each dopamine cell is assumed to have the same relative scaling for positive and negative RPEs (left). This causes each value prediction (or value baseline) to be the mean of the outcome distribution (middle). Dotted lines indicate zero RPE or pre-stimulus firing. **b**, In our proposed model, distributional TD, different channels have different relative scaling for positive (α^+) and negative (α^-) RPEs. Red shading indicates $\alpha^+ > \alpha^-$, and blue shading indicates $\alpha^- > \alpha^+$. An imbalance between α^+ and α^- causes each channel to learn a different value prediction. This set of value predictions collectively represents the distribution over possible rewards. **c**, We analyse data from two tasks. In the variable-magnitude task, there is a single cue, followed by a reward of unpredictable magnitude. **d**, In the variable-probability task, there are three cues, which each signal a different probability of reward, and the reward magnitude is fixed.

distributional RL, the reversal point differs across dopamine neurons according to their degree of optimism.

We tested for such reversal-point diversity in optogenetically verified dopaminergic VTA neurons, focusing on responses to receipt of liquid rewards, the volume of which was drawn at random on each trial from seven possible values (Fig. 1c). As anticipated by distributional RL, but not by the standard theory, we found that dopamine neurons had substantially different reversal points, ranging from cells that reversed between the smallest two rewards to cells that reversed between the largest two rewards (Fig. 2a, b). This diversity was not owing to noise, as the reversal point estimated on a random half of the data was a robust predictor of the reversal point estimated on the other half of the data ($R = 0.58$, $P = 1.8 \times 10^{-5}$ by linear regression; Fig. 2c). In fact, in response to the 5 µl reward, 13 out of 40 cells had significantly above-baseline responses and 10 out of 40 cells had significantly below-baseline responses. Note that while some cells appeared pessimistic and others appeared optimistic, there was also a population of cells with approximately neutral responses, as predicted by the distributional RL model (compare with Fig. 2a, right).

A stronger test of our theory is whether this diversity also exists within a single animal. Most animals had too few cells for analysis, but within the single animal with the highest number of recorded cells, reversal points estimated on half of the data were robustly predictive of reversal points estimated on the other half ($P = 0.008$). Furthermore, in response to a single reward magnitude (5 µl), 6 out of 16 cells had significantly above-baseline responses and 5 out of 16 cells had significantly below-baseline responses. Finally, Fig. 2d shows rasters of two example cells from this animal, exhibiting consistently opposite responses to the same reward.

Because the diversity we observe is reliable across trials, it cannot be explained by adding measurement noise to non-distributional TD models. As detailed in section 2 of the Supplementary Information (see also Extended Data Fig. 4), we also analysed several more elaborate alternative models, and whereas some of these can give rise to the appearance of reversal-point diversity under some analysis methods, the same models are contradicted by other aspects of the experimental data, which we report below.

Our first prediction dealt with the relationship between dopaminergic signalling and reward magnitude; dopaminergic RPE signals also scale with reward probability[2,18], and distributional RL also leads to a prediction in this domain. Pursuing this, we analysed data from a second task in which sensory cues indicated the probability of an upcoming liquid reward (Fig. 1d). One cue indicated a 10% probability of reward, a different cue indicated a 50% probability, and a third a 90% probability. The standard RPE theory predicts that, considering responses at the time the cue is presented, all dopamine neurons should have the same relative spacing between 10%, 50% and 90% cue responses. (Under neutral risk preferences, the 50% cue response should be midway between the 10% and 90% cues. Under different risk preferences, the 50% cue response might be at a different position between 10% and 90%, but it should be the same for all neurons). Distributional RL predicts, instead, that dopamine neurons should vary in their responses to the 50% cue: some neurons should respond optimistically, emitting a RPE nearly as large as to the 90% cue. Others should respond pessimistically, emitting a RPE closer to the 10% cue response (Fig. 3a). Labelling these two cases as optimistically and pessimistically biased, respectively, distributional RL predicts that as a population, dopamine neurons should show concurrent optimistic and pessimistic coding for reward probability.

To test this prediction, we analysed responses of dopaminergic VTA neurons in the cued probability task just described (see Methods for more details). As predicted by distributional RL, but not by the standard theory, dopamine neurons differed in their patterns of response across the three reward-probability cues, with both optimistic and pessimistic probability coding observed (Fig. 3b left, Extended Data Figs. 6, 7). Again, this diversity was not due to noise, as 10 out of 31 cells were significantly optimistic and 9 out of 31 cells were significantly pessimistic, at a $P < 0.05$ threshold (see Methods). By comparison, at a 0.05 threshold, approximately 3 out of 31 cells in a non-distributional TD system are expected by chance to appear either significantly optimistic or pessimistic. At the group level, the null hypothesis of no diversity was rejected by one-way analysis of variance (ANOVA) ($F(30, 3335) = 4.31$, $P = 6 \times 10^{-14}$). Notably, both forms of probability coding were observed side by side in individual animals. In the animal with the largest number of recorded cells, 4 out of 17 cells were consistently optimistic and 5 out of 17 cells were consistently pessimistic. This was also significant by ANOVA ($F(15, 1652) = 4.02$, $P = 3 \times 10^{-7}$).

Because most cells were recorded in different sessions, it was important to examine whether global changes in reward expectations between sessions might explain the observed diversity in optimism. To this end, we analysed patterns of anticipatory licking. Here we found that, although within-session fluctuations in licking were predictive of within-session fluctuations in dopamine cell firing, there was no relationship between optimism and licking on a cell-by-cell basis (Extended Data Fig. 9). This observation makes it unlikely that the diverse responses we observed in dopamine neurons are explained by session-to-session variability in global reward expectation. That interpretation is further undermined by the fact that reversal-point diversity was observed in the one case where several cells were recorded simultaneously in one animal (Fig. 3c and Supplementary Information).

GABAergic neurons make diverse reward predictions

In distributional RL, diversity in RPE signalling arises because different RPE channels listen to different reward predictions, which vary

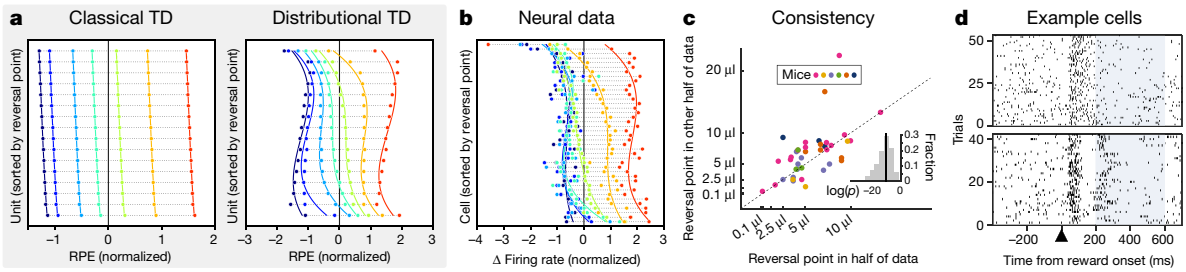


Fig. 2 | Different dopamine neurons consistently reverse from positive to negative responses at different reward magnitudes. Variable-magnitude task from ref. [30]. On each trial, the animal experiences one of seven possible reward magnitudes (0.1, 0.3, 1.2, 2.5, 5, 10 or 20 µl), selected randomly. **a**, RPEs produced by classical and distributional TD simulations. Each horizontal bar is one simulated neuron. Each dot colour corresponds to a particular reward magnitude. The *x* axis is the cell's response (change in firing rate) when reward is delivered. Cells are sorted by reversal point. In classical TD, all cells carried approximately the same RPE signal. Note that the slight differences between cells arose from Gaussian noise added to the simulation; the differences between cells in the classical TD simulation were not statistically reliable. Conversely, in distributional TD, cells had reliably different degrees of optimism. Some responded positively to almost all rewards, and others responded positively to only the very largest reward. **b**, Responses recorded from light-identified dopamine neurons in behaving mice. Neurons differed markedly in their reversal points. **c**, To assess whether this diversity was reliable, we randomly partitioned the data into two halves and estimated reversal points independently in each half. We found that the reversal point estimated in one half was correlated with that estimated in the other half ($P = 1.8 \times 10^{-5}$ by linear regression). **d**, Spike rasters for two example dopamine neurons from the same animal, showing responses to all trials when the 5 µl reward was delivered. We analysed data from 200 to 600 ms after reward onset (highlighted), to exclude the initial transient that was positive for all magnitudes. During this epoch, the cell on the bottom fires above its baseline rate, while the cell on the top pauses.

in their degree of optimism. From a neuroscientific perspective, it should thus be possible to track the effects we have identified at the level of VTA dopamine neurons back to upstream neurons signalling reward predictions. Previous work strongly suggests that VTA GABAergic (γ-aminobutyric acid) neurons have precisely this role, and that the reward prediction used to compute the RPE is reflected in their firing rates[19]. Therefore, we predicted that, in the same task described above, the population of VTA GABAergic neurons should also contain concurrent optimistic and pessimistic probability coding. As predicted, consistent differences in probability coding were observed across putative GABAergic neurons, again with concurrent optimism and pessimism (Fig. 3b, right). In the animal with the largest number of cells recorded, 12 out of 36 cells were consistently optimistic and 11 out of 36 cells were consistently pessimistic (example cells shown in Fig. 3d).

Distribution coding from asymmetric RPE scaling

The results reported in the preceding sections suggest that a distribution of value predictions is coded in the neural circuits underlying RL. How might such coding arise in the first place? Recent AI work on distributional RL[15] has shown that distributional coding arises automatically if a single change is made to the classical TD learning mechanism.

In classical TD, positive and negative errors are given equal weight. As a result, positive and negative errors are in equilibrium when the learned prediction equals the mean of the reward distribution. Therefore, classical TD learns to predict the average over future rewards.

By contrast, in distributional TD, different RPE channels place different relative weights on positive versus negative RPEs (see Fig. 1b). In channels that overweight positive RPEs, reaching equilibrium requires these positive errors to become less frequent, so the learning dynamics converge on a

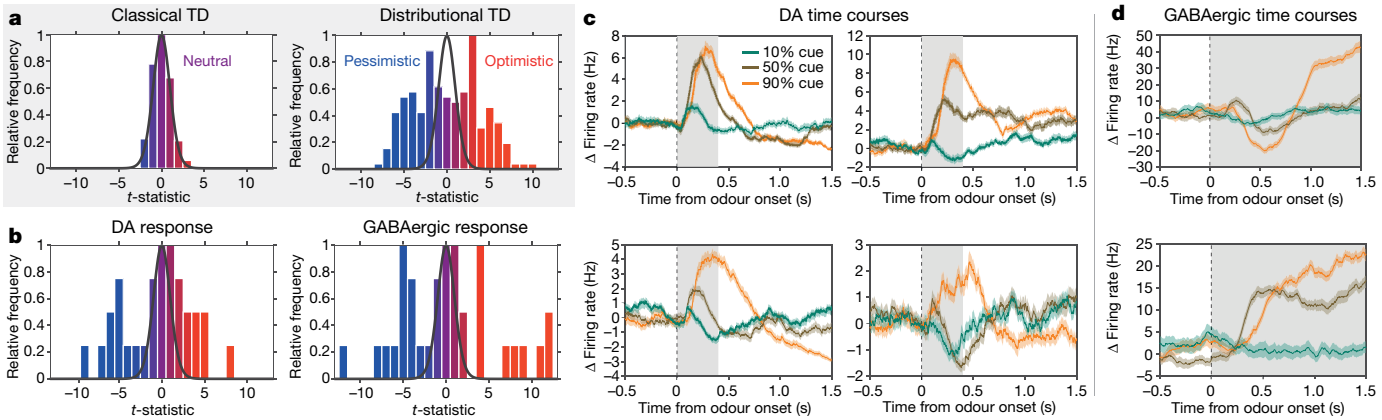


Fig. 3 | Optimistic and pessimistic probability coding occur concurrently in dopamine and VTA GABAergic neurons. Data from variable-probability task. **a**, Histogram (across simulated cells) of *t*-statistics which compare each cell's 50% cue response against the mean 50% cue response across cells. Qualitatively identical results hold when comparing the 50% cue response against the midpoint of 10% and 90% responses. The superimposed black curve shows the *t*-distribution with the corresponding degrees of freedom. Distributional TD predicts simultaneous optimistic and pessimistic coding of probability, whereas classical TD predicts that all cells have the same coding. Colour indicates the degree of optimism or pessimism. **b**, Same as **a**, but using data from real dopamine and putative GABAergic neurons. The pattern of results closely matches the predictions from the distributional TD model. **c**, Responses of four example dopamine neurons recorded simultaneously in a single animal. Each trace is the average response to one of the three cues. Shaded area shows s.e.m. Time zero is the onset of the odour cue. Some cells code the 50% cue similarly to the 90% cue, while others simultaneously code it similarly to the 10% cue. Grey areas show epoch averaged for summary analyses. **d**, Responses of two example VTA GABAergic cells from the same animal.

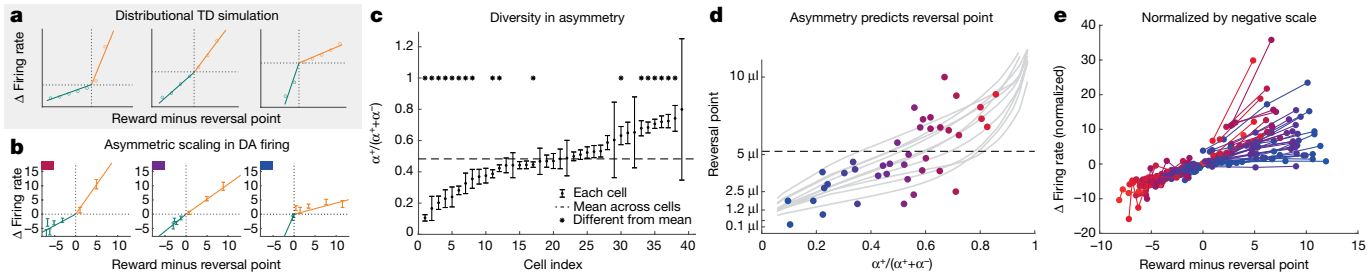


Fig. 4 | Relative scaling of positive and negative dopamine responses predicts reversal point. a, Three simulated dopamine neurons—each with a different asymmetry—in the variable-magnitude task. For each unit, we empirically estimated the reversal point where responses switch from negative to positive. The *x* axis shows reward minus the per-cell reversal point, effectively aligning each cell's responses to its respective reversal point. Baseline-subtracted response to reward is plotted on the *y* axis. Responses below the reversal point are shown in green and those above are shown in orange. Solid curves show linear functions fit separately to the above-reversal and below-reversal domains of each cell. **b**, Same as **a**, but showing three real example dopamine cells. **c**, The diversity in relative scaling of positive and negative responses in dopamine cells is statistically reliable (one-way ANOVA; $F(38, 234) = 2.93$, $P = 4 \times 10^{-7}$). The mean and 95% confidence intervals of $\alpha^+/(\alpha^+ + \alpha^-)$ are displayed, where α^+ and α^- are the slopes estimated above. **d**, Relative scaling of positive and negative responses predicts that cell's reversal point ($P = 8.1 \times 10^{-5}$ by linear regression). Each point represents one dopamine cell. Dashed line is the mean over cells. Light grey traces show reversal points measured in distributional TD simulations of the same task, and show variability over simulation runs. **e**, All 40 dopamine cells plotted in the same fashion as in **b**, except normalized by the slope estimated in the negative domain. Thus, the observed variability in slope in the positive domain corresponds to diversity in relative scaling of positive and negative responses. Cells are coloured by reversal point, to illustrate the relationship between reversal point and asymmetric scaling. In all panels, reward magnitudes are in estimated utility space (see Methods).

more optimistic reward prediction. Conversely, in channels overweighting negative RPEs, a more pessimistic prediction is needed to attain equilibrium (Fig. 4a, Extended Data Fig. 1a). Together, the set of predictions learned across all channels encodes the full shape of the reward distribution.

When distributional RL is considered as a model of the dopamine system, these points translate into two testable predictions. First, dopamine neurons should differ in their relative scaling of positive and negative RPEs. To test this prediction, we analysed activity from VTA dopamine neurons in the variable-magnitude task described above. We first estimated a reversal point for each cell as previously described. Then, for each cell, we separately estimated two slopes: α^+ for responses in the positive domain (that is, above the reversal point), and α^- for the negative domain (Fig. 4b). This revealed reproducible differences across dopamine neurons in the relative magnitude of positive versus negative RPEs (Extended Data Fig. 5). Across all animals, the mean value of the ratio $\alpha^+/(\alpha^+ + \alpha^-)$ was 0.48. However, many cells had a value significantly above or below this mean (Fig. 4c; see Methods for details of statistical test). At the group level, there was significant diversity between cells by one-way ANOVA ($F(38, 234) = 2.93$, $P = 4 \times 10^{-7}$). In the animal with the largest number of recorded cells, 3 out of 15 cells were significantly below the mean and 3 out of 15 were significantly above the mean; ANOVA again rejected the null hypothesis of no diversity between cells ($F(14, 90) = 4.06$, $P = 2 \times 10^{-5}$).

Second, RPE asymmetry should correlate, across dopamine neurons, with reversal point. Dopamine neurons that scale positive RPEs more steeply relative to negative RPEs should be linked with relatively optimistic reward predictions, and so should have reversal points at relatively high reward magnitudes. Dopamine neurons that scale positive RPEs less steeply should have relatively low reversal points. Again using data from the variable-magnitude task, we found a strong correlation between RPE asymmetry and reversal point ($P = 8.1 \times 10^{-5}$ by linear regression; Fig. 4d, e), validating this prediction. Furthermore, this effect survived when only considering data from the single animal with the largest number of recorded cells ($P = 0.002$).

Decoding reward distributions

As we have discussed, the distributional TD model correctly predicts that dopamine neurons should show diverse reversal points and response asymmetries, and that these should correlate. Finally, we consider the most detailed prediction of the model. The specific reversal points observed in any experimental situation, together with the particular response asymmetries in the corresponding neurons, should encode an approximate representation of the anticipated probability distribution over future rewards.

If this is the case, then with sufficient data it should be possible to decode the full value distribution from the responses of dopamine neurons. As a final test of the distributional RL hypothesis, we attempted this type of decoding. The distributional TD model implies that, if dopaminergic responses are approximately linear in the positive and negative domains, then the resultant learned reward predictions will correspond to expectiles of the reward distribution[20] (expectiles are a statistic of distributions, which generalize the mean in the same way that quantiles generalize the median).

We therefore treated the reversal points and response asymmetries measured in the variable-magnitude task as defining a set of expectiles, and we transformed these expectiles into a probability density (see Methods). As shown in Fig. 5a–c, the resulting density captured multiple modes of the ground-truth value distribution. Decoding the RPEs produced by a distributional TD simulation, but not a classical TD simulation, produced the same pattern of results.

Parallel analyses focusing on the variable-probability task (see Methods) yielded similarly good matches to the ground-truth distributions in that task (Fig. 5d, e). In both tasks, successful decoding depended on the specific pattern of variability in the neural data, and not on the presence of variability per se (Extended Data Fig. 8).

It is worth emphasizing that none of the effects we have reported are anticipated by the standard RPE theory of dopamine, which implies that all dopamine neurons should transmit essentially the same RPE signal. Why have the present effects not been observed before? In some cases, relevant data have been hiding in plain sight. For example, a number of studies have reported marked variability in the relative magnitude of positive and negative RPEs across dopamine neurons; however, they have treated this as an incidental finding or a reflection of measurement error, or viewed it as a problem for the RPE theory[17]. One of the earliest studies of reward-probability coding in dopaminergic RPEs remarked on apparent diversity across dopamine neurons, but only in a footnote[18]. A more general issue is that the forms of variability we have reported are masked by traditional analysis techniques, which typically focus on average responses across dopamine neurons (see Supplementary Information and Extended Data Fig. 10).

Distributional RL offers a range of untested predictions. Dopamine neurons should maintain their ordering of relative optimism across task

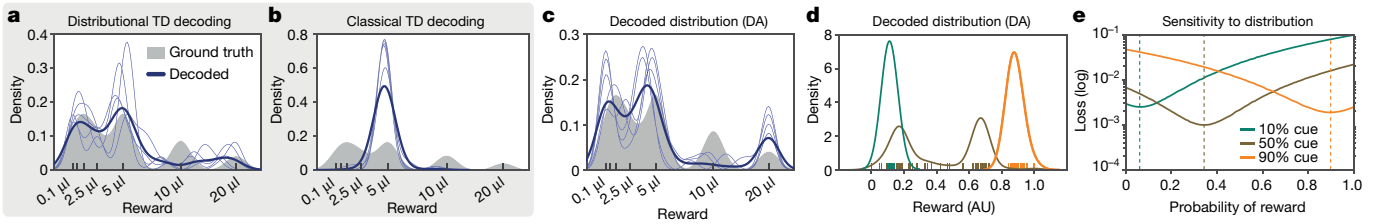


Fig. 5 | Decoding reward distributions from neural responses. **a**, Distributional TD simulation trained on the variable-magnitude task, whose actual (smoothed) distribution of rewards is shown in grey. After training the model, we interpret the learned values as a set of expectiles. We then decode the set of expectiles into a probability density (blue traces). Multiple solutions are shown in light blue, and the average across solutions is shown in dark blue. (See Methods for more details.) **b**, Same as **a**, but with a classical TD simulation. **c**, Same as **a**, but using data from recorded dopamine cells. The expectiles are defined by the reversal points and the relative scaling from the slopes of positive and negative RPEs, as shown in Fig. 4. Unlike the classical TD simulation, the real dopamine cells collectively encode the shape of the reward distribution that animals have been trained to expect. **d**, Same decoding analysis, using data from each of the cue conditions in the variable-probability task, based on cue responses of dopamine neurons (decoding for GABAergic neurons shown in Extended Data Fig. 8i, j). **e**, The neural data for both dopamine and GABAergic neurons were best fit by Bernoulli distributions closely approximating the ground-truth reward probabilities in all three cue conditions.

contexts, even as the specific distribution of rewards changes. If RPE channels with particular levels of optimism are selectively activated with optogenetics, this should sculpt the learned distribution, which should in turn be detectable with behavioural measures of sensitivity to moments of the distribution. We list further predictions in the Supplementary Information.

Distributional RL also gives rise to a number of broader questions. What are the circuit- or cellular-level mechanisms that give rise to a diversity of asymmetry in positive versus negative RPE scaling? It is also worth considering whether other mechanisms, aside from asymmetric scaling of RPEs, might contribute to distributional coding. It is well established, for example, that positive and negative RPEs differentially engage striatal D_1 and D_2 dopamine receptors[21], and that the balance of these receptors varies anatomically[22–24]. This suggests a second potential mechanism for differential learning from positive versus negative RPEs[25]. Moreover, how do different RPE channels anatomically couple with their corresponding reward predictions (see Extended Data Fig. 4i–k)? Finally, what effects might distributional coding have downstream, at the level of action learning and selection? With this question in mind, it is notable that some current theories in behavioural economics centre on risk measures that can be easily read out from the kind of distributional codes that the present work has considered.

Finally, we speculate on the implications of the distributional hypothesis of dopamine for the mechanisms of mental disorders such as addiction and depression. Mood has been linked with predictions of future reward[26], and it has been proposed that both depression and bipolar disorder may involve biased forecasts concerning value-laden outcomes[27]. It has recently been proposed that such biases may arise from asymmetries in RPE coding[28,29]. There are clear potential connections between these ideas and the phenomena we have reported here, presenting opportunities for further research.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1924-6.

1. Schultz, W., Stauffer, W. R. & Lak, A. The phasic dopamine signal maturing: from reward via behavioural activation to formal economic utility. *Curr. Opin. Neurobiol.* **43**, 139–148 (2017).
2. Glimcher, P. W. Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proc. Natl Acad. Sci. USA* **108**, 15647–15654 (2011).
3. Watabe-Uchida, M., Eshel, N. & Uchida, N. Neural circuitry of reward prediction error. *Annu. Rev. Neurosci.* **40**, 373–394 (2017).
4. Morimura, T., Sugiyama, M., Kashima, H., Hachiya, H. & Tanaka, T. Parametric return density estimation for reinforcement learning. In *Proc. 26th Conference on Uncertainty in Artificial Intelligence* (eds Grunwald, P. & Spirtes, P.) http://dl.acm.org/citation.cfm?id=3023549.3023592 (2010).
5. Bellemare, M. G., Dabney, W., & Munos, R. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning* (eds Precup, D. & The, Y. W.) 449–458 (2017).
6. Dabney, W. Rowland, M. Bellemare, M. G. & Munos, R. Distributional reinforcement learning with quantile regression. In *AAAI Conference on Artificial Intelligence* (2018).
7. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: an Introduction* Vol. 1 (MIT Press, 1998).
8. Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
9. Silver, D. et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
10. Hessel, M. et al. Rainbow: combining improvements in deep reinforcement learning. In *32nd AAAI Conference on Artificial Intelligence* (2018).
11. Botvinick, M. M., Niv, Y. & Barto, A. G. Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. *Cognition* **113**, 262–280 (2009).
12. Wang, J. X. et al. Prefrontal cortex as a meta-reinforcement learning system. *Nat. Neurosci.* **21**, 860–868 (2018).
13. Song, H. F., Yang, G. R. & Wang, X. J. Reward-based training of recurrent neural networks for cognitive and value-based tasks. *eLife* **6**, e21492 (2017).
14. Barth-Maron, G. et al. Distributed distributional deterministic policy gradients. In *International Conference on Learning Representations* https://openreview.net/forum?id=SyZipzbCb (2018).
15. Dabney, W., Ostrovski, G., Silver, D. & Munos, R. Implicit quantile networks for distributional reinforcement learning. In *International Conference on Machine Learning* (2018).
16. Pouget, A., Beck, J. M., Ma, W. J. & Latham, P. E. Probabilistic brains: knowns and unknowns. *Nat. Neurosci.* **16**, 1170–1178 (2013).
17. Lammel, S., Lim, B. K. & Malenka, R. C. Reward and aversion in a heterogeneous midbrain dopamine system. *Neuropharmacology* **76**, 351–359 (2014).
18. Fiorillo, C. D., Tobler, P. N. & Schultz, W. Discrete coding of reward probability and uncertainty by dopamine neurons. *Science* **299**, 1898–1902 (2003).
19. Eshel, N. et al. Arithmetic and local circuitry underlying dopamine prediction errors. *Nature* **525**, 243–246 (2015).
20. Rowland, M., et al. Statistics and samples in distributional reinforcement learning. In *International Conference on Machine Learning* (2019).
21. Frank, M. J., Seeberger, L. C. & O'Reilly, R. C. By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science* **306**, 1940–1943 (2004).
22. Hirvonen, J. et al. Striatal dopamine D1 and D2 receptor balance in twins at increased genetic risk for schizophrenia. *Psychiatry Res. Neuroimaging* **146**, 13–20 (2006).
23. Piggott, M. A. et al. Dopaminergic activities in the human striatum: rostrocaudal gradients of uptake sites and of D1 and D2 but not of D3 receptor binding or dopamine. *Neuroscience* **90**, 433–445 (1999).
24. Rosa-Neto, P., Doudet, D. J. & Cumming, P. Gradients of dopamine D1- and D2/3-binding sites in the basal ganglia of pig and monkey measured by PET. *Neuroimage* **22**, 1076–1083 (2004).
25. Mikhael, J. G. & Bogacz, R. Learning reward uncertainty in the basal ganglia. *PLOS Comput. Biol.* **12**, e1005062 (2016).
26. Robb, B. et al. A computational and neural model of momentary subjective well-being. *Proc. Natl Acad. Sci. USA* **111**, 12252–12257 (2014).
27. Huys, Q. J., Daw, N. D. & Dayan, P. Depression: a decision-theoretic analysis. *Annu. Rev. Neurosci.* **38**, 1–23 (2015).
28. Bennett, D. & Niv, Y. Opening Burton's clock: psychiatric insights from computational cognitive models. Preprint at https://doi.org/10.31234/osf.io/y2vzu (2018).
29. Tian, J. & Uchida, N. Habenula lesions reveal that multiple mechanisms underlie dopamine prediction errors. *Neuron* **87**, 1304–1316 (2015).
30. Eshel, N., Tian, J., Bukwich, M. & Uchida, N. Dopamine neurons share common response function for reward prediction error. *Nat. Neurosci.* **19**, 479–486 (2016).

Methods

Distributional RL model

The model for distributional RL we use throughout the work is based on the principle of asymmetric regression and extends recent results in AI[5,6,15]. We present a more detailed and accessible introduction to distributional RL in the Supplementary Information. Here we outline the method in brief.

Let $f: \mathbb{R} \to \mathbb{R}$ be a response function. In each observed state x, let there be a set of value predictions $V_i(x)$ which are updated with learning rates $\alpha_i^+, \alpha_i^- \in \mathrm{R}^+$. Then given a state x, next-state x', resulting reward signal r and time discount $\gamma \in [0, 1)$, the distributional TD model computes distributional TD errors

$$\delta_i = r + \gamma V_j(x') - V_i(x) \quad (1)$$

where $V_j(x')$ is a sample from the distribution $V(x')$. The model then updates the baselines with

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

$$V_i(x) \leftarrow V_i(x) + \alpha_i^- f(\delta_i) \text{ for } \delta_i \le 0 \quad (3)$$

When performed with a tabular representation, asymmetry uniformly distributed, and $f(\delta) = \mathrm{sgn}(\delta)$, this method converges to the τ_i quantile, $\tau_i = \frac{\alpha_i^+}{\alpha_i^+ + \alpha_i^-}$, of the distribution over discounted returns at x (ref. [6]). Similarly, asymmetric regression with response function $f(\delta) = \delta$ corresponds to expectile regression[31]. Like quantiles, expectiles fully characterize the distribution and have been shown to be particularly useful for measures of risk[32,33].

Finally, we note that throughout the paper, we use the terms optimistic and pessimistic to refer to return predictions that are above or below the mean (expected) return. Importantly, these predictions are optimistic in the sense of corresponding to particularly good outcomes from the set of possible outcomes. They are not optimistic in the sense of corresponding to outcomes that are impossibly good.

Artificial agent results

Atari results are on the Atari-57 benchmark using the publicly available Arcade Learning Environment[34]. This is a set of 57 Atari 2600 games and human-performance baselines. Refer to previous work for details on deep Q-networks (DQN) and computation of human-normalized scores[8]. The distributional TD agent uses our proposed model and a DQN with multiple ($n = 200$) value predictors, each with a different asymmetry, spaced uniformly in [0, 1]. The training objective of DQN, the Huber loss, is replaced with the asymmetric quantile-Huber loss, which corresponds to the κ-saturating response function $f(\delta) = \max(\min(\delta, \kappa), -\kappa)$, with $\kappa = 1$.

Finally, at each update we train all channels based on the immediate reward and the predicted future returns from all next-state value predictors. Further details can be found in ref. [6]. The physics-based motor-control task requires control of a 28 degrees-of-freedom humanoid to complete a 3D obstacle course in minimal time[35]. Full details for the D3PG and distributional D3PG agents are as described[14]. Distributions over return shown in Extended Data Fig. 2d, f are based on the network-predicted distribution in each of the given frames.

Tabular simulations

Tabular simulations of the classical TD and distributional TD models used a population of learning rates selected uniformly at random, $\alpha_i^+ \sim \mathrm{U}(0, 1)$ for each cell i. In all cases the only algorithmic difference between the classical and distributional TD models was that the distributional model used a separately varying learning rate for negative prediction errors, $\alpha_i^- \sim \mathrm{U}(0, 1)$ for each cell i. Both methods used a linear response function. Qualitatively similar results were also obtained with other response functions (for example, Hill function[30] or κ–saturating), despite these leading to semantically different estimators of the distribution. The population sizes were chosen for clarity of presentation and to provide similar variability as observed in the neuronal data. Each cell was paired with a different state-dependent value estimate $V_i(x)$. Note that while these simulations focused on immediate rewards, the same algorithm also learns distributions over multi-step returns.

In the variable-probability task, each cue corresponded to a different value estimate and reward probability (90%, 50% or 10%). When rewarded, the agent received numerical reward of 1.0, and when omitted, it received 0.0. Both agents were trained for 100 trials of 5,000 updates, and both simulated $n = 31$ cells (separate value estimates). The learning rates were all selected uniformly at random between [0.001, 0.2]. Cue response was taken to be the temporal difference from a constant zero baseline to the value estimate.

In the variable-magnitude task, all rewards were taken to be the water magnitude measured in microlitres (qualitatively same results obtained with utilities instead of magnitudes). For Fig. 2 we ran 10 trials of 25,000 updates each for 150 estimators with random learning rates in [0.001, 0.02]. These smaller learning rates and larger number of updates were intended to ensure the values converged fully with low error. We then report temporal difference errors for ten cells taken uniformly to span the range of value estimates for each agent. Reported errors (simulating change in firing rate) are the utility of a reward minus the value estimate and scaled by the learning rate. As with the neuronal data, these are reported averaged over trials and normalized by variance over reward magnitudes. Distributional TD RPEs are computed using asymmetric learning rates, with a small constant (floor) added to the learning rates.

Distribution decoding

For both real neural data and TD simulations, we performed distribution decoding. The distributional and classical TD simulations used for decoding in the variable-magnitude task each used 40 value predictors, to match the 40 recorded cells in the neural data (neural analyses were pooled across the six animals). In the distributional TD simulation, each value predictor used a different asymmetric scaling factor $\tau_i = \frac{\alpha_i^+}{\alpha_i^+ + \alpha_i^-}$, and therefore learned a different value prediction V_i.

The decoding analyses began with a set of reversal points, V_i, and asymmetric scaling factors τ_i. For the neural data, these were obtained as described elsewhere. For the simulations, they were read directly from the simulation. These numbers were interpreted as a set of expectiles, with the τ_i-th expectile having value V_i. We decoded these into probability densities by solving an optimization problem to find the density most compatible with the set of expectiles[20]. For optimization, the density was parameterized as a set of samples. For display in Fig. 5, the samples are smoothed with kernel density estimation.

Animals and behavioural tasks

The rodent data we re-analysed here were first reported in ref. [19]. Methods details can be found in that paper and in ref. [30]. We give a brief description of the methods below.

Five mice were trained on a 'variable-probability' task, and six different mice on a 'variable-magnitude' task. In the variable-probability task, in each trial the animal first experienced one of four odour cues for 1 s, followed by a 1-s pause, followed by a reward (3.75 µl water), an aversive airpuff or nothing. Odour 1 signalled a 90% chance of reward, odour 2 signalled a 50% chance of reward, odour 3 signalled a 10% chance of reward and odour 4 signalled a 90% chance of airpuff. Odour meanings were randomized across animals. Inter-trial intervals were exponentially distributed.

An infrared beam was positioned in front of the water delivery spout, and each beam break was recorded as one lick event. We report the average lick rate over the entire interval between the cue and the outcome (that is, 0–2,000 ms after cue onset).

In the variable-magnitude task, in 10% of trials an odour cue was delivered that indicated that no reward would be delivered on that trial. In the remaining 90% of trials, one of the following reward magnitudes was delivered, at random: 0.1, 0.3, 1.2, 2.5, 5, 10 or 20 µl. In half of these trials, this reward was preceded by 1,500 ms by an odour cue (which indicated that a reward was forthcoming but did not disclose its magnitude). In the other half, it was unsignalled.

In order to identify dopamine neurons while recording, neurons in the VTA were tagged with channelrhodopsin-2 (ChR2) by injecting adeno-associated virus (AAV) that expresses ChR2 in a Cre-dependent manner into the VTA of transgenic mice that express Cre recombinase under the promoter of the dopamine transporter (DAT) gene *Slc6a3* (B6.SJL-Slc6a3tm1.1(cre)Bkmn/J, The Jackson Laboratory)[36]. Mice were implanted with a head plate and custom-built microdrive containing 6–8 tetrodes (Sandvik) and optical fibre, as described[37].

All experiments were performed in accordance with the US National Institutes of Health Guide for the Care and Use of Laboratory Animals and approved by the Harvard Institutional Animal Care and Use Committee.

Neuronal data and analysis

Extracellular recordings were made from VTA using a data acquisition system (DigiLynx, Neuralynx). VTA recording sites were verified histologically. The identity of dopaminergic cells was confirmed by recording the electrophysiological responses of cells to a brief blue light pulse train, which stimulates only DAT-expressing cells. Spikes were sorted using SpikeSort3D (Neuralynx) or MClust-3.5 (A.D. Redish). Putative GABAergic neurons in the VTA were identified by clustering of firing patterns as described previously[30,37]. All confidence intervals are s.e.m. unless otherwise noted.

Data analyses were performed using NumPy 1.15 and MATLAB R2018a (Mathworks). Spike times were collected in 1-ms bins to create peristimulus time histograms. These histograms were then smoothed by convolving with the function $(1 - e^{-t}) \cdot e^{-t/T}$, where T was a time constant, set to 20 ms as in ref. [30]. For single-cell traces, we set T to 200 ms for display purposes.

After smoothing, the data were baseline-corrected by subtracting from each trial and each neuron independently the mean over that trial's activity from −1,000 to 0 ms relative to stimulus onset (or relative to reward onset in the unexpected reward condition).

Variable-probability task. $n = 31$ cells were recorded from five animals, with the following number of cells per animal: 1, 4, 16, 1 and 9. Responses to cue for dopamine neurons were defined as the average activity from 0 to 400 ms after cue onset. This interval was chosen to match ref. [30]. Responses to cue for putative GABAergic neurons were defined as the average activity from 0 to 1,500 ms after cue onset. This longer interval was chosen because these neurons had much slower responses, often ramping up slowly over the first 500 or 1,000 ms after cue onset[37] (Fig. 3d).

We were interested in whether there was between-cell diversity in responses to the 50% cue. We first normalized the responses to the 50% cue on a per-cell basis as follows: $c_{50}^{\mathrm{norm}} = (c_{50} - \mathrm{mean}(c_{10}))/(\mathrm{mean}(c_{90}) - \mathrm{mean}(c_{10}))$, where mean indicates the mean over trials within a cell. In order to be agnostic about the risk preferences of the animal, we then performed a two-tailed t-test of the cell's normalized responses to the 50% cue against the average of all cells' normalized responses to the 50% cue. This is the test for optimistic or pessimistic probability coding that we report in the main text. Note that these t-statistics would be t-distributed if the differences between cells were due to chance. We also report ANOVA results where we evaluate the null hypothesis that all cells' normalized 50% responses have the same mean.

The same pattern of results held when instead comparing responses to the 50% cue against the midway point between responses to the 10% cue and responses to the 90% cue.

The per-cell cue responses shown in Extended Data Fig. 7 were normalized to zero mean and unit variance, to allow direct comparison of cells with different response variability. Each cell appears in one of three panels based on the outcome of two single-tailed Mann–Whitney tests evaluating the rank order for $c_{10} < c_{50}$ and $c_{50} < c_{90}$ (see Supplementary Information section 3.3 for further details). The left, centre and right panels correspond to outcomes ($P \geq 0.05$, $P < 0.05$), ($P < 0.05$, $P < 0.05$ or $P \geq 0.05$, $P \geq 0.05$) and ($P < 0.05$, $P \geq 0.05$), respectively.

Variable-magnitude task. $n = 40$ cells were recorded from five animals, with the following number of cells per animal: 3, 6, 9, 16 and 6. Responses to reward were defined as the average activity from 200 to 600 ms after reward onset. This time interval was selected to match ref. [30] as closely as possible, while excluding the initial response to the feeder click[30,38,39], which was not selective to reward magnitude and was positive for all reward magnitudes. This enabled us to find the reward magnitudes for which the dopamine response was either boosted or suppressed relative to baseline.

The reversal point (that is, the reward magnitude that would elicit neither a positive nor a negative deflection in firing relative to baseline) for each cell was defined as the magnitude M_R that maximized the number of positive responses to rewards greater than M_R plus the number of negative responses to rewards less than M_R. To obtain statistics for reliability of cell-to-cell differences in reversal point, we partitioned the data into random halves and estimated the reversal point for each cell separately in each half. We repeated this procedure 1,000 times with different random partitions, and we report the mean R value and geometric mean p value across these 1,000 folds.

After measuring reversal points, we fit linear functions separately to the positive and negative domains of each cell. To obtain confidence intervals, we divided the data into seven random partitions (seven being the smallest number of trials in any condition for any cell), subject to the constraint that every condition for every cell contain at least one trial in each partition. In each partition, we repeated the procedure for estimating reversal points and finding slopes in the positive and negative domains. Our confidence interval on $\tau = \alpha^+/(\alpha^+ + \alpha^-)$ was then the s.e.m. of the values calculated across the seven partitions. ANOVAs are also reported testing the null hypothesis that means (across partitions) were not different between cells.

Fitting linear functions to dopamine responses was more logical in utility space than in reward volume space. We relied on ref. [38] to approximate the underlying utility function from the dopamine responses to rewards of varying magnitudes. We used these empirical utilities instead of raw reward magnitudes for the analyses shown in Fig. 4. However, none of the reported results were sensitive to this choice of utility function. We also ran the analyses using other utility functions, and these results are reported in Extended Data Fig. 5. One cell was excluded from analyses in Fig. 5: because it had no positive responses to any reward magnitude, a slope could not be fit in the positive domain.

When measuring the correlation (across cells) between reversal point and τ, we first randomly split the data into two disjoint halves of trials. In one half, we first calculated reversal points RP_1 and used these reversal points to calculate α^+ and α^-. In the other half, we calculated reversal points RP_2. The correlation we report is between RP_2 and $\tau = \alpha^+/(\alpha^+ + \alpha^-)$. We did this to avoid confounds associated with using the same data to estimate both slopes and intercepts.

Reporting summary

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

Data availability

The neuronal data analysed in this work are available at https://doi.org/10.17605/OSF.IO/UX5RG.

Code availability

The analysis code from our value-distribution decoding and code used to generate model predictions for distributional TD are available at https://doi.org/10.17605/OSF.IO/UX5RG.

31. Newey, W. K. & Powell, J. L. Asymmetric least squares estimation and testing. *Econometrica* **55**, 819–847 (1987).
32. Chris Jones, M. Expectiles and m-quantiles are quantiles. *Stat. Probab. Lett.* **20**, 149–153 (1994).
33. Ziegel, J. F. Coherence and elicitability. *Math. Finance* **26**, 901–918 (2016).
34. Bellemare, M. G., Naddaf, Y., Veness, J. & Bowling, M. The arcade learning environment: an evaluation platform for general agents. *J. Artif. Intell. Res.* **47**, 253–279 (2013).
35. Heess, N. et al. Emergence of locomotion behaviours in rich environments. Preprint at https://arxiv.org/abs/1707.02286 (2017).
36. Bäckman, C. M., et al. Characterization of a mouse strain expressing cre recombinase from the 3′ untranslated region of the dopamine transporter locus. *Genesis* **44**, 383–390 (2006).
37. Cohen, J. Y. et al. Neuron-type-specific signals for reward and punishment in the ventral tegmental area. *Nature* **482**, 85–88 (2012).
38. Stauffer, W. R., Lak, A. & Schultz, W. Dopamine reward prediction error responses reflect marginal utility. *Curr. Biol.* **24**, 2491–2500 (2014).
39. Fiorillo, C. D., Song, M. R. & Yun, S. R. Multiphasic temporal dynamics in responses of midbrain dopamine neurons to appetitive and aversive stimuli. *J. Neurosci.* **33**, 4710–4725 (2013).
40. Schaul, T., Quan, J., Antonoglou, I. & Silver, D. Prioritized experience replay. In *International Conference on Learning Representations* (2016).
41. Van Hasselt, H., Guez, A. & Silver, D. Deep reinforcement learning with double *q*-learning. In *AAAI Conference on Artificial Intelligence* (2016).
42. Krizhevsky, A. & Hinton, G. *Learning Multiple Layers of Features from Tiny Images* (Univ. of Toronto, 2009).

Acknowledgements We thank K. Miller, P. Dayan, T. Stepleton, J. Paton, M. Frank, C. Clopath, T. Behrens and the members of the Uchida laboratory for comments on the manuscript; and N. Eshel, J. Tian, M. Bukwich and M. Watabe-Uchida for providing data.

Author contributions W.D. conceived the project. W.D., Z.K.-N. and M.B. contributed ideas for experiments and analysis. W.D. and Z.K.-N. performed simulation experiments and analysis. N.U. and C.K.S. provided neuronal data for analysis. W.D., Z.K.-N. and M.B. managed the project. M.B., N.U., R.M. and D.H. advised on the project. M.B., W.D. and Z.K.-N. wrote the paper. W.D., Z.K.-N., M.B., N.U., C.K.S., D.H. and R.M. provided revisions to the paper.

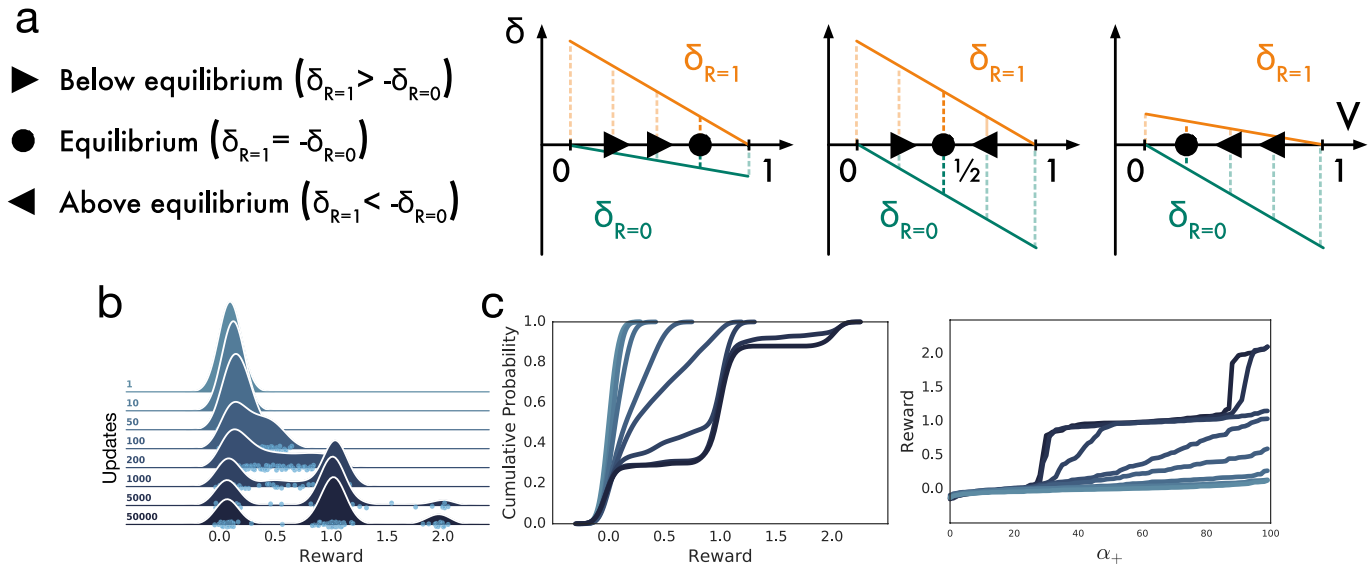
Competing interests The authors declare no competing interests.

Additional information

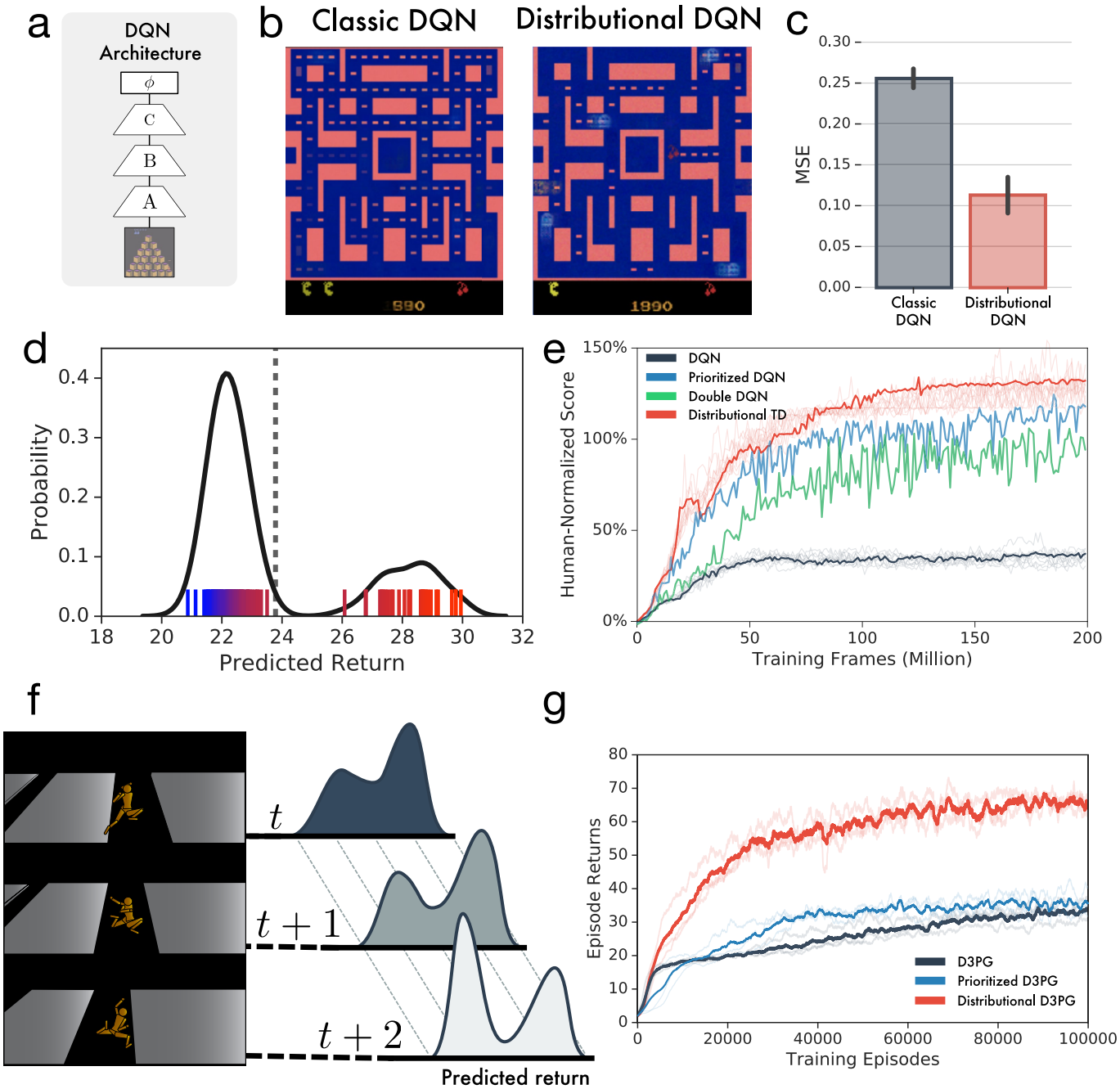
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1924-6.

Correspondence and requests for materials should be addressed to W.D.

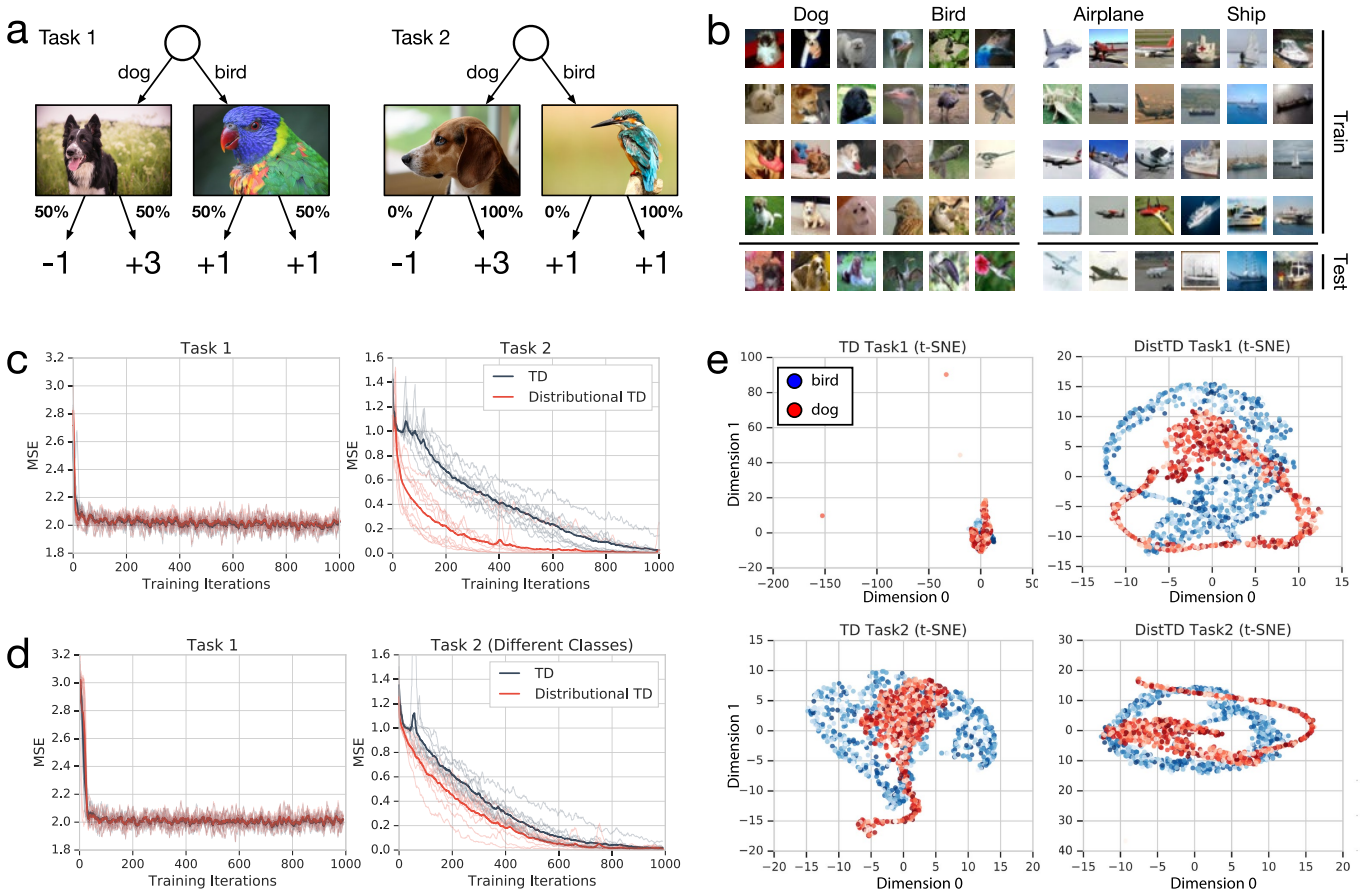
Peer review information *Nature* thanks Rui Costa, Michael Littman and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



Extended Data Fig. 1 | Mechanism of distributional TD. a, The degree of asymmetry in positive to negative scale determines the equilibrium where positive and negative errors balance. Equal scaling equilibrates at the mean, whereas a larger positive (negative) scaling produces an equilibrium above (below) the mean. **b**, Distributional prediction emerges through experience. Quantile (sign function) version is displayed here for clarity. Model is trained on arbitrary task with trimodal reward distribution. **c**, Same as **b**, viewed in terms of cumulative distribution (left) or learned value for each predictor (quantile function) (right).

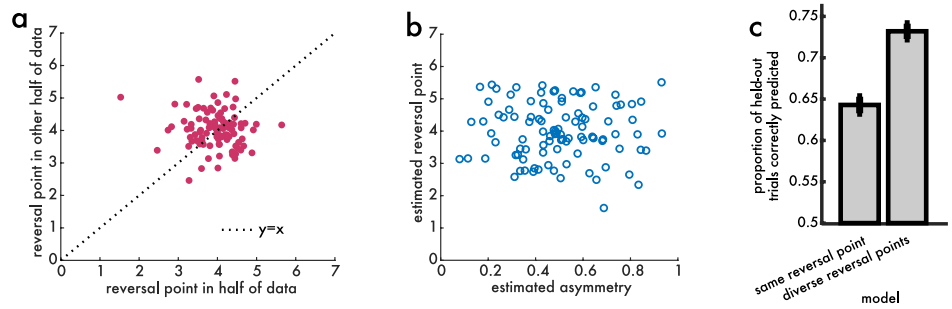


Extended Data Fig. 2 | Learning the distribution of returns improves performance of deep RL agents across multiple domains. a, DQN and distributional TD share identical nonlinear network structures. **b**, **c**, After training classical or distributional DQN on MsPacman, we freeze the agent and then train a separate linear decoder to reconstruct frames from the agent's final layer representation. For each agent, reconstructions are shown. The distributional model's representation allows substantially better reconstruction. **d**, At a single frame of MsPacman (not shown), the agent's value predictions together represent a probability distribution over future rewards. Reward predictions of individual RPE channels shown as tick marks ranging from pessimistic (blue) to optimistic (red), and kernel density estimate shown in black. **e**, Atari-57 experiments with single runs of prioritized experience replay[40] and double DQN[41] agents for reference. Benefits of distributional learning exceed other popular innovations. **f**, **g**, The performance pay-off of distributional RL can be seen across a wide diversity of tasks. Here we give another example, a humanoid motor-control task in the MuJoCo physics simulator. Prioritized experience replay agent is shown for reference[14]. Traces show individual runs; averages are in bold.

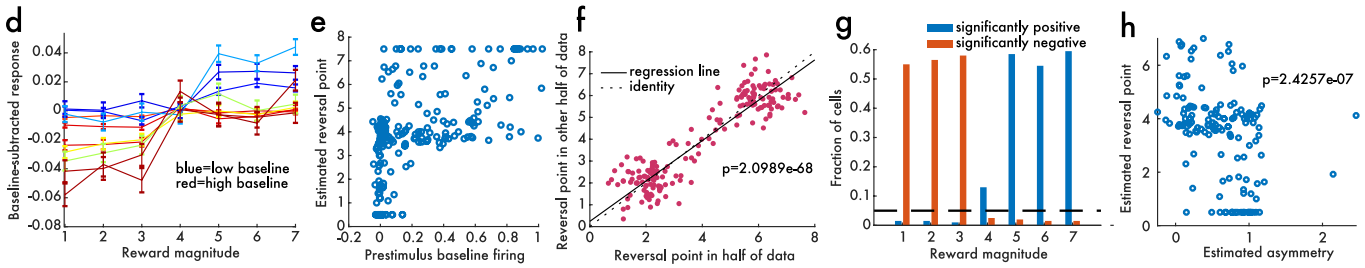


Extended Data Fig. 3 | Simulation experiment to examine the role of representation learning in distributional RL. a, Illustration of tasks 1 and 2. **b**, Example images for each class used in our experiment[42] **c**, Experimental results, where each of ten random seeds yields an individual run shown with traces; average over seeds is shown in bold. **d**, Same as **c**, but for control experiment. **e**, Bird–dog *t*-SNE visualization of final hidden layer of network, given different input images (blue, bird; red, dog). Left, classical TD; right, distributional TD; top row, representation after training on task 1; bottom row, representation after training on task 2.

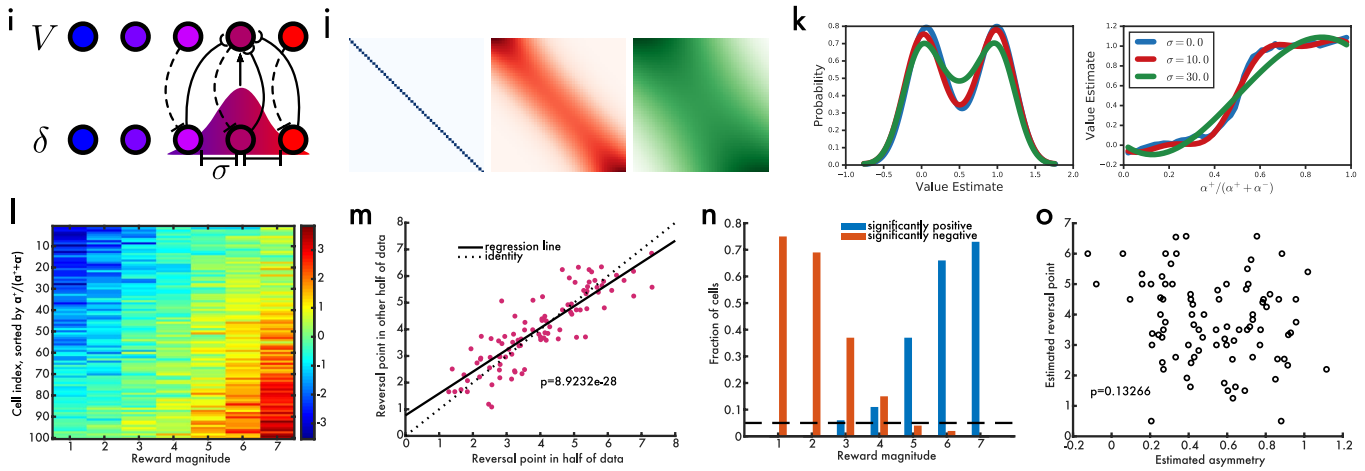
Model 1: RPE+noise



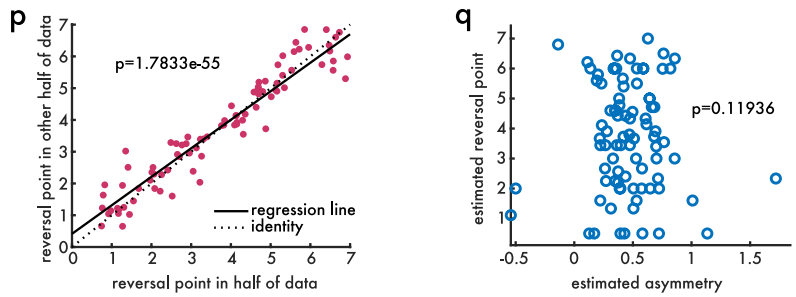
Model 3: σ(RPE+bias)+noise



Model 5: "One-sided" distributional RL

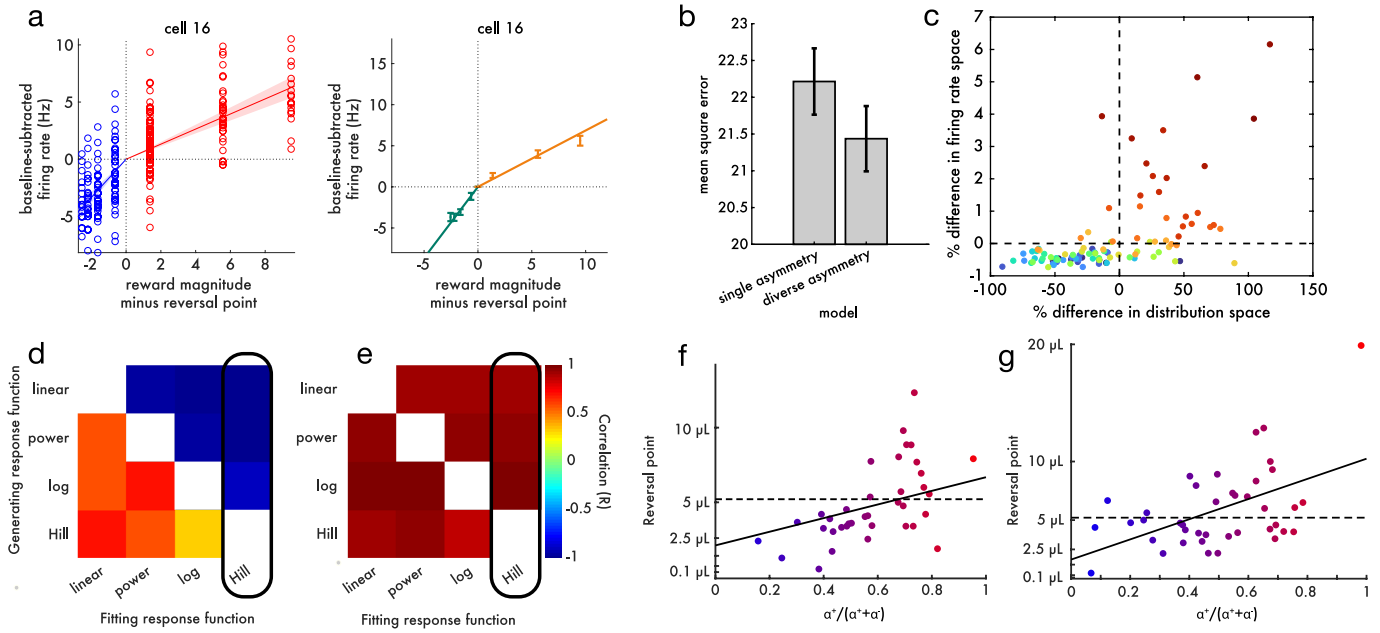


Model 6: Synaptic asymmetries

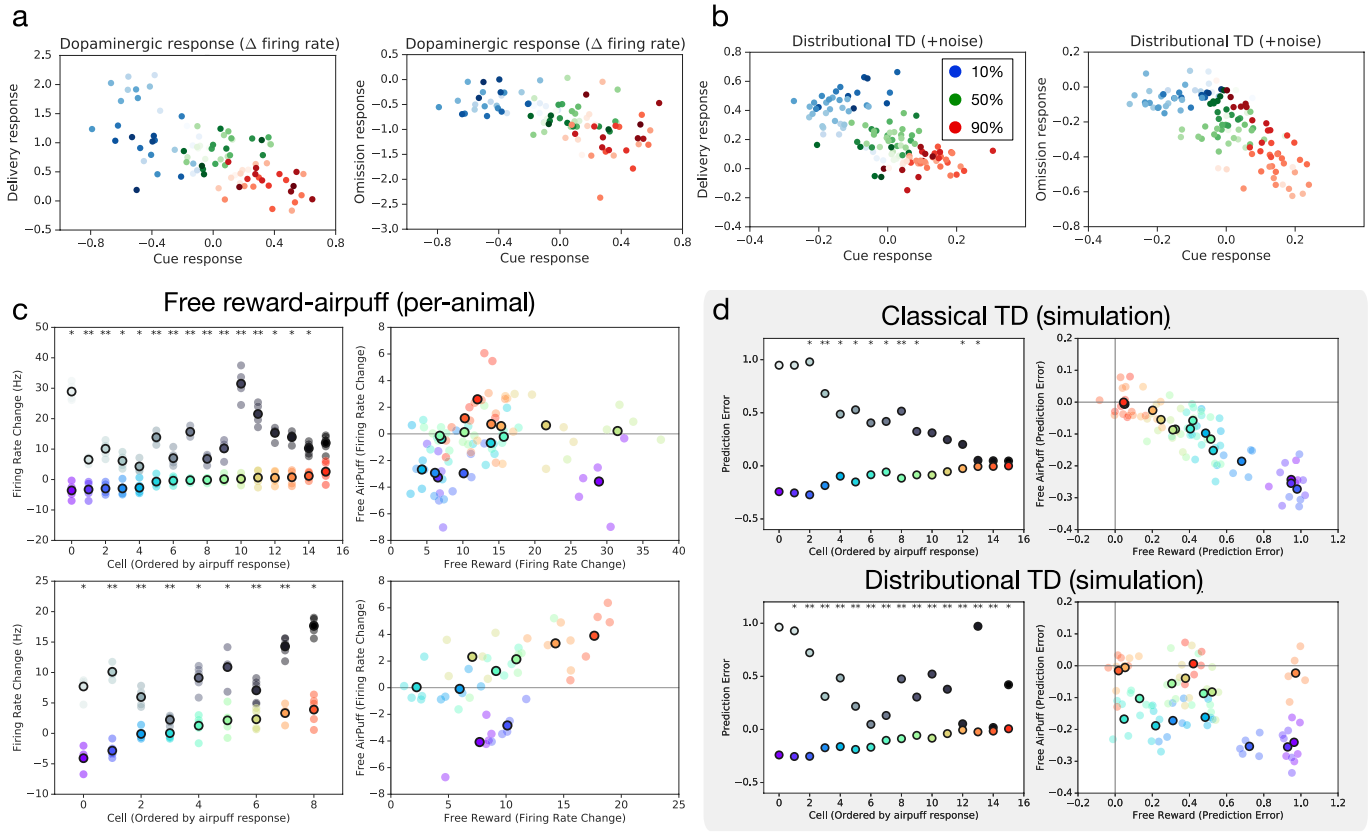


Extended Data Fig. 4 | See next page for caption.

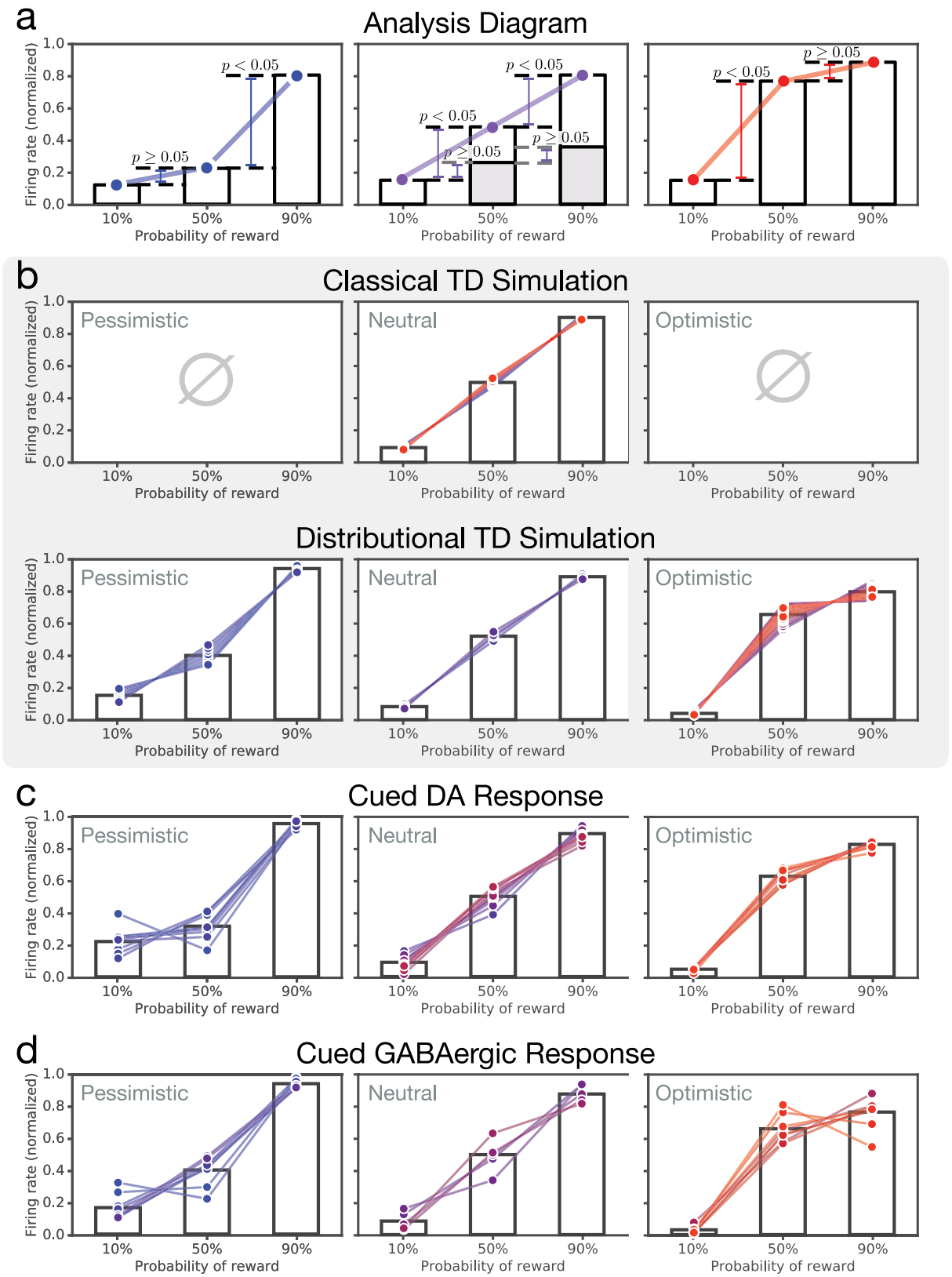
Extended Data Fig. 4 | Null models. a, Classical TD plus noise does not give rise to the pattern of results observed in real dopamine data in the variable-magnitude task. When reversal points were estimated in two independent partitions there was no correlation between the two ($P = 0.32$ by linear regression). **b**, We then estimated asymmetric scaling of responses and found no correlation between this and reversal point ($P = 0.78$ by linear regression). **c**, Model comparison between 'same', a single reversal point, and 'diverse', separate reversal points. In both, the model is used to predict whether a held-out trial has a positive or negative response. **d**, Simulated baseline-subtracted RPEs, colour-coded according to the ground-truth value of bias added to that cell's RPEs. **e**, Across all simulated cells, there was a strong positive relationship between pre-stimulus baseline firing and the estimated reversal point. **f**, Two independent measurements of the reversal point were strongly correlated. **g**, The proportion of simulated cells that have significantly positive (blue) or negative (red) responses showed no magnitudes with both positive and negative responses. **h**, In the simulation, there was a significant negative relationship between the estimated asymmetry of each cell and its estimated reversal point (opposite that observed in neural data). **i**, Diagram illustrating a Gaussian-weighted topological mapping between RPEs and value predictors. **j**, Varying the standard deviation of this Gaussian modulates the degree of coupling. **k**, In a task with equal chance of a reward 1.0 or 0.0, distributional TD with different levels of coupling shows robustness to the degree of coupling. **l**, When there is no coupling, a distributional code is not learned, but asymmetric scaling can cause spurious detection of diverse reversal points. **m**, Even though every cell has the same reward prediction they appear to have different reversal points. **n**, With this model, some cells may have significantly positive responses, and others significantly negative responses, in response to the same reward. **o**, But this model is unable to explain a positive correlation between asymmetric scaling and reversal points. **p**, Simulation of 'synaptic' distributional RL, in which learning rates but not firing rates are asymmetrically scaled. This model predicts diversity in reversal points between dopamine neurons. **q**, The model predicts no correlation between asymmetric scaling of firing rates and reversal point.



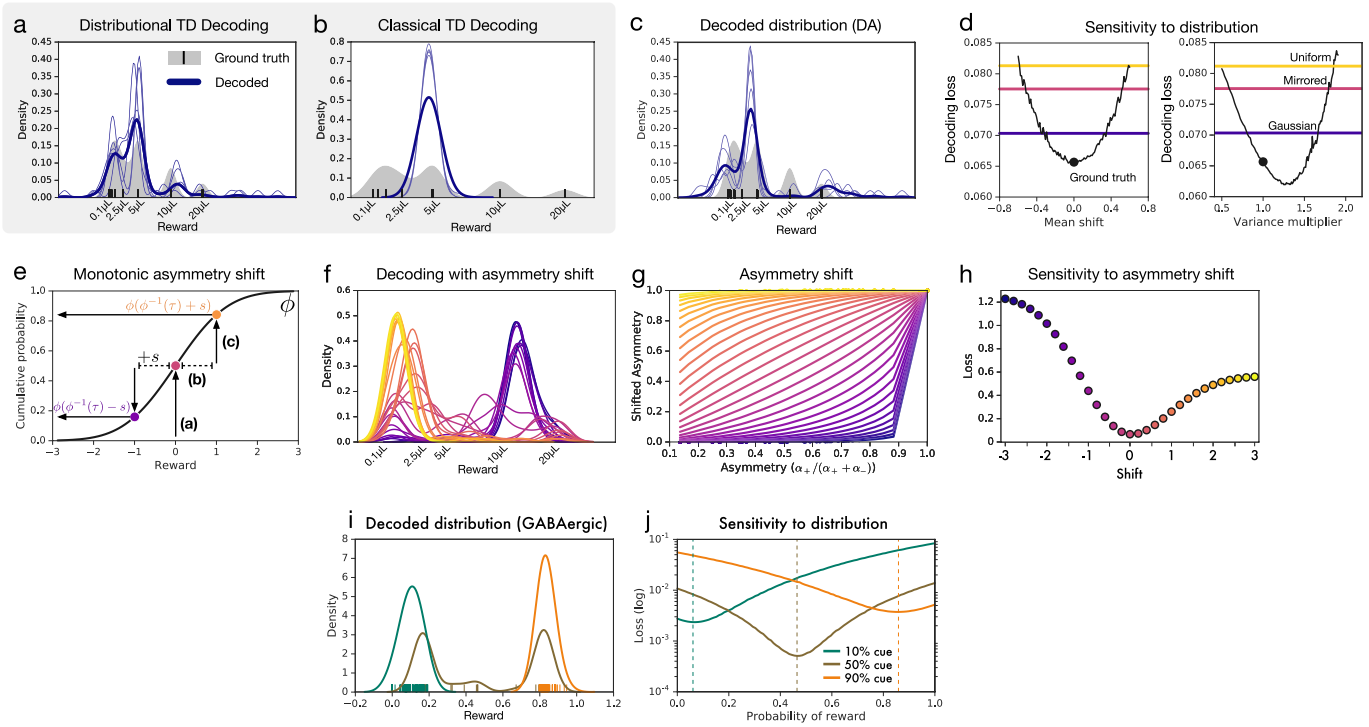
Extended Data Fig. 5 | Asymmetry and reversal. a, Left, all data points (trials) from an example cell. The solid lines are linear fits to the positive and negative domains, and the shaded areas show 95% confidence intervals calculated with Bayesian regression. Right, the same cell plotted in the format of Fig. 4b. **b**, Cross-validated model comparison on the dopamine data favours allowing each cell to have its own asymmetric scaling ($P = 1.4 \times 10^{-11}$ by paired t-test). The standard error of the mean appears large relative to the P value because the P value is computed using a paired test. **c**, Although the difference between single-asymmetry and diverse-asymmetry models was small in firing-rate space, such small differences correspond to large differences in decoded distribution space (more details in Supplementary Information). Each point is a TD simulation; colour indicates the degree of diversity in asymmetric scaling within that simulation. **d**, We were interested in whether an apparent correlation between reversal point and asymmetry could arise as an artefact, owing to a mismatch between the shape of the actual dopamine response function and the function used to fit it. Here we simulate the variable-magnitude task using a TD model without a true correlation between asymmetric scaling and reversal point. We then apply the same analysis pipeline as in the main paper, to measure the correlation (colour axis) between asymmetric scaling and reversal point. We repeat this procedure 20 times with different dopamine response functions in the simulation, and different functions used to fit the positive and negative domains of the simulated data. The functions are sorted in increasing order of concavity. An artefact can emerge if the response function used to fit the data is less concave than the response function used to generate the data. For example, when generating data with a Hill function but fitting with a linear function, a positive correlation can be spuriously measured. **e**, When simulating data from the distributional TD model, where a true correlation exists between asymmetric scaling and reversal point, it is always possible to detect this positive correlation, even if the fitting response function is more concave than the generating response function. The black rectangle highlights the function used to fit real neural data in **c**. **f**, Here we analyse the real dopamine cell data identically to Fig. 4d, but using Hill functions instead of linear functions to fit the positive and negative domains. Because the correlation between asymmetric scaling and reversal point still appears under these adversarial conditions, we can be confident it is not driven by this artefact. **g**, Same as Fig. 4d, but using linear response function and linear utility function (instead of empirical utility).



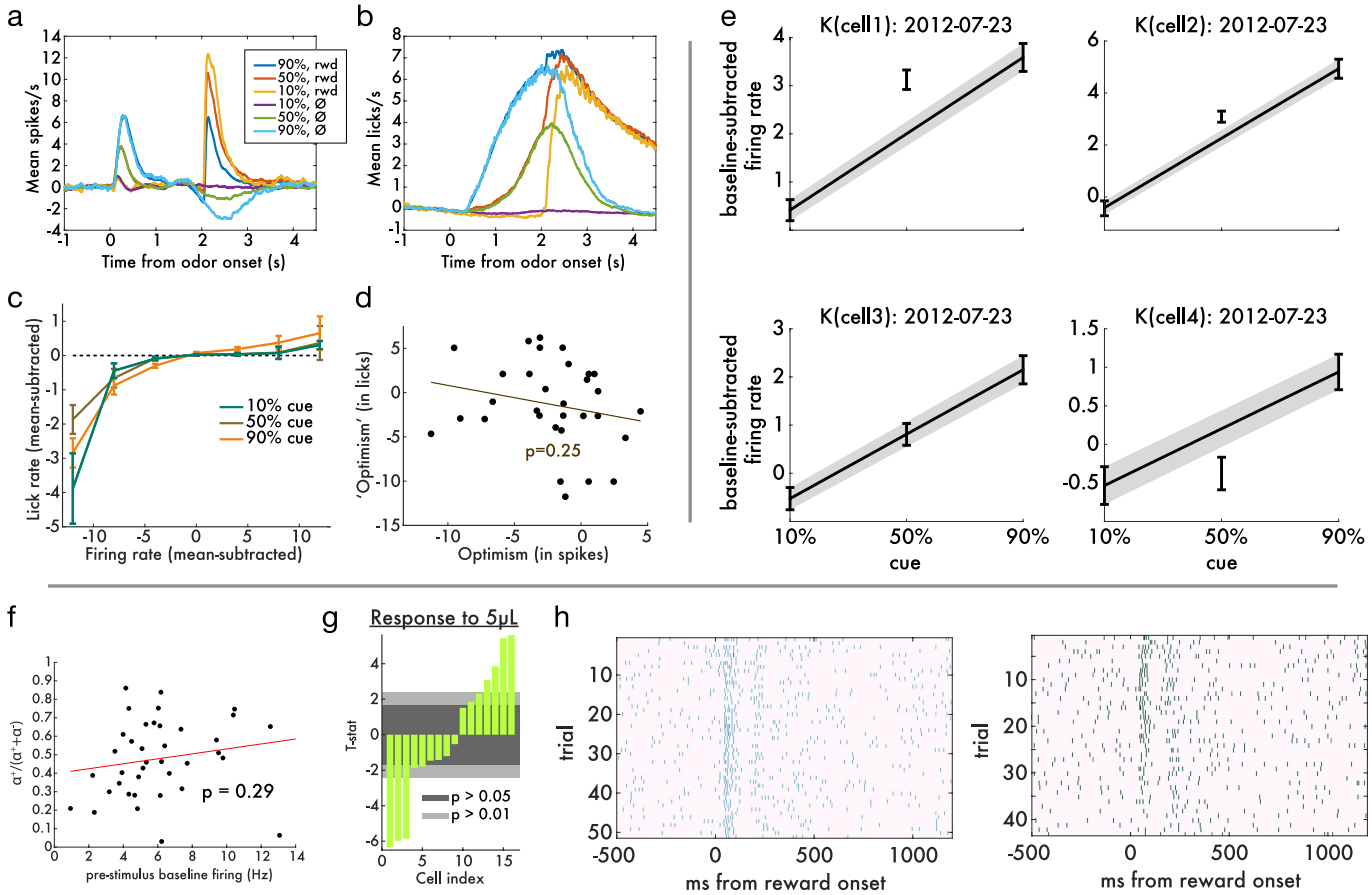
Extended Data Fig. 6 | Cue responses versus outcome responses, and more evidence for diversity. a, In the variable-probability task: firing at cue, versus firing at reward (left) or omission (right). Colour brightness denotes asymmetry. **b**, Same as **a**, but showing RPEs from distributional TD simulation. **c**, Data from ref. [30] also included unpredicted rewards and unpredicted airpuffs. Top two panels show responses for all the cells recorded in one animal and bottom two panels show responses for all the cells of another animal. Left, the *x* axis is the baseline-subtracted response to free reward and the *y* axis is the baseline-subtracted response to airpuff. Dots with black outlines are per-cell means, and un-outlined dots are means of disjoint subsets of trials indicating consistency of asymmetry. Right, the same data plotted in a different way, with cells sorted along the *x* axis by response to airpuff. Response to reward is shown in greyscale dots. Asterisks indicate significant difference in firing rates from one or both neighbouring cells. **d**, Simulations for distributional but not classical TD produce diversity in relative response.



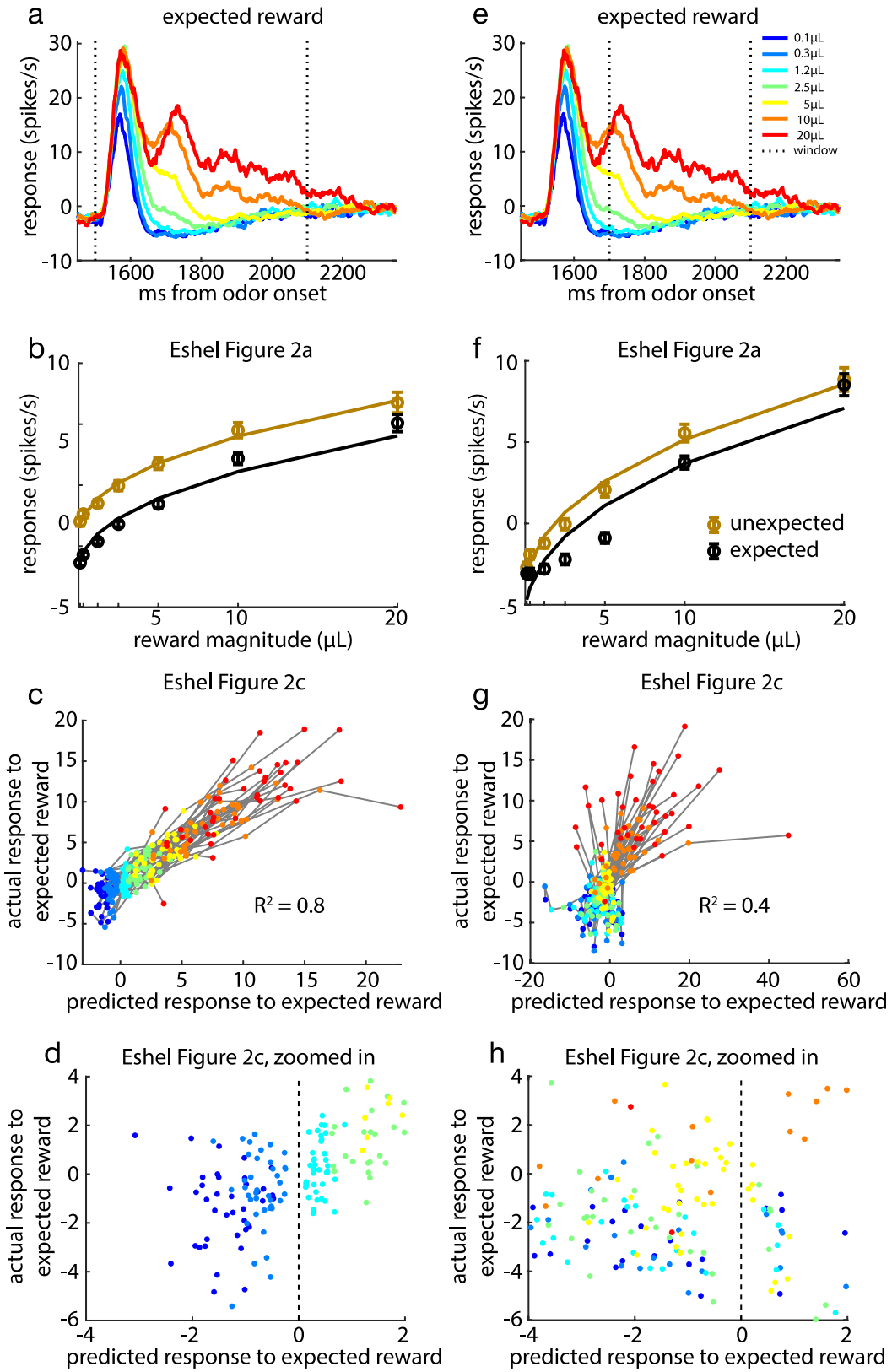
Extended Data Fig. 7 | More details of data in variable-probability task. **a**, Details of analysis method. Of the four possible outcomes of the two Mann–Whitney tests (Methods), two outcomes correspond to interpolation (middle) and one each to the pessimistic (left) and optimistic (right) groups. **b**, Simulation results for the classical TD and distributional TD models. *y* axis shows the average firing-rate change, normalized to mean zero and unit variance, in response to each of the three cues. Each curve is one cell. The cells are split into panels according to a statistical test for type of probability coding (see Methods for details). Colour indicates the degree of optimism or pessimism. Distributional TD predicts simultaneous optimistic and pessimistic coding of probability, whereas classical TD predicts all cells have the same coding. **c**, Same as **b**, but using data from real dopamine neurons. The pattern of results closely matches the predictions from the distributional TD model. **d**, Same as **b**, using data from putative VTA GABAergic interneurons.



Extended Data Fig. 8 | Further distribution decoding analysis. This figure pertains to the variable-magnitude experiment. **a**–**c**, In the decoding shown in the main text, we constrained the support of the distribution to the range of the rewards in the task. Here, we applied the decoding analysis without constraining the output values. We find similar results, although with increased variance. **d**, We compare the quality of the decoded distribution against several controls. The real decoding is shown as black dots. In coloured lines are reference distributions (uniform and Gaussian with the same mean and variance as the ground truth; and the ground truth mirrored). Black traces shift or scale the ground-truth distribution by varying amounts. **e**, Nonlinear functions used to shift asymmetries, to measure degradation of decoded distribution. The normal cumulative distribution function ϕ is used to transform asymmetry τ. This is shifted by some value s and transformed back through the normal quantile function ϕ^{-1}. Positive values s increase the value of τ and negative values decrease the value of τ. **f**, Decoded distributions under different shifts, s. **g**, Plot of shifted asymmetries for values of s used. **h**, Quantification of match between decoded and ground-truth distribution, for each s. **i**, **j**, Same as Fig. 5d, e, but for putative GABAergic cells rather than dopamine cells.



Extended Data Fig. 9 | Simultaneous diversity. a, **b**, Variable-probability task. Mean spiking (**a**) and licking (**b**) activity in response to each of the three cues (indicating 10%, 50% or 90% probability of reward) at time 0, and in response to the outcome (reward or no reward) at time 2,000 ms. **c**, Trial-to-trial variations in lick rates were strongly correlated with trial-to-trial variations in dopamine firing rates. Mean of each cell is subtracted from each axis, and the *x* axis is binned for ease of visualization. **d**, Dopaminergic coding of the 50% cue relative to the 10% and 90% cues (as shown in **b**) was not correlated with the same measure computed on lick rates. Therefore, between-session differences in cue preference, measured by anticipatory licking, cannot explain between-cell differences in optimism. **e**, Four simultaneously recorded dopamine neurons. These are the same four cells whose time courses are shown in Fig. 3c. **f**, Variable-magnitude task. Across cells, there was no relationship between asymmetric scaling of positive versus negative prediction errors, and baseline firing rates ($R = 0.18$, $P = 0.29$). Each point is a cell. These data are from dopamine neurons at reward delivery time. **g**, *t*-statistics of response to 5 µl reward compared with baseline firing rate, for all 16 cells from animal D. Some cells respond significantly above baseline and others significantly below. Cells are sorted by *t*-statistic. **h**, Spike rasters showing all trials in which the 5 µl reward was delivered. The two panels are two example cells from the same animal with rasters shown in Fig. 2.



Extended Data Fig. 10 | Relationship of results to original analysis. Here we reproduce results for the variable-magnitude task in ref. [30] with two different time windows. **a**, Change in firing rate in response to cued reward delivery averaged over all cells. **b**, Comparing Hill-function fit and response averaged over all cells for expected (cued) and unexpected reward delivery. **c**, Correlation between response predicted by scaled common response function and actual response to expected reward delivery. **d**, Zooming in on **c** shows correlation driven primarily by larger reward magnitudes. **e**–**h**, Repeating the above analysis for a window of 200–600 ms.

nature research

Corresponding author(s): Will Dabney, Zeb Kurth-Nelson, Matthew Botvinick

Last updated by author(s): Oct. 31, 2019

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	Simulation experiments were built with custom code and use the following components: Python 2.7 and Tensorflow 1.13.1. Artificial agent experiments are based upon previously published methods.
Data analysis	Data analysis was performed using MATLAB R2018a, NumPy 1.15, and SciPy 1.2.1. Analysis code from our value-distribution decoding analyses, as well as code used to generate model predictions for distributional TD, are available at https://doi.org/10.17605/OSF.IO/UX5RG.

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

Data

Policy information about availability of data

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

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

The neuronal data analyzed in this work have been uploaded to OSF and are available at https://doi.org/10.17605/OSF.IO/UX5RG.

Field-specific reporting

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

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

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

Sample size	All data were taken from a previously-published set of experiments (Eshel et al 2015, 2016).
Data exclusions	No data were excluded.
Replication	Cross-validation was used to avoid over-fitting.
Randomization	There were no between-animal manipulations. Trial sequences were randomized.
Blinding	There were no design factors to which blinding would be relevant.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	The rodent data we re-analyzed here were first reported in Eshel et al. (2015). We used 5 adult male mice, backcrossed for more than 5 generations with C57/BL6J mice, that were heterozygous for Cre recombinase under the control of either the DAT gene or the Vgat gene.
Wild animals	The study did not involve wild animals.
Field-collected samples	The study did not involve field-collected samples.
Ethics oversight	All experiments were performed in accordance with the US National Institutes of Health Guide for the Care and Use of Laboratory Animals and approved by the Harvard Institutional Animal Care and Use Committee.

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