An opponent striatal circuit for distributional reinforcement learning

https://doi.org/10.1038/s41586-024-08488-5

Received: 2 January 2024

Accepted: 4 December 2024

Published online: 19 February 2025

Check for updates

Adam S. Lowet[1,2,3], Qiao Zheng[1,4], Melissa Meng[1,2], Sara Matias[1,2], Jan Drugowitsch[1,4]✉ & Naoshige Uchida[1,2]✉

Machine learning research has achieved large performance gains on a wide range of tasks by expanding the learning target from mean rewards to entire probability distributions of rewards—an approach known as distributional reinforcement learning (RL)[1]. The mesolimbic dopamine system is thought to underlie RL in the mammalian brain by updating a representation of mean value in the striatum[2], but little is known about whether, where and how neurons in this circuit encode information about higher-order moments of reward distributions[3]. Here, to fill this gap, we used high-density probes (Neuropixels) to record striatal activity from mice performing a classical conditioning task in which reward mean, reward variance and stimulus identity were independently manipulated. In contrast to traditional RL accounts, we found robust evidence for abstract encoding of variance in the striatum. Chronic ablation of dopamine inputs disorganized these distributional representations in the striatum without interfering with mean value coding. Two-photon calcium imaging and optogenetics revealed that the two major classes of striatal medium spiny neurons—D1 and D2—contributed to this code by preferentially encoding the right and left tails of the reward distribution, respectively. We synthesize these findings into a new model of the striatum and mesolimbic dopamine that harnesses the opponency between D1 and D2 medium spiny neurons[4–9] to reap the computational benefits of distributional RL.

Midbrain dopamine neurons and their primary target, the striatum, constitute an evolutionarily ancient neural circuit that is critical for motivated behaviours[10]. Computationally, dopamine has long been thought to signal reward prediction error (RPE)[2], reminiscent of the teaching signals used in many RL algorithms[11]. Consistent with this idea, dopamine is also known to modulate plasticity of corticostriatal synapses[12–14], allowing neurons in the striatum to learn a representation of average anticipated reward[15,16], often called ‘value’.

Despite the simplicity and popularity of this model, many aspects of the mesolimbic circuit remain unexplained. First, value representations reside not only in the striatum but also throughout the entire brain[17–19]. Second, the striatum is far from uniform, containing various interneuron subtypes as well as D1 and D2 medium spiny neurons (MSNs), whose plasticity is modulated in opposite directions by dopamine[12–14], and consequently, whose coding properties[4,5] and effects on behaviour[6–9] differ. Third, dopamine activity is much more complex than a simple scalar RPE, varying both qualitatively across dopamine projection systems[20,21] and quantitatively within systems[22,23]. Whether such diversity is cause to revise RPE-based accounts of dopamine[3,24,25] or discard them altogether[26,27] is currently the subject of intense debate.

In parallel to these questions about the neuronal representation of value, the striatum—and particularly the ventral striatum—has long been associated with decision-making under risk. Lesions in the ventral striatum[28] and dopaminergic drugs[29] can both impair risky decision-making, with some groups suggesting a particular role for D2 MSNs in the ventral striatum[30]. Nonetheless, RL models of the basal ganglia typically ignore the role of risk, and most theoretical investigations of uncertainty focus on sensory noise rather than intrinsic, irreducible environmental stochasticity[31,32].

Borrowing from tremendous successes in machine learning[33,34], it has recently been proposed[3] that the residual heterogeneity within RPE-coding dopamine neurons[35–37] and perhaps other neuronal populations[38] resembles the predictions of so-called expectile distributional RL (EDRL)[39]. This algorithm dramatically improves performance relative to traditional RL while unifying the learning of value and risk within the same framework. However, it fails to explain the molecular and functional diversity within the striatum and to rule out alternative explanations for the same dopamine data[40–42].

Here we developed a novel computational model that combines these diverse dopamine inputs[3] with opponent plasticity rules[12–14] to allow D1 and D2 MSNs to learn the right and left tails of the reward distribution, respectively. Our model makes several new experimental predictions about the representational geometry of the striatum, which we confirm using Neuropixels recordings, dopamine lesions, two-photon calcium imaging and optogenetics. Together, this study improves our understanding of the computational principles underlying the reward circuitry of the brain and tightens the bonds between natural and artificial intelligence.

[1]Center for Brain Science, Harvard University, Cambridge, MA, USA. [2]Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA, USA. [3]Program in Neuroscience, Harvard University, Boston, MA, USA. [4]Department of Neurobiology, Harvard Medical School, Boston, MA, USA. ✉e-mail: jan_drugowitsch@hms.harvard.edu; uchida@mcb.harvard.edu

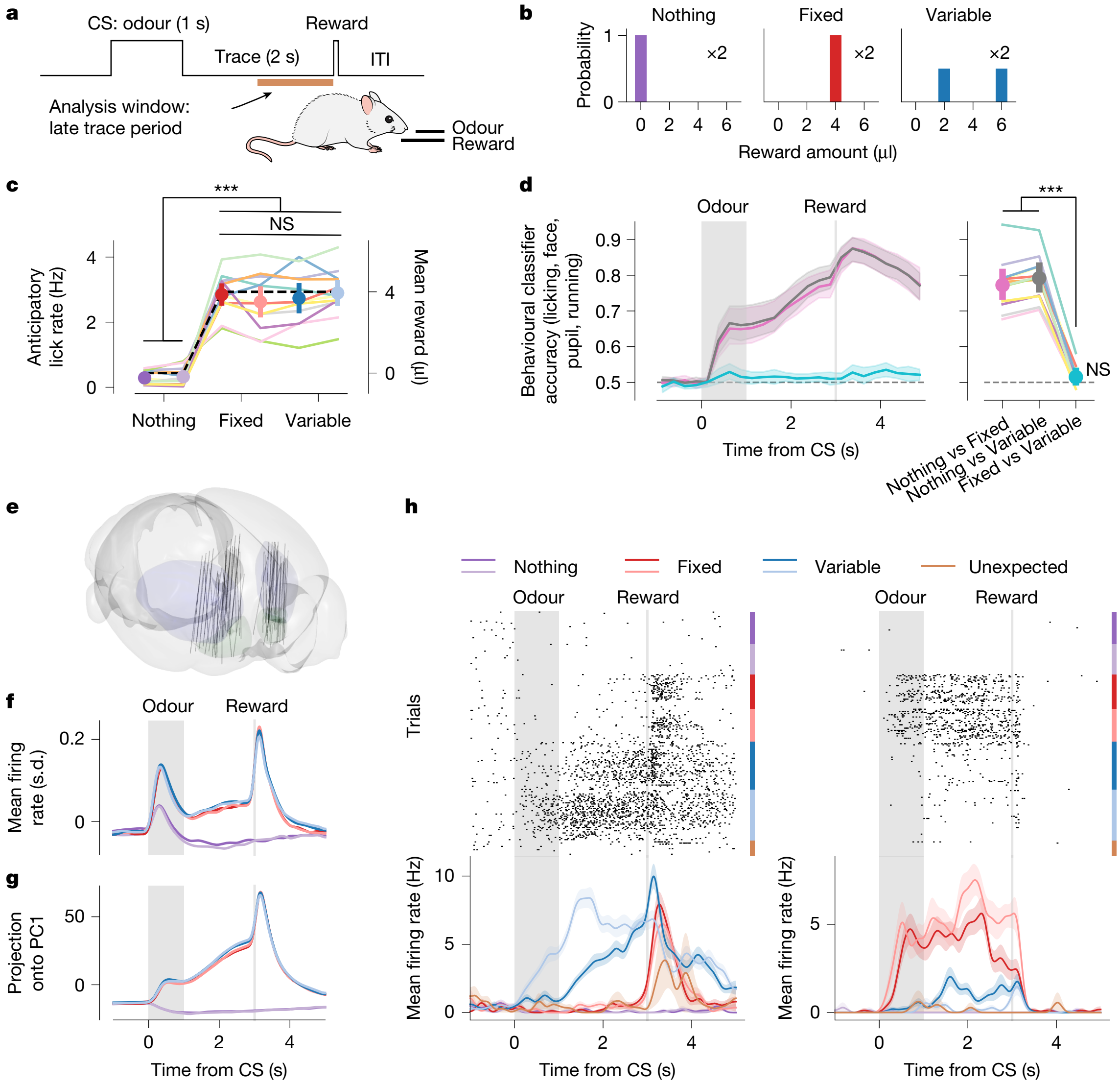


Fig. 1 | A classical conditioning task and recording setup to investigate distributional RL. a, Head-fixed mice were trained to associate odours with stochastic rewards. CS, conditioned stimulus; ITI, inter-trial interval. **b**, Probability distributions over reward amounts, each of which was paired with two unique odours. **c**, Anticipatory lick rates for each trial type, computed during the late trace period (Nothing odours: $P < 0.001$ versus all others; Fixed 1: $P = 0.502$, 0.925 and 0.419 versus Fixed 2, Variable 1 and Variable 2, respectively; $n = 12$ mice, 104 sessions). The dashed line indicates the mean reward for that trial, given on the secondary *y* axis. Error bars indicate 95% confidence interval (c.i.). **d**, Cross-validated accuracy of a linear support vector classifier trained to predict distribution (pooled across odours) on the basis of licking, pupil area, whisking, running and face motion. Behavioural classifier accuracy across time (left), and quantification of classifier accuracy when trained separately on the entire late trace period (right; Fixed versus Variable: $P < 0.001$ versus others, $P = 0.053$ compared with chance level of 50%; $n = 12$ mice, 101 sessions) are shown. Shaded area (left) and error bars (right) denote 95% c.i. across mice. **e**, Reconstructed Neuropixels probe trajectories, aligned to the Allen Mouse Brain Common Coordinate Framework. **f**, Grand average of the *z*-scored firing rates of individual neurons. **g**, Time course of activity across trial types, projected onto PC1. **h**, Example peri-stimulus time histograms of two simultaneously recorded neurons in the ventromedial striatum. Spike rasters, aligned to odour onset and sorted by trial type (top), and mean ± s.e.m. firing to each trial type (bottom) are shown, including trials in which reward was delivered without being preceded by an odour (Unexpected, brown). Where indicated, statistical significance is derived from a linear mixed effects model across sessions with a random intercept (and, if applicable, random slope) for each mouse: ***$P < 0.001$, **$P < 0.01$, *$P < 0.05$ and not significant (NS) at $\alpha = 0.05$.

A behavioural task to investigate distributional RL

Single-unit representations of reward variance have been previously observed in a number of brain regions[43,44], but reports in the striatum have been limited[45,46]. We therefore designed a classical conditioning task in which mice were trained to associate random odour cues with probability distributions over reward amounts (Fig. 1a). Three different probability distributions (Fig. 1b) were used: Nothing (100% chance of a 0 µl reward), Fixed (100% chance of a 4 µl reward) and Variable (50% chance of a 2 µl or 6 µl reward). Fixed and Variable distributions had the same mean but different variance, so distributional RL predicts systematic differences in their underlying neural representations, whereas traditional RL does not. To ensure that any such differences did not reflect idiosyncratic odour preferences, two unique odours predicted each of the three distributions, allowing us to compare representations of different odours both across distributions and within distributions.

Crucially, while the animals' anticipatory licking revealed a clear preference for rewarded (Fixed and Variable) over unrewarded (Nothing) odours, it did not differ between the Fixed and Variable distributions (Fig. 1c). Additional behavioural data, including face motion, whisking, pupil area and running also did not support reliably distinguishing

Fixed from Variable trials[47] (Fig. 1d and Extended Data Fig. 1a–e). This implies that any ability to decode these trial types from neural data must be due to the associated probability distributions and not to differences in motivational value.

The animals' anticipatory licking discriminated all trial types, including Nothing odours, from baseline (Extended Data Fig. 1f), suggesting that meaningful associations were formed with all six odours. Behavioural responses showed minimal trial-by-trial updating; licking (Extended Data Fig. 1g) as well as other behavioural variables (Extended Data Fig. 1h) did not change based on whether the previous Variable reward was greater or less than expected, probably because we were recording from expert mice in a stationary environment.

Striatum represents both mean and variance

Next, we used high-density electrophysiological probes (Neuropixels) to record activity from across a broad swathe of the anterior striatum (Fig. 1e and Extended Data Fig. 2a; $n = 12$ mice, 71 sessions, 13,997 neurons). Consistent with previous work[15,16], we found that both the average firing rate of all neurons (Fig. 1f and Extended Data Fig. 2b) and the time course of trial-type-averaged activity projected onto the first principal component (PC1; Fig. 1g) cleanly separated rewarded from unrewarded odours. Furthermore, a substantial fraction of individual neurons correlated significantly with expected reward, allowing us to reliably predict mean value from neural (pseudo-)population activity across all striatal subregions (Extended Data Fig. 2c–e). Other striatal neurons correlated significantly with RPE during the outcome period[48], but these formed a smaller and mostly independent subset (Extended Data Fig. 2f–h).

However, not all neurons obeyed this simple pattern seen at the level of population averages. Some single neurons consistently preferred Variable odours, whereas others—even when recorded simultaneously—preferred Fixed odours (Fig. 1h). Such neurons fired similarly to both instances of the Fixed and Variable odours, suggesting that they abstracted over odour-specific details to instead encode information about variance—even as the population as a whole contained ample odour information (Extended Data Fig. 2i–l).

To determine whether such distribution coding generalized to the complete population, we compared the cosine distances between the average population activity vectors in the 1-s window before reward delivery (late trace period) for each of the rewarded trial types (representational dissimilarity analysis (RDA)). We found that the distances between across-distribution pairs were greater, on average, than between within-distribution pairs, consistent with distributional RL (Fig. 2a). The same was true for the performance of single-trial linear classifiers applied to pairs of rewarded trial types (Extended Data Fig. 3a,b) or to trial-type groupings that either respected or violated their distribution identities (Extended Data Fig. 3c,d). These latter analyses also confirmed that distributional decoding was orthogonal to mean value coding (Extended Data Fig. 3e–g) and stable over time (Extended Data Fig. 3h–j).

Distributional decoding was strongest in the more ventral and lateral parts of the striatum, particularly the lateral nucleus accumbens shell (lAcbSh; Extended Data Fig. 4a–d). An artificial neural network-based decoder trained on single pseudo-trial population activity from this distribution-coding subpopulation successfully predicted complete reward distributions and generalized to unseen odours (Extended Data Fig. 4e–l). Although we do not claim that the brain decodes distributions in the same manner, this shows, in principle, that there is sufficient information contained in striatal populations to perform distributional RL.

To further exclude alternative explanations for distributional coding, we fit a generalized linear model (GLM) to our data with separate regressors for trial history, reward, (distributional) reward prediction, sensory and motor-related variables (Extended Data Fig. 5; see Methods). Although motor activity explained a high fraction of deviance overall, as seen in previous work[49,50], this trend was not uniform across brain regions. In particular, the ventrolateral parts of the striatum had the relatively weakest encoding of action and the strongest encoding of reward prediction, consistent with the preferential association of the dorsal striatum with motor control and the ventral striatum with state value[15]. Furthermore, motor encoding was uncorrelated with other task variables, whereas trial history and reward responses were positively correlated with distribution coding, suggesting that striatal neurons multiplex certain additional variables, but not behaviour, with reward prediction (see Extended Data Fig. 2h). Nonetheless, the magnitude of reward and (especially) trial history coding was weaker than that of reward prediction, making trial-by-trial updates unlikely to drive the observed differences between Fixed and Variable trials (Extended Data Fig. 5g–i).

Finally, an alternative hypothesis is that rather than represent reward variance in their mean firing rates within a trial, neurons instead encode reward variance in their spiking variability across trials[51]. However, across-trial variability was the same across trial types with different variances, ruling out such 'sampling-based codes' in this instance[52] (Extended Data Fig. 6).

Variance coding is abstract and at the population level

The preceding analyses show that the neural activities evoked by odours identifying the same distribution are more similar to one another than to those evoked by odours identifying distributions with the same mean but different variances. Let us now ask about the relationship between Fixed and Variable odour representations—specifically, whether variance is represented in an 'abstract format', that is, in a consistent way across odours that would support generalization to unseen situations[53]. To do so, we adapted two previously defined metrics[53] to our task: parallelism score and cross-condition generalization performance (CCGP; see Methods).

The parallelism score is simply the average cosine similarity between the two difference vectors pointing from Variable to Fixed population activity, one for each odour identifying the respective distribution. Across sessions and mice, these difference vectors were significantly more aligned than would be expected by chance (Fig. 2b). Similarly, a decoder trained on one Fixed versus Variable dichotomy and then tested on the held-out dichotomy achieved above-chance CCGP, averaged across all four possible dichotomies (Fig. 2c).

Consistent coding of variance could arise due to linear encoding of reward variance in single-neuron firing rates, as has been observed in other brain regions[43–45]. However, unlike these previous studies, we found fewer striatal neurons encoding variance (or conditional value at risk, another risk measure) than would be predicted simply from the combination of mean reward and odour coding alone (Extended Data Fig. 2m–r). Variance coding in the striatum is thus an intrinsically population-level phenomenon.

Using striatal opponency for distributional RL

Next, we explored how such an abstract representation might be acquired. Although there exist multiple theories for how the brain could learn abstract reward distributions[33,40], EDRL[39] is especially promising because it requires only minimal modifications to existing, empirically tested models of the basal ganglia[3]. EDRL proposes not just a single value predictor but an entire family of predictors, V_i (parameterized by τ_i, the degree of optimism), each of which converges to an 'expectile' of the reward distribution. Expectiles generalize the mean just as quantiles generalize the median, and collectively, they completely characterize a probability distribution[54] (Fig. 2d; see Methods).

Although EDRL has some appealing properties, it ignores the cellular diversity within the striatum, most notably the presence of D1 and D2 MSNs[55]. Instead, we start with the same piecewise linear

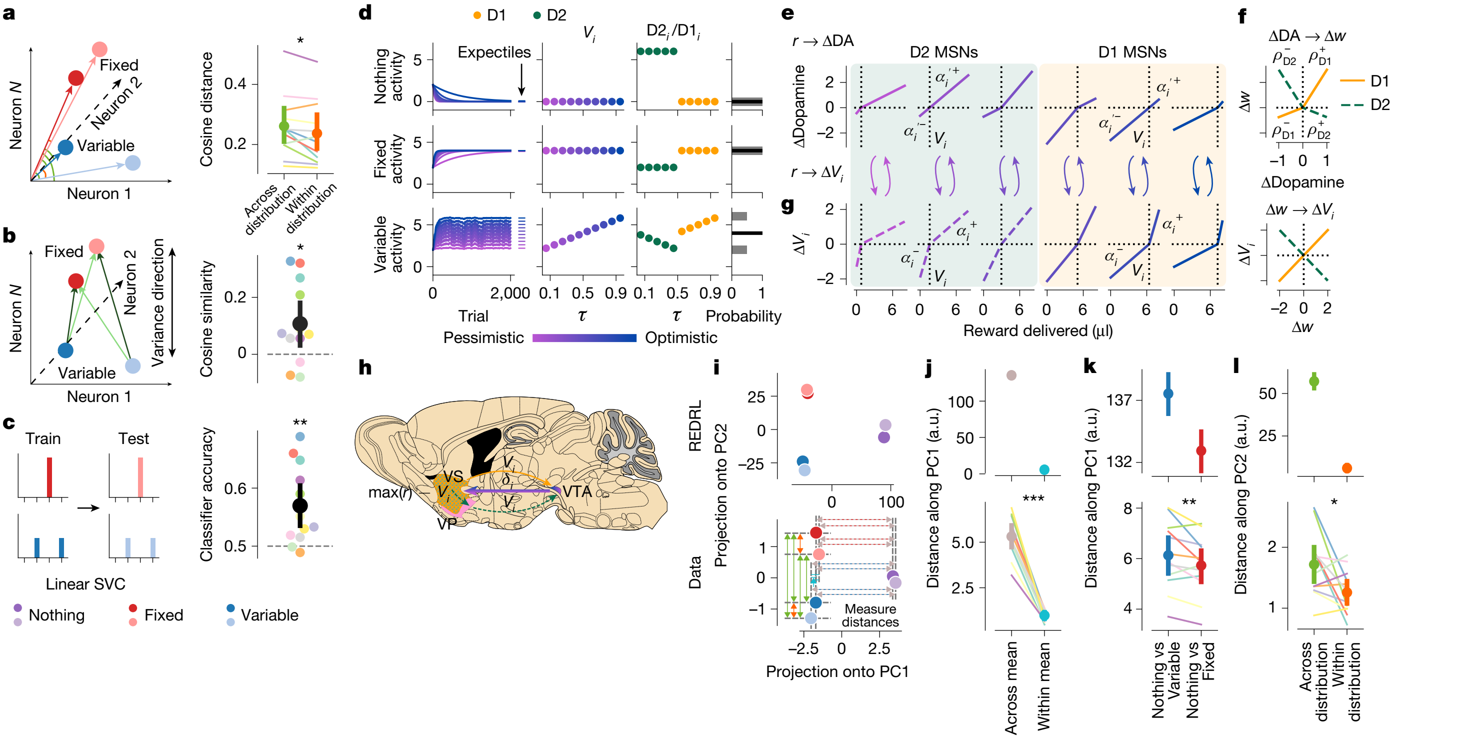


Fig. 2 | REDRL explains distributional coding across the striatum. **a**, Schematic (left) and quantification (right) of representational dissimilarity between across-distribution (green) and within-distribution (orange) pairs ($P = 0.027$). **b**, Schematic (left) and quantification (right) of parallelism score, defined as the cosine similarity between the difference between Fixed − Variable vectors, averaged over the two possible combinations (dark and light green; $P = 0.015$ relative to chance value of zero, grey dashed line). **c**, Schematic (left) and quantification (right) of CCGP ($P = 0.001$ relative to chance value of 0.5, grey dashed line). SVC, support vector classifier. **d**, Algorithmic REDRL model. With learning, value predictors (V_i) converge to the τ_i-th expectiles of the associated reward distribution. D1 MSN activity ($\tau > 0.5$) is equal to V_i, whereas D2 activity ($\tau < 0.5$) is sign-flipped and offset. **e**–**g**, Implementation of REDRL. Dopaminergic neurons are modelled using piecewise linear functions, with slopes $\alpha_i'^-$ and $\alpha_i'^+$ in the negative and positive domain, respectively, and zero-crossing points equal to the τ_i-th expectile (vertical dotted lines; **e**)[3]. D1 and D2 MSNs have complementary, nonlinear plasticity rules[13] (**f**, top), leading to positive and negative encoding of their respective value predictions[4,5] (**f**, bottom). The net result is that D1 and D2 MSNs are biased optimistically and pessimistically relative to their dopamine input asymmetries (**g**). r, reward; w, synaptic weight. **h**, Hypothesized circuit basis[74] of REDRL. VTA dopaminergic neurons convey distributional RPEs (δ_i) to the ventral striatum (VS). D1 MSNs feedback directly to optimistic VTA neurons, whereas D2 MSNs are routed to pessimistic VTA neurons via the ventral pallidum (VP). The schematic in panel **h** was adapted from ref. 74, Elsevier. **i**, 2D principal component projection of REDRL value predictors (top) and an example recording session (bottom). Grey dashed lines denote positions along PC1 and PC2 from which distances (coloured arrows) were measured. **j**–**l**, Euclidean distances along the indicated principal components are consistent between model (top) and data (bottom; PC1 across versus within mean: $P < 0.001$ (**j**); PC1 Nothing versus Fixed or Variable: $P = 0.005$ (**k**); and PC2 across versus within distribution: $P = 0.012$ (**l**)). Data in **a**–**c** and **j**–**l** (bottom) are mean ± 95% c.i. across mice, with statistical significance as indicated in the caption for Fig. 1.

heterogeneity in dopamine responses[3] (Fig. 2e) but combine it with an opponent plasticity rule (Fig. 2f, top) in which D1 MSNs increase their synaptic weights more from positive RPEs (ρ_m^+), whereas D2 MSNs increase their synaptic weights more from negative RPEs[12–14] (ρ_m^-). Because of the symmetry between D1 and D2 plasticity functions, we call our implementation reflected EDRL (REDRL).

The opponency of the plasticity rule gives rise to opposite directions of value coding (Fig. 2f, bottom), with D1 MSNs[4,5,56,57] and D2 MSNs[4,5,58] primarily correlating positively and negatively, respectively, with value. Meanwhile, its piecewise linear nature has the effect of extremizing value predictors—D1 MSNs are more optimistic, and D2 MSNs are more pessimistic, than their individual dopamine inputs would create on their own—while nonetheless converging mathematically to expectile estimates (Fig. 2g). The ventral pallidum, which predominantly receives projections from D2 MSNs[59], adds an extra inhibitory synapse and thereby flips the sign of this input before feeding these pessimistic value predictions back to dopamine neurons (Fig. 2h).

Validating the geometry of REDRL in striatal data

REDRL not only gives rise to abstract coding of variance in the striatum (Fig. 2a–c) but also makes specific predictions about the population geometry of striatal representations, which can then be compared to data and to alternative models (Extended Data Fig. 7a–m). Specifically, we projected either the REDRL value predictors, or the trial-type-averaged firing rates of each session, onto their first and second principal components (accounting for 73.3 ± 2.3% and 10.9 ± 0.9% of the variance across trial types, respectively; mean ± s.e.m. across mice; see Methods). We then measured the Euclidean distances in principal component space along each dimension (Fig. 2i).

PC1 mainly separated trial types according to their means, as expected (Fig. 2j). More surprisingly, but also consistent with REDRL, Variable odours elicited higher average firing rates than Fixed odours (Extended Data Fig. 7n) and so were more distant from Nothing odours along PC1 (Fig. 2k). Fixed and Variable odours also separated out along PC2, such that there was a greater distance between across-distribution odour pairs than within-distribution odour pairs (Fig. 2l). Across the population, substantial fractions of neurons correlated positively or negatively with expected reward across trials (Extended Data Fig. 7o), as would be expected for D1 and D2 MSNs, respectively, in REDRL. Although other distributional RL formulations predicted some of these effects, only REDRL and its close cousin, reflected quantile DRL, predicted all of them.

We demonstrated further support of REDRL across three additional classical conditioning tasks in three independent cohorts of mice, in which REDRL continued to predict the population geometry (Extended Data Fig. 8a–e) and single-neuron encoding properties (Extended Data Fig. 8f) of our recordings. In particular, we replicated the core findings that PC1, along with a sizeable portion of individual neurons, represents mean reward for these particular distributions. Distributions with the same mean but different higher-order moments separate out along PC2, without many individual neurons linearly encoding reward variance, including in the Bernoulli task, in which mean and variance were orthogonal by design. These features were most similar to the theoretical predictions of REDRL (Extended Data Fig. 8d).

A different set of distributions, which we call the Fourth Moments task, featured pairs of distributions (Uniform and Bimodal) with the same mean, variance and skewness, differing only at the fourth moment and above. Licking to the Bimodal distribution was modestly weaker than to Uniform in this cohort of mice, leading to separation along PC1 in addition to PC2, in contrast to theoretical predictions (Extended Data Fig. 8b–e). Nonetheless, the structure of this task allowed us to run more rigorous tests of distribution coding—CCGP, parallelism score, pairwise decoding and congruency analysis—following our approach for the main task. Ventrolateral subregions of the striatum, particularly the lAcbSh and core, continued to show signatures of distributional representations even with these more closely matched distributions, extending such coding as high as the fourth moment (Extended Data Fig. 8g–j).

Thus, REDRL provides a mechanistic account of distributional RL which quantitatively matches the structure of striatal representations across a diverse range of probability distributions. This permits us to reinterpret single-neuron activities in the striatum as linearly encoding specific (linear combinations of) expectiles of the reward distribution, explaining our ability to decode reward variance from neuronal populations in the absence of strong single-neuron variance correlations (Extended Data Figs. 2m–o and 8f).

Dopamine is necessary for distributional RL

If striatal representations are updated incrementally by dopamine RPEs as predicted by REDRL, then eliminating dopamine before learning should disrupt these distributional representations (Fig. 3a). To test this hypothesis, we injected the neurotoxin 6-hydroxydopamine (6-OHDA) unilaterally into the lateral ventral striatum in naive mice, which resulted in local lesions of dopamine neurons projecting to the injection site (Fig. 3b,c and Extended Data Fig. 9a). After recovery, we trained the mice on the original task and then recorded neurons in both the control and the lesioned hemisphere ($n = 5$ mice, 20 sessions, 2,283 neurons from control; 19 sessions, 2,596 neurons from lesion). Unilateral lesions modestly impaired our ability to distinguish rewarded and unrewarded odours based on behavioural predictors, but mice nonetheless learned the task (Extended Data Fig. 9b,c), and neural encoding of motor behaviour and other variables was similar in the two hemispheres, as measured by our GLM (Extended Data Fig. 9d–f).

Projecting striatal activity from each hemisphere independently into principal component space suggested that distributions were less well separated in the lesioned hemisphere than in the control hemisphere (Fig. 3d). Indeed, when we quantified distances as before, we found unrewarded (Nothing) and rewarded (Fixed and Variable) odours to be equally well separated along PC1 for both hemispheres (Fig. 3e), but Fixed and Variable odours to be less well separated along PC2 in the lesioned hemisphere (Fig. 3f). Analogous effects were seen for parallelism score (Fig. 3g) and representational dissimilarity (Fig. 3h), with stronger (and abstract) variance coding in the control than in the lesioned hemisphere. The persistence of mean value coding in the lesioned hemisphere may reflect the inability of unilateral 6-OHDA to kill all dopamine neurons within the targeted hemisphere, the interhemispheric broadcasting of mean value information once it reaches the cortex[17–19], or, more radically, the dispensability of dopamine for learning about mean value entirely.

In addition to supporting our mechanistic REDRL model, the selective disruption of variance coding by 6-OHDA gives us an experimental tool with which to probe the role of distributional RL in the brain. When paired with deep neural networks, distributional RL is thought to boost performance mainly by improving state representations[1,3,60]. Because odour-specific information is multiplexed alongside reward distributions in the striatum (Extended Data Fig. 2i–l), it is possible to ask whether dopamine lesions—by perturbing distributional RL—also impair striatal representations of stimulus identity. We used multinomial logistic regression to decode odour identity from neural activity during the 1-s window following odour onset. Although we could decode odour identity well above chance for both hemispheres, decoding performance was significantly worse in the lesioned than in the control hemisphere (Fig. 3i). The lesion impaired decoding performance across nearly all trial types, with the main driver being increased confusion between Fixed and Variable odours (Fig. 3j,k). These results are consistent with distributional RL shaping the representation of sensory inputs in biological brains, similarly to its role in artificial neural networks.

Opponent contributions of D1 and D2 MSNs to REDRL

We next tested the distinct contributions of D1 and D2 MSNs predicted by REDRL. A Ca^{2+} indicator, jGCaMP7s, was expressed in D1 or D2 MSNs in the lAcbSh (D1 $n = 4$ mice, 27 sessions, 945 neurons; D2 $n = 4$ mice, 38 sessions, 1,106 neurons), and single-neuron activity was monitored using two-photon calcium imaging through implanted gradient refractive index lenses (Fig. 4a–c; 31.6 ± 17.4 cells per field of view, mean ± s.d. across sessions).

We observed different patterns of activity across D1 and D2 populations[56–58] despite the fact that behaviour did not differ across groups (Extended Data Fig. 10a–c). Many D1 MSNs were activated more to rewarded than to unrewarded odours and outcomes, whereas the reverse was true, albeit less strongly, in D2 MSNs (Fig. 4d–f). Also consistent with our model, significant fractions of D1 and D2 MSNs increased and decreased their activities relative to baseline, respectively, more on rewarded than on unrewarded trials; however, the pattern in D2 MSNs was again more heterogeneous than in D1 MSNs, with less consistent variability across trial types (Extended Data Fig. 10d,e).

We also found neurons that, like those that we recorded using electrophysiology, reliably distinguished between Fixed and Variable odours during the late trace period (Fig. 4g). To test whether these trends were systematic, we performed the same analyses (CCGP, RDA and principal component analysis (PCA)) separately on D1 and D2 MSNs, while pooling across all mice to compensate for the lower cell counts and higher variability of Ca^{2+} signals. Consistently across disjoint subsets of pseudo-trials in both D1 and D2 MSNs, variance was encoded in an abstract format (Fig. 4h), and across-distribution pairs were represented more dissimilarly than within-distribution pairs (Fig. 4i).

REDRL not only predicts the existence of distributional coding in D1 and D2 MSNs independently but also specifies the ways in which this coding should differ. For example, pessimistic ($\tau < 0.5$) REDRL predictors associate Variable odours with lower-than-average rewards. We therefore expect their representation of Nothing odours to be more similar to that of Variable odours than to Fixed odours, whether assessed via PCA or RDA. Meanwhile, the opposite should be true of optimistic ($\tau > 0.5$) predictors (Fig. 4j–m). D1 and D2 MSNs

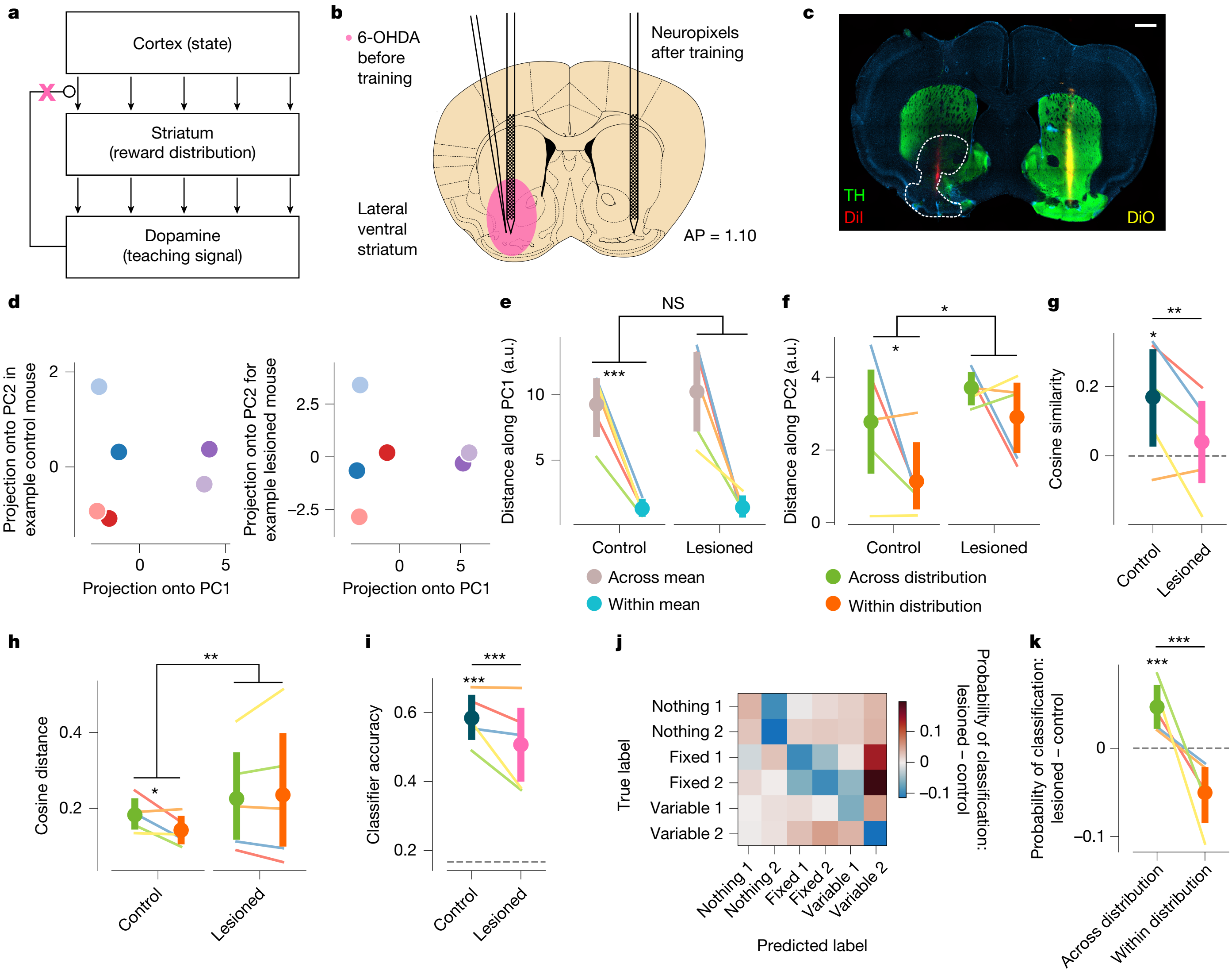


Fig. 3 | Dopamine is necessary for learning distributional representations. **a**, Dopamine lesions (pink 'x') are predicted to disrupt representations of the reward distribution in the striatum. **b**, Schematic[74] of the dopamine lesion experiment ($n = 5$ mice). AP, anteroposterior. The schematic was adapted from ref. 74, Elsevier. **c**, Histology from an example 6-OHDA mouse showing Neuropixels probe tracks (red and yellow), dopaminergic axons (green; tyrosine hydroxylase (TH)) and lesion boundary (white dashed line). Scale bar, 500 µm. **d**, Principal component projection from the control (left) and lesioned (right) hemispheres for an example mouse. **e**, Distance along PC1, although significantly higher for across-mean than for within-mean pairs ($P < 0.001$), does not differ between hemispheres ($P = 0.676$). **f**, By contrast, the difference in distance along PC2 between across-distribution and within-distribution pairs is significantly positive ($P = 0.033$) and greater in the control than in the lesioned hemisphere ($P = 0.026$). **g**, Parallelism score is significantly positive ($P = 0.029$) and greater in the control than in the lesioned hemisphere ($P = 0.009$). **h**, Similarly, the difference in representational dissimilarity between across-distribution and within-distribution pairs is significantly positive ($P = 0.036$) and greater in the control than in the lesioned hemisphere ($P = 0.005$). **i**, Six-way odour classification accuracy during the odour period is above chance ($P < 0.001$) and is higher in the control than in the lesioned hemisphere ($P < 0.001$). **j**, Difference in odour classifier confusion matrices between the lesioned and control hemispheres. The probability of correct classification (main diagonal) decreases for nearly all trial types upon lesioning. **k**, The decrement in odour coding due to the lesion is mainly due to an increase in across-distribution, within-mean classification errors ($P < 0.001$) and a concomitant decrease in within-distribution classification ($P < 0.001$ for across-distribution versus within-distribution difference). Data in **e**–**i**, **k** are mean ± 95% c.i. across mice, with statistical significance as indicated in the caption for Fig. 1.

mirrored these predictions precisely (Fig. 4n–q), strongly supporting the notion that they encode the right and left tails of the reward distribution, respectively.

Perturbing REDRL with optogenetics

As a final test of REDRL, we sought to independently manipulate D1 and D2 MSNs while mice performed a similar classical conditioning task. To do so, we expressed either the excitatory opsin CoChR ($n = 12$ mice, 96 sessions) or the inhibitory opsin GtACR1 ($n = 13$ mice, 92 sessions) in D1 or D2 MSNs and implanted an optical fibre in lAcbSh (Fig. 5a). We then manipulated these neurons during the 2-s trace period and quantified licking just before reward delivery (Fig. 5b).

To generate model predictions for these manipulations, we clamped the simulated values of inhibited and excited predictors, respectively, at 0 µl and 8 µl, the maximum reward size that we delivered in these experiments. We performed these simulated manipulations separately on optimistic and pessimistic predictors and computed the predicted value estimate of the mouse as the mean across all predictors (Fig. 5c,d and Extended Data Fig. 11a–i; see Methods). We then took the difference

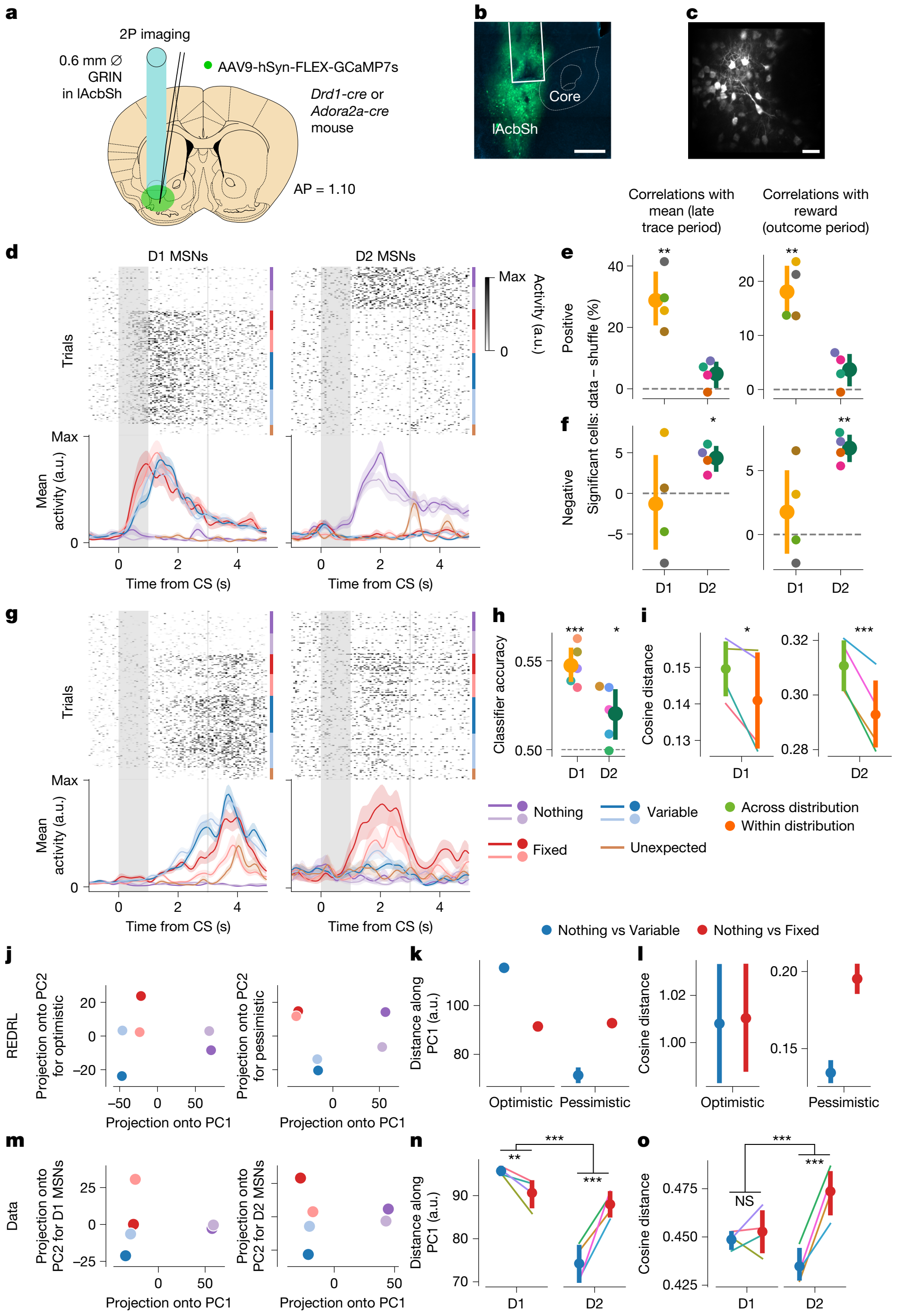


Fig. 4 | Opponent contributions of D1 and D2 MSNs to distributional coding. **a**, Schematic of the two-photon (2P) calcium imaging experiment. GRIN, gradient refractive index lens. The schematic was adapted from ref. 74, Elsevier. **b**,**c**, Example slice (**b**) and field of view (**c**) showing expression of GCaMP7s in the lAcbSh of a *Drd1-cre* mouse. Scale bars: 500 µm (**b**) and 50 µm (**c**). **d**, Deconvolved Ca^{2+} activity from example D1 (left) and D2 (right) MSNs, as in Fig. 1h. Shaded grey bar denotes period of odour delivery. Bottom, shaded regions denote mean ± s.e.m. across trials. **e**, Percentage of significant cells that correlate positively with mean (left) or reward (right) during the late trace and outcome periods, respectively. There are more cells than expected by chance (grey dashed line) for D1 (paired samples Student's *t*-test: $P = 0.009$ and 0.006; mean ± s.e.m. = 28.79 ± 4.78 and 18.06 ± 2.58 for mean and reward, respectively), but not D2 ($P = 0.113$ and 0.107; mean ± s.e.m. = 4.90 ± 2.21 and 3.68 ± 1.61, for mean and reward, respectively; $n = 8$ mice). **f**, Same as panel **e**, but for significant negative correlations. There are more cells than expected for D2 ($P = 0.013$ and 0.001; mean ± s.e.m. = 4.39 ± 0.81 and 6.76 ± 0.56, for mean and reward, respectively) but not D1 ($P = 0.736$ and 0.433; mean ± s.e.m. = −1.29 ± 3.48 and 1.76 ± 1.95, for mean and reward, respectively). **g**, Same as panel **d**, but showing MSNs that discriminate Fixed and Variable odours. **h**, CCGP is above chance (grey dashed line) for both D1 (one-sample Student's *t*-test, $P < 0.001$, mean ± s.e.m. = 0.0473 ± 0.0051) and D2 ($P = 0.048$; mean ± s.e.m. = 0.0202 ± 0.0072; $n = 5$ pseudo-populations per genotype). **i**, For representational dissimilarity, cosine distance is greater for across-distribution than for within-distribution pairs for both D1 ($P = 0.022$; $n = 4$ pseudo-populations per genotype) and D2 ($P < 0.001$; $n = 4$ pseudo-populations per genotype). **j**, 2D principal component plots for simulated optimistic and pessimistic REDRL value predictors. **k**,**l**, Predicted distance along PC1 (**k**) and RDA (**l**) for the REDRL model, averaged across $n = 4$ odour pairs. **m**–**o**, Same as panels **j**,**k**, but showing data collected from D1 and D2 MSNs (distance along PC1: $P = 0.001$ for D1, $P < 0.001$ for D2 and the relative differences; RDA: $P = 0.489$ for D1, $P < 0.001$ for D2 and the relative differences; $n = 4$ pseudo-populations per genotype). Unless otherwise noted, data in **h**,**i**,**n**,**o** are mean ± 95% c.i. across pseudo-populations, and statistical significance is as indicated in the caption for Fig. 1.

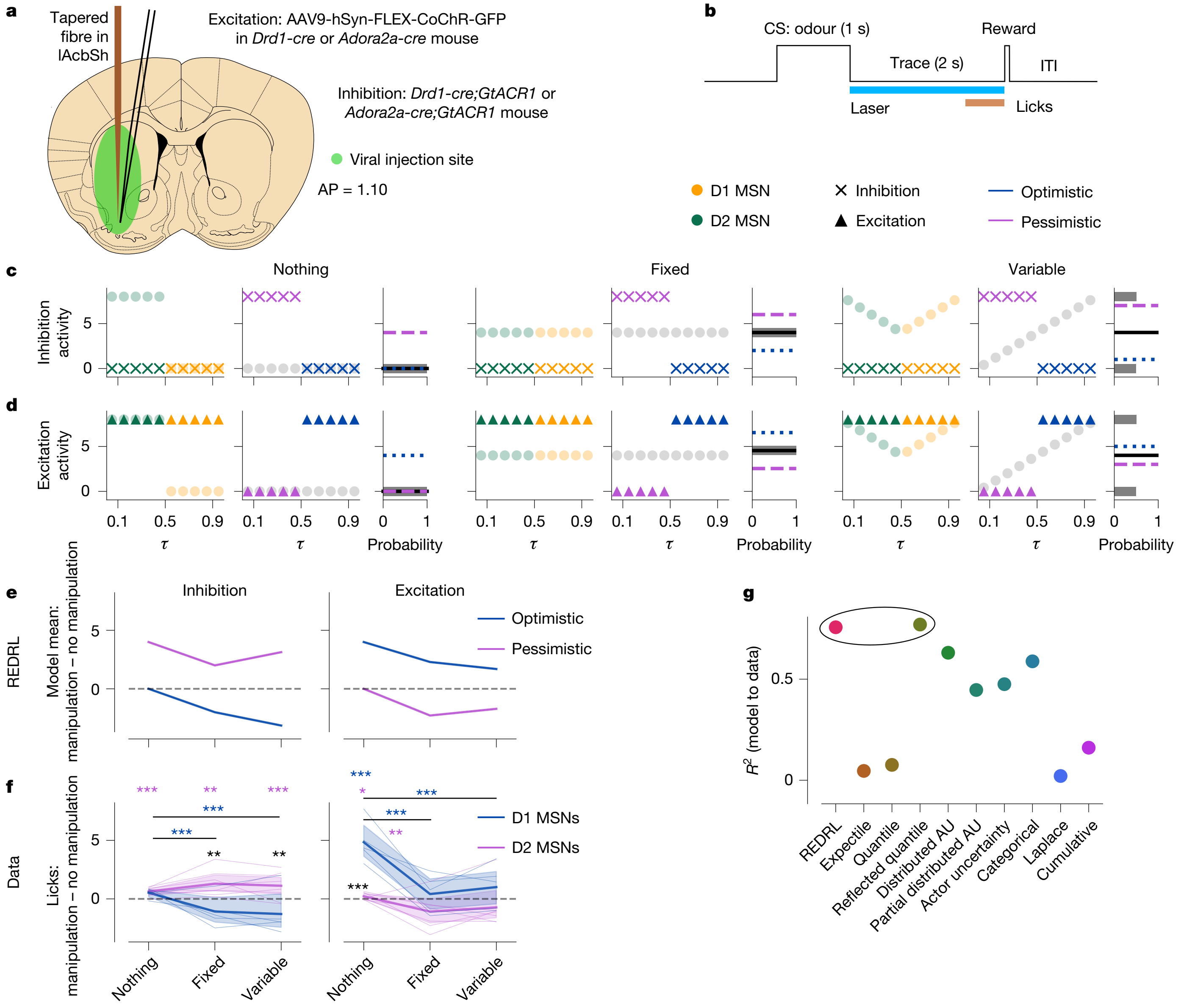


Fig. 5 | Causal contributions of D1 and D2 MSNs to REDRL. a,**b**, Schematic of the optogenetic experiments (**a**) and trial structure (**b**). The schematic in panel **a** was adapted from ref. 74, Elsevier. **c**, Approach for simulating the effects of optogenetic inhibition in the REDRL model (see Methods). Within each group of panels (Nothing, Fixed and Variable), the left column shows the predicted D1 (yellow) and D2 (green) activities for the no manipulation (grey faded circles) and manipulation ('x') conditions. The middle column shows the resulting effect on the encoded V_i. The right column illustrates the effect that this change in V_i has on the encoded mean (blue and purple horizontal dashed lines), relative to the unperturbed distribution (grey histogram, with mean shown in black). **d**, Same as panel **c**, but for optogenetic excitation (triangles) rather than inhibition. **e**, Summary of REDRL model predictions for differences between manipulation versus no manipulation trials (zero difference, grey dashed line). **f**, Difference in anticipatory licking between manipulation and no manipulation trials, computed within session and then averaged across sessions and within mice (thin lines) and coloured as in panel **e**. The coloured asterisks with horizontal lines denote significant differences in the effect of manipulation between trial types within the indicated genotype (D1 inhibition: $P < 0.001$ Nothing versus Fixed or Variable; D1 excitation: $P < 0.001$ Nothing versus Fixed or Variable; D2 excitation: $P = 0.007$, Nothing versus Fixed). The coloured asterisks over single trial types indicate significant differences relative to zero (grey dashed line) for that genotype (D2 inhibition: $P < 0.001$ Nothing, $P = 0.002$ Fixed, $P < 0.001$ Variable; D1 excitation: $P < 0.001$ Nothing; D2 excitation, $P = 0.032$ Nothing). The black asterisks over single trial types indicate significant differences between genotypes (inhibition: $P = 0.001$ Fixed, $P = 0.005$ Variable; excitation: $P < 0.001$ Nothing), and statistical significance is as indicated in the caption for Fig. 1. **g**, Summary panel showing the mean coefficient of determination for each model, when predicting the average difference in licking across trials. AU, actor uncertainty.

between the estimated mean values of the model in manipulation versus no manipulation trials for each trial type (Fig. 5e and Extended Data Fig. 11j,k) and compared them with the differences in anticipatory licking by the mouse.

REDRL not only captured the main effects of 'go' and 'no-go' pathways[61] but also predicted precise patterns of licking across trial types, even for the same type of manipulation (Fig. 5f). This could not be explained simply by ceiling effects, as the increase in licking was sometimes greater for rewarded than for unrewarded odours, and average lick rates were far below physiological limits (Extended Data Fig. 11n). Quantitative comparison confirmed that reflected expectile-like models outperformed alternatives in fitting the licking data (Fig. 5g), arguing that the value predictions learned by REDRL are used online to guide behaviour.

Discussion

Here we have combined large-scale electrophysiology with cell-type-specific recordings and manipulations to develop the REDRL model of the basal ganglia. This model maintains the algorithmic advantages of distributional RL[1] while lending itself to a biological implementation that is consistent with observed dopamine population activity[3] and dopamine-mediated plasticity rules[12–14], as well as the hypothesized computational role of dopamine as a RPE signal (as opposed to directly influencing causal associations[26] or learning rate[27]; see Supplementary Discussion). This dopamine activity is also what led us to favour REDRL over the conceptually similar reflected quantile code, which made similar predictions for MSNs across the distributions that we tested, but differed at the level of dopamine neurons (see Supplementary Discussion).

The most notable feature of REDRL is the distinct roles played by D1 and D2 MSNs, which specialize in the right and left tails of the reward distribution, respectively. This bifurcated layout resembles other neural systems, such as ON/OFF pathways in vision, and probably has similar benefits, such as efficient coding[62], flexibility[63] and perhaps robustness to noise (Extended Data Fig. 8d). For example, certain computations, such as expected value estimation, would benefit from combining information from D1 and D2 MSNs, but others, such as risk-sensitive behaviour, might depend on just one tail (and thus neuronal cell type) or the other. Furthermore, this architecture simplifies the problem of connectivity: genetically defined subsets of dopamine neurons[64] could form independent closed loops with D2 (via the ventral pallidum) and D1 MSNs, thereby helping to keep separate pessimistic and optimistic RPE channels (Fig. 2h). These predictions should form the basis of future anatomical investigations into the mesolimbic dopamine circuitry, as well as theories of alternative architectures that might obviate this need[65], which is shared by EDRL. It will also be important to record from dopamine neurons in similar task settings, to ensure that their degree of optimism is consistent across different cues[35] and perhaps organized topographically[22].

At the level of the striatum, REDRL helps to unify previous approaches to understanding D1 and D2 MSNs within a single, normative framework. Although D1 and D2 MSNs are known to frequently behave in an opponent manner[4–9], this has generally been attributed to go/no-go pathways and modelled using a single value predictor or action channel[61,66]. Here we have shown how, far from being a bug or redundancy in the RL architecture, such diversity could actually be a feature, biasing convergence to optimistic or pessimistic value predictors. More speculatively, it could also explain why D1 and D2 MSNs are not simply inverses of each other[67–69]. The tendency for both pathways to activate before movement onset, for example, may not only be a consequence of the role of the (dorsal) striatum in action selection; such co-activation would also be predicted for the ventral striatum if these transition points coincide with increases in the predicted variance (and thus the density on both the left and the right tails) of the reward distribution.

The present studies could only infer motivational value, and not risk attitudes, from the conditioned responding of mice, and many questions remain as to how the brain transforms high-dimensional reward distributions into a single choice. Nonetheless, it is tempting to speculate that this process corresponds to the dimensionality reduction that takes place throughout the various nuclei of the basal ganglia[70], ultimately collapsing onto a unitary value estimate in the thalamus that defines the choice axis. Notably, such a 'distributional critic'—centred here in the lAcbSh, a region that receives RPE-like dopamine input[20–22]—could integrate seamlessly into a broader RL framework[71], with the dorsal striatum likely playing the role of the 'actor' and choosing actions in continuous, high-dimensional spaces. More work is needed in operant or unconstrained contexts—including those requiring response inhibition[9,58]—to establish the ubiquity of distributional representations of reward, explore how such coding intersects with choice and tease apart candidate neural architectures. Furthermore, it remains to be clarified how distributional information may help to tune state representations in the cortex without making use of backpropagation—and, more generally, how MSN activity evolves with learning.

Modifications of the encoded reward distribution, such as by dopaminergic drugs[29], or of the downstream basal ganglia circuit, might bias risky choice on rapid or developmental timescales[30]. Various psychopathologies—such as depression, in which patients learn more from losses than gains[72], or addiction, in which patients systematically overweight the right tail of the reward distribution[73]—could similarly stem from the dysfunction of this core distributional RL circuitry. Thus, REDRL can serve as a bridge between RL, behavioural economics, computational psychiatry and systems neuroscience, demonstrating how the circuit logic of the striatum can combine with vector-valued dopamine signals to realize the computational benefits of distributional RL.

Online content

Any methods, additional references, Nature Portfolio 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/s41586-024-08488-5.

1. Bellemare, M. G., Dabney, W. & Rowland, M. *Distributional Reinforcement Learning* (MIT Press, 2023).
2. Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science* **275**, 1593–1599 (1997).
3. Dabney, W. et al. A distributional code for value in dopamine-based reinforcement learning. *Nature* **577**, 671–675 (2020).
4. Shin, J. H., Kim, D. & Jung, M. W. Differential coding of reward and movement information in the dorsomedial striatal direct and indirect pathways. *Nat. Commun.* **9**, 404 (2018).
5. Nonomura, S. et al. Monitoring and updating of action selection for goal-directed behavior through the striatal direct and indirect pathways. *Neuron* **99**, 1302–1314.e5 (2018).
6. Hikida, T., Kimura, K., Wada, N., Funabiki, K. & Nakanishi, S. Distinct roles of synaptic transmission in direct and indirect striatal pathways to reward and aversive behavior. *Neuron* **66**, 896–907 (2010).
7. Kravitz, A. V., Tye, L. D. & Kreitzer, A. C. Distinct roles for direct and indirect pathway striatal neurons in reinforcement. *Nat. Neurosci.* **15**, 816–818 (2012).
8. Tai, L.-H., Lee, A. M., Benavidez, N., Bonci, A. & Wilbrecht, L. Transient stimulation of distinct subpopulations of striatal neurons mimics changes in action value. *Nat. Neurosci.* **15**, 1281–1289 (2012).
9. Cruz, B. F. et al. Action suppression reveals opponent parallel control via striatal circuits. *Nature* **607**, 521–526 (2022).
10. Floresco, S. B. The nucleus accumbens: an interface between cognition, emotion, and action. *Annu. Rev. Psychol.* **66**, 25–52 (2015).
11. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction* Vol. 2 (MIT Press, 2018).
12. Yagishita, S. et al. A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science* **345**, 1616–1620 (2014).
13. Iino, Y. et al. Dopamine D2 receptors in discrimination learning and spine enlargement. *Nature* **579**, 555–560 (2020).
14. Lee, S. J. et al. Cell-type-specific asynchronous modulation of PKA by dopamine in learning. *Nature* **590**, 451–456 (2021).
15. Ito, M. & Doya, K. Distinct neural representation in the dorsolateral, dorsomedial, and ventral parts of the striatum during fixed- and free-choice tasks. *J. Neurosci.* **35**, 3499–3514 (2015).
16. Shin, E. J. et al. Robust and distributed neural representation of action values. *eLife* **10**, e53045 (2021).
17. Hattori, R., Danskin, B., Babic, Z., Mlynaryk, N. & Komiyama, T. Area-specificity and plasticity of history-dependent value coding during learning. *Cell* **177**, 1858–1872.e15 (2019).
18. Hirokawa, J., Vaughan, A., Masset, P., Ott, T. & Kepecs, A. Frontal cortex neuron types categorically encode single decision variables. *Nature* **576**, 446–451 (2019).
19. Ottenheimer, D. J., Hjort, M. M., Bowen, A. J., Steinmetz, N. A. & Stuber, G. D. A stable, distributed code for cue value in mouse cortex during reward learning. *eLife* **12**, RP84604 (2023).
20. Watabe-Uchida, M. & Uchida, N. Multiple dopamine systems: weal and woe of dopamine. *Cold Spring Harb. Symp. Quant. Biol.* **83**, 83–95 (2018).
21. de Jong, J. W. et al. A neural circuit mechanism for encoding aversive stimuli in the mesolimbic dopamine system. *Neuron* **101**, 133–151.e7 (2019).
22. Tsutsui-Kimura, I. et al. Distinct temporal difference error signals in dopamine axons in three regions of the striatum in a decision-making task. *eLife* **9**, e62390 (2020).
23. Engelhard, B. et al. Specialized coding of sensory, motor and cognitive variables in VTA dopamine neurons. *Nature* **570**, 509–513 (2019).
24. Akiti, K. et al. Striatal dopamine explains novelty-induced behavioral dynamics and individual variability in threat prediction. *Neuron* **110**, 3789–3804.e9 (2022).

25. Lee, R. S., Sagiv, Y., Engelhard, B., Witten, I. B. & Daw, N. D. A feature-specific prediction error model explains dopaminergic heterogeneity. *Nat. Neurosci.* **27**, 1574–1586 (2024).
26. Jeong, H. et al. Mesolimbic dopamine release conveys causal associations. *Science* **378**, eabq6740 (2022).
27. Coddington, L. T., Lindo, S. E. & Dudman, J. T. Mesolimbic dopamine adapts the rate of learning from action. *Nature* **614**, 294–302 (2023).
28. Costa, V. D., Dal Monte, O., Lucas, D. R., Murray, E. A. & Averbeck, B. B. Amygdala and ventral striatum make distinct contributions to reinforcement learning. *Neuron* **92**, 505–517 (2016).
29. St Onge, J. R. & Floresco, S. B. Dopaminergic modulation of risk-based decision making. *Neuropsychopharmacology* **34**, 681–697 (2009).
30. Zalocusky, K. A. et al. Nucleus accumbens D2R cells signal prior outcomes and control risky decision-making. *Nature* **531**, 642–646 (2016).
31. Ma, W. J., Beck, J. M., Latham, P. E. & Pouget, A. Bayesian inference with probabilistic population codes. *Nat. Neurosci.* **9**, 1432–1438 (2006).
32. Walker, E. Y. et al. Studying the neural representations of uncertainty. *Nat. Neurosci.* **26**, 1857–1867 (2023).
33. Bellemare, M. G., Dabney, W. & Munos, R. A distributional perspective on reinforcement learning. In *Proc. 34th International Conference on Machine Learning* (eds Precup, D. & Teh, Y. W.) 449–458 (PMLR, 2017).
34. Wurman, P. R. et al. Outracing champion Gran Turismo drivers with deep reinforcement learning. *Nature* **602**, 223–228 (2022).
35. Rothenhoefer, K. M., Hong, T., Alikaya, A. & Stauffer, W. R. Rare rewards amplify dopamine responses. *Nat. Neurosci.* **24**, 465–469 (2021).
36. Avvisati, R. et al. Distributional coding of associative learning in discrete populations of midbrain dopamine neurons. *Cell Rep.* **43**, 114080 (2024).
37. Sousa, M., Bujalski, P., Cruz, B. F., Louie, K. & Paton, J. J. Dopamine neurons encode a multidimensional probabilistic map of future reward. Preprint at *bioRxiv* https://doi.org/10.1101/2023.11.12.566727 (2023).
38. Muller, T. H. et al. Distributional reinforcement learning in prefrontal cortex. *Nat. Neurosci.* **27**, 403–408 (2024).
39. Rowland, M. et al. Statistics and samples in distributional reinforcement learning. In *Proc. 36th International Conference on Machine Learning* (eds Chaudhuri, K. & Salakhutdinov, R.) 5528–5536 (PMLR, 2019).
40. Tano, P., Dayan, P. & Pouget, A. A local temporal difference code for distributional reinforcement learning. In *Proc. Advances in Neural Information Processing Systems 33* (eds Larochelle, H. et al.) 13662–13673 (NeurIPS, 2020).
41. Louie, K. Asymmetric and adaptive reward coding via normalized reinforcement learning. *PLoS Comput. Biol.* **18**, e1010350 (2022).
42. Schütt, H. H., Kim, D. & Ma, W. J. Reward prediction error neurons implement an efficient code for reward. *Nat. Neurosci.* **27**, 1333–1339 (2024).
43. O'Neill, M. & Schultz, W. Coding of reward risk by orbitofrontal neurons is mostly distinct from coding of reward value. *Neuron* **68**, 789–800 (2010).
44. Monosov, I. E. & Hikosaka, O. Selective and graded coding of reward uncertainty by neurons in the primate anterodorsal septal region. *Nat. Neurosci.* **16**, 756–762 (2013).
45. White, J. K. & Monosov, I. E. Neurons in the primate dorsal striatum signal the uncertainty of object–reward associations. *Nat. Commun.* **7**, 12735 (2016).
46. Yanike, M. & Ferrera, V. P. Representation of outcome risk and action in the anterior caudate nucleus. *J. Neurosci.* **34**, 3279–3290 (2014).
47. Yamada, K. & Toda, K. Pupillary dynamics of mice performing a Pavlovian delay conditioning task reflect reward-predictive signals. *Front. Syst. Neurosci.* **16**, 1045764 (2022).
48. Tian, J. et al. Distributed and mixed information in monosynaptic inputs to dopamine neurons. *Neuron* **91**, 1374–1389 (2016).
49. Stringer, C. et al. Spontaneous behaviors drive multidimensional, brainwide activity. *Science* **364**, 255 (2019).
50. Musall, S., Kaufman, M. T., Juavinett, A. L., Gluf, S. & Churchland, A. K. Single-trial neural dynamics are dominated by richly varied movements. *Nat. Neurosci.* **22**, 1677–1686 (2019).
51. Hoyer, P. & Hyvärinen, A. Interpreting neural response variability as Monte Carlo sampling of the posterior. In *Proc. Advances in Neural Information Processing Systems 15* (eds Becker, S. et al.) 293–300 (MIT Press, 2002).
52. Orbán, G., Berkes, P., Fiser, J. & Lengyel, M. Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron* **92**, 530–543 (2016).
53. Bernardi, S. et al. The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell* **183**, 954–967.e21 (2020).
54. Lowet, A. S., Zheng, Q., Matias, S., Drugowitsch, J. & Uchida, N. Distributional reinforcement learning in the brain. *Trends Neurosci.* **43**, 980–997 (2020).
55. Gerfen, C. R. & Surmeier, D. J. Modulation of striatal projection systems by dopamine. *Annu. Rev. Neurosci.* **34**, 441–466 (2011).
56. Faust, T. W., Mohebi, A. & Berke, J. D. Reward expectation selectively boosts the firing of accumbens D1$^+$ neurons during motivated approach. Preprint at *bioRxiv* https://doi.org/10.1101/2023.09.02.556060 (2023).
57. Martiros, N., Kapoor, V., Kim, S. E. & Murthy, V. N. Distinct representation of cue-outcome association by D1 and D2 neurons in the ventral striatum's olfactory tubercle. *eLife* **11**, e75463 (2022).
58. Nishioka, T. et al. Error-related signaling in nucleus accumbens D2 receptor-expressing neurons guides inhibition-based choice behavior in mice. *Nat. Commun.* **14**, 2284 (2023).
59. Kupchik, Y. M. et al. Coding the direct/indirect pathways by D1 and D2 receptors is not valid for accumbens projections. *Nat. Neurosci.* **18**, 1230–1232 (2015).
60. Such, F. P. et al. An Atari model zoo for analyzing, visualizing, and comparing deep reinforcement learning agents. In *Proc. 28th International Joint Conference on Artificial Intelligence* (ed. Kraus, S.) 3260–3267 (IJCAI, 2019).
61. Collins, A. G. E. & Frank, M. J. Opponent actor learning (OpAL): modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. *Psychol. Rev.* **121**, 337–366 (2014).
62. Gjorgjieva, J., Sompolinsky, H. & Meister, M. Benefits of pathway splitting in sensory coding. *J. Neurosci.* **34**, 12127–12144 (2014).
63. Ichinose, T. & Habib, S. ON and OFF signaling pathways in the retina and the visual system. *Front. Ophthalmol.* **2**, 989002 (2022).
64. Poulin, J.-F., Gaertner, Z., Moreno-Ramos, O. A. & Awatramani, R. Classification of midbrain dopamine neurons using single-cell gene expression profiling approaches. *Trends Neurosci.* **43**, 155–169 (2020).
65. Wenliang, L. K. et al. Distributional Bellman operators over mean embeddings. In *Proc. 41st International Conference on Machine Learning* (eds Salakhutdinov, R. et al.) 52839–52868 (PMLR, 2024).
66. Mikhael, J. G. & Bogacz, R. Learning reward uncertainty in the basal ganglia. *PLoS Comput. Biol.* **12**, e1005062 (2016).
67. Cui, G. et al. Concurrent activation of striatal direct and indirect pathways during action initiation. *Nature* **494**, 238–242 (2013).
68. Markowitz, J. E. et al. The striatum organizes 3D behavior via moment-to-moment action selection. *Cell* **174**, 44–58 (2018).
69. Tan, B. et al. Dynamic processing of hunger and thirst by common mesolimbic neural ensembles. *Proc. Natl Acad. Sci. USA* **119**, e2211688119 (2022).
70. Bar-Gad, I., Morris, G. & Bergman, H. Information processing, dimensionality reduction and reinforcement learning in the basal ganglia. *Prog. Neurobiol.* **71**, 439–473 (2003).
71. Barth-Maron, G. et al. Distributed distributional deterministic policy gradients. In *Proc. 6th International Conference on Learning Representations* 4855–4870 (ICLR, 2018).
72. Brown, V. M. et al. Reinforcement learning disruptions in individuals with depression and sensitivity to symptom change following cognitive behavioral therapy. *JAMA Psychiatry* **78**, 1113–1122 (2021).
73. Gueguen, M. C. M., Schweitzer, E. M. & Konova, A. B. Computational theory-driven studies of reinforcement learning and decision-making in addiction: what have we learned? *Curr. Opin. Behav. Sci.* **38**, 40–48 (2021).
74. Paxinos, G. & Franklin, K. B. J. *Paxinos and Franklin's the Mouse Brain in Stereotaxic Coordinates* (Academic Press, 2019).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Experimental procedures

Mice. A total of 56 adult C57BL/6J (Jackson Laboratory) male and female mice were used in these experiments. Twelve wild-type animals (6 males and 6 females) were used for Neuropixels recordings in the original task, of which 5 (2 males and 3 females) were also included in unilateral 6-OHDA experiments. The Bernoulli, Diverse Distributions and Fourth Moments tasks made use of three (one male and two female), three (one male and two female) and five (2 male and 3 female) animals, respectively. For two-photon imaging, four *Drd1-cre* (B6.FVB(Cg)-Tg(Drd1-cre)EY262Gsat/Mmucd, RRID: MMRRC_030989-UCD; three male and one female) and four *Adora2a-cre* (B6.FVB(Cg)-Tg(Adora2a-cre)KG139Gsat/Mmucd, RRID: MMRRC_036158-UCD; one male and three female) mice were used[75–77]. For optogenetic excitation, we used five *Drd1-cre* (two male and three female) and seven *Adora2a-cre* (three male and four female) animals. For optogenetic inhibition, we crossed these lines with a Cre-dependent *GtACR1* reporter mouse[78,79] (R26-CAG-LNL-GtACR1-ts-FRed-Kv2.1, RRID: IMSR_JAX:033089). Five *Drd1-cre;GtACR1* (two male and three female) and eight *Adora2a-cre;GtACR1* (four male and four female) mice were used. All transgenic mice used for experiments were backcrossed with C57BL/6J and heterozygous for the relevant allele (or alleles). Sample size was chosen based on similar experiments performed previously in the laboratory. No randomization or blinding was performed, other than randomization of odours to distributions.

Animals were housed on a 12-h dark–12-h light cycle and performed the task at the same time each day (±1 h), during the dark period. Ambient temperature was kept at 75 ± 5 °F, and humidity was kept below 50%. Animals were group-housed (2–5 animals per cage) until surgery, then individually housed throughout training and testing. All procedures 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.

Surgeries. All surgeries were performed under aseptic conditions. Mice (over 8 weeks of age) were anaesthetized with isoflurane (3.5% induction, followed by 1–2% maintenance at $1\,\mathrm{l\,min^{-1}}$), and local anaesthetic (lidocaine, 2%) was administered subcutaneously at the incision site. Analgesia (buprenorphine for pre-operative treatment, $0.1\,\mathrm{mg\,kg^{-1}}$, intraperitoneal (i.p.); ketoprofen for post-operative treatment, $5\,\mathrm{mg\,kg^{-1}}$ i.p.) was administered for 2 days after surgery. After levelling, cleaning and drying the skull, we affixed a custom-made titanium head plate to the skull with adhesive cement[80] (C&B Metabond, Parkell).

For all injections, the solution (6-OHDA or virus) was backfilled into a pulled glass pipette (5-000-1001-X, Drummond), followed by mineral oil and a plunger. A small craniotomy (less than 1 mm in diameter) was made using a dental drill, and then the pipette assembly was mounted on the stereotaxic holder, lowered to the desired coordinate and injected slowly (approximately $100\,\mathrm{nl\,min^{-1}}$) to minimize damage to the surrounding tissue (MO-10, Narishige). After each injection, we waited at least 10 min to allow the solution to diffuse away from the pipette tip before slowly going up to the next coordinate or retracting the pipette from the brain. Target coordinates for the lAcbSh were the same across experiments: anteroposterior 1.1 mm from bregma, mediolateral 1.7 mm and dorsoventral 4.2 mm from the pial surface.

6-OHDA procedure. To unilaterally ablate dopamine neurons projecting to the lateral ventral striatum, we followed an existing protocol[24,81]. The following solution was injected (i.p.) into animals at $10\,\mathrm{mg\,kg^{-1}}$ immediately before surgery: 14.25 mg desipramine (D3900-1G, Sigma-Aldrich); 3.1 mg pargyline (P8013-500MG, Sigma-Aldrich); and 5 ml distilled water.

Most animals (weighing roughly 25 g) received approximately 250 µl of this solution, which was given to prevent dopamine uptake in noradrenaline neurons and to increase the selectivity of uptake by dopamine neurons. We additionally prepared a solution of $10\,\mathrm{mg\,ml^{-1}}$ 6-OHDA (H116-5MG, Sigma-Aldrich) and 0.2% ascorbic acid in saline (0.9% NaCL; PHR1008-2G, Sigma-Aldrich). The ascorbic acid in this solution helps to prevent 6-OHDA from breaking down. The control hemisphere was either injected with vehicle ascorbic acid solution or uninjected; we observed no differences between these groups and so combined them. To further prevent 6-OHDA from breaking down, we kept the solution on ice, wrapped in aluminium foil and used it within 3 h of preparation. If the solution turned brown during this time (indicating that 6-OHDA had broken down), it was discarded and a fresh solution was made. In total, 225 nl of 6-OHDA (or vehicle) was injected unilaterally into the lAcbSh.

Surgeries occurred at least 1 week before the start of behavioural training. We lesioned nine mice and included control hemisphere data for all of them in the main dataset. However, four of these mice either died before we could record from the lesioned hemisphere or were not correctly targeted for the lesion and/or recording, and so were excluded from the lesion dataset.

Viruses. To express constructs specifically in D1 or D2 MSNs, we injected viruses into *Drd1-cre* and *Adora2a-cre* mice. For imaging experiments, we unilaterally injected 450 nl AAV9-hSyn-flex-GCaMP7s (at least $1\times10^{13}\,\mathrm{vg\,ml^{-1}}$, Addgene)[82] into the lAcbSh. For optogenetic activation experiments, we bilaterally injected AAV9-hSyn-flex-CoChR-GFP ($5.1\times10^{12}\,\mathrm{vg\,ml^{-1}}$, UNC Vector Core, NC)[83] at anteroposterior 1.1 mm and mediolateral ±1.7 mm in 300-nl increments at four separate depths below the pial surface: 4.2, 3.4, 2.6 and 1.8 mm.

GRIN lens and fibre implantations. Before the GRIN lens surgery, we injected animals i.p. with 50 µl dexamethasone ($2\,\mathrm{mg\,ml^{-1}}$; Vedco) to reduce inflammation. Before virus injection, a needle was mounted on the stereotaxic holder, connected to light suction and lowered to 3.4 mm below the pial surface to gently aspirate away the overlying brain tissue. After virus injection, a singlet GRIN lens (0.5 NA, 0.6 mm diameter, 7.3 mm length, 0–200 µm working distance, 3/2 pitch; 1050-004597, Inscopix) was mounted onto a stereotaxic cannula holder (Doric) and then slowly lowered over at least 30 min to its target depth, 200 µm above the injection site and 3.8 mm below the pial surface. Metabond was used to secure the GRIN lens on all sides and allowed to dry completely before removing the cannula holder and covering everything with another layer of Metabond mixed with charcoal powder to block out light. Finally, a plastic cap was attached with Kwik-Cast (World Precision Instruments) to protect the lens from damage.

For optogenetic manipulation, we bilaterally implanted tapered fibres[84] (0.66 NA, 200 µm diameter, 3 mm emitting length, 5 mm implant length; Optogenix) in the lAcbSh after virus injection, at a depth of 4 mm. Each fibre was secured using Metabond and then protected with a fitted cap.

Behavioural setup and tasks. Behavioural events were controlled (and licking was monitored) using custom-written software in MATLAB (Mathworks) and the Bpod library (Sanworks) interfacing with the Bpod state machine (1024 and 1027, Sanworks), valve module (1015, Sanworks) and port interface board (1020, Sanworks)/water valve (LHDA1233115H, Lee Company) assembly. Odours were delivered using a custom olfactometer[85], which directed air through one of eight solenoid valves (LHDA1221111H, Lee Company) mounted on a manifold (LFMX0510528B, Lee Company). Each odour was dissolved in mineral oil at 10% dilution, and 30 µl of diluted odour solution was applied to a syringe filter (2.7 µm pore, 13 mm diameter; 6823-1327, Whatman). Wall air was passed through a hydrocarbon filter (HT200-4, Agilent Technologies) and split into a $100\,\mathrm{ml\,min^{-1}}$ odour stream and $900\,\mathrm{ml\,min^{-1}}$ carrier stream using analogue flowmeters (MFLX32460-40 and MFLX32460-42, Cole-Parmer), which were recombined at the odour manifold before being delivered to the nose of the mouse. Licking was monitored using an infrared emitter–photodiode pair positioned just in front of the plastic lick spout, positioned at the mouth of the mouse. Following previous work, we assumed that the level of Pavlovian conditioned responding provides a readout of the motivational value estimates of the mouse[80,86].

Animals used for Neuropixels recording and two-photon imaging were conditioned with five to six (depending on the task) different neutral odours, chosen at random from these seven: isoamyl acetate, *p*-cymene, ethyl butyrate, (*S*)-(+)-carvone, (±)-citronellal, α-ionone and L-fenchone. Mice in the optogenetic manipulation experiment used only the first three odours. In all experiments, the mapping between physical odour and conceptual trial type was randomized across mice. Each trial began with a 1-s odour presentation, followed by 2-s trace period and then reward delivery. There was a minimum of 4.6 s before the next trial (4.1 s for optogenetic manipulation mice), plus a variable inter-trial interval (ITI) drawn from a truncated exponential distribution with a mean of 2 s, minimum of 0.1 s and maximum of 10 s. For two-photon imaging experiments, this was extended to a mean of 10.5 s, minimum of 6.5 s and maximum of 18.5 s, to account for the slower kinetics of the calcium indicator relative to electrophysiology.

The main recording task consisted of three different reward distributions: Nothing, Fixed and Variable (Fig. 1b). Each distribution was then paired with two unique odours, for a total of six odours. The distributions were as follows:

- Nothing: 100% chance of 0 μl water
- Fixed: 100% chance of 4 μl water
- Variable: 50% chance of 2 μl water or 50% chance of 6 μl water.

The task used for optogenetic manipulation was simplified in two ways. First, we used only one odour per distribution, for a total of three odours. Second, we modified the Variable distribution to be 50/50% between 0 μl and 8 μl, because our model predicted that increasing the variance would lead to a greater behavioural difference between Fixed and Variable odours.

Behavioural training. Water restriction began no earlier than 5 days after recovery from surgery. The condition of the mice was monitored daily to ensure that mice did not dip below 85% of their free-drinking body weight, including supplementing with additional water after the task to bring their total daily intake to approximately 1.2 ml. Over the course of three successive habituation days, mice were (1) handled gently for several minutes in their home cage, (2) permitted to freely roam around the platform in the behaviour rig to collect water, and then (3) head-fixed while receiving frequent (inter-reward interval of 4–5 s) 6 μl water rewards.

The optogenetic manipulation task proceeded in only one phase, with up to 110 Nothing, 110 Fixed and 114 Variable trials, randomly interleaved. By contrast, training for the recording task took place in three phases, each with a maximum of 300 trials.

- Phase 1: both Nothing odours and both Fixed odours with equal probabilities
- Phase 2: all six odours, but with the Variable odours 5.5× more frequent than the others
- Phase 3: all six odours at a final ratio of 4:4:7 (Nothing:Fixed:Variable), to increase the statistical power for analysing responses to different reward sizes.

We trained animals in three additional tasks to test the generality of REDRL (Extended Data Fig. 8). Each of these tasks also had its own shaping procedure.

Bernoulli task:

- Phase 1: 0% and 100% odours with equal probabilities
- Phase 2: 0%, 50% and 100% odours at a ratio of 3:10:3
- Phase 3: 0%, 20%, 50% and 80% odours at a ratio of 2:15:8:15
- Phase 4: all five odours at a final ratio of 1:2:2:2:1.

Diverse Distributions task:

- Phase 1: conditioned stimulus 1, 2 and 6 with equal probabilities
- Phase 2: conditioned stimulus 1, 2, 3 and 6 at a ratio of 4:3:50:3
- Phase 3: conditioned stimulus 1, 3 and 4 at a ratio of 1:5:9
- Phase 4: conditioned stimulus 1, 2, 3, 4 and 6 at a ratio of 4:9:8:30:9
- Phase 5: conditioned stimulus 1, 3, 4 and 5 at a ratio of 2:2:5:21
- Phase 6: all six odours at a ratio of 9:9:20:50:203:9
- Phase 7: all six odours at a final ratio of 23:25:80:70:77:25.

Fourth Moments task:

- Phase 1: both Nothing and both Uniform odours at a ratio of 74:77
- Phase 2: all six odours, but with the Bimodal odours 5.5× more frequent than the others
- Phase 3: all six odours at a final ratio of 39:56:56.

On recording days, animals experienced a maximum of 20 additional unexpected reward trials, in which 4 μl of water was delivered without being preceded by an odour cue. All trials were randomly interleaved in all phases.

For all tasks, animals completed at least 150 trials per day, and almost always more than 250. The experiment might have been terminated early by the experimenter if the mice stopped licking in anticipation (or consumption) of the rewards due to satiety. A behavioural session was considered significant if the lick rate during the last half second before reward delivery was significantly different between rewarded (Fixed and Variable) and unrewarded (Nothing) odours (Mann–Whitney *U*-test, $\alpha = 0.05$) and the effect size was at least 0.75 licks per second. Animals were advanced to the next phase, or to habituation for recording or manipulation, after at least 2 consecutive days with significant behaviour. On recording or manipulation days, only significant behavioural sessions were included for neural or behavioural analysis.

Neuropixels recordings. The day before recording, mice were habituated to the recording setup by covering their heads with a plastic sheet to block their view of the probe and manipulator. We then turned on the lamp, ran the brushed motor controller (KDC101 and Z825B, Thorlabs) up and down for about 30 s, tapped on the skull several times with fine forceps, and left the mouse head-fixed for at least 30 min before beginning the behavioural protocol. If necessary, we repeated this habituation protocol every day until the behaviour of the mouse was significant (see ‘Behavioural training’). After this, we anaesthetized the mouse to make a small craniotomy, which was then covered with Kwik-Cast. The craniotomy was guided by fiducial marks made at the target sites for probe insertion during headplate implantation using a fine-tipped pen. Target coordinates included: anteroposterior 0.9 mm and mediolateral 1.7 mm (lAcbSh); anteroposterior 1.1 mm and mediolateral 1.4 mm (nucleus accumbens core); and anteroposterior 1.4 mm and mediolateral 0.6 mm (medial accumbens shell (mAcbSh)). For the first craniotomy, a ground pin was inserted into the posterior cortex and a custom-made plastic recording chamber was fixed to the top of the headplate, both using 5-min epoxy (Devcon).

The next day, we head-fixed the mouse, covered its head as before, removed the Kwik-Cast and flushed the craniotomy with saline. For the first recording in each craniotomy, we coated the probe in lipophilic dye at 10 mg ml^{-1}. DiI (1,1′-dioctadecyl-3,3,3′,3′-tetramethylindocarbocyanine perchlorate; 100 mg; 42364, Sigma-Aldrich) and DiD (1,1′-dioctadecyl-3,3,3′,3′-tetramethylindodicarbocyanine, 4-chlorobenzenesulfonate; 10 mg; 60014, Biotium) were dissolved in 100% ethanol (V1001, Koptec), and DiO (3,3′-dioctadecyloxacarbocyanine perchlorate; D275, Thermo Fisher) was dissolved in 100% *N*,*N*-dimethylformamide (Sigma-Aldrich, D4254). The coated Neuropixels 1.0 (ref. 87) or four-shank Neuropixels 2.0 (ref. 88) probe was then mounted on the manipulator and connected to the ground pin via a wire soldered onto the reference pad and shorted to ground. In the event the external reference was unstable, we used tip referencing instead. All recordings were performed in SpikeGLX software (https://github.com/billkarsh/SpikeGLX), with a sampling rate of 30 kHz, local field potential gain of 250 and action potential gain of 500, and we analysed only the action potential channel (which was high-pass filtered in hardware with a cut-off frequency of 300 Hz).

We inserted the probe into the brain at 9 µm s^{-1} before slowing to 2 µm s^{-1} when we were 500 µm above the target depth. We stopped insertion when we saw ventral pallidal activity, characterized by large-amplitude, high-frequency spikes, on the first 40 channels or so (or 5 channels for Neuropixels 2.0). This point was usually reached around 5.2 mm below the visually identified pial surface. After reaching the target depth, the probe was allowed to settle for 30 min before starting the experiment and Neuropixels recording. Behavioural and neural recordings were synchronized using a transistor–transistor logic (TTL) pulse sent from the Bpod to the PXIe acquisition module SMA input at the start of every trial. After the experiment, the probe was retracted at 9 µm s^{-1}, and the craniotomy was resealed with Kwik-Cast. Neuropixels data were spike sorted offline with Kilosort 3 (ref. 89) with default parameters, followed by manual curation in Phy (https://github.com/cortex-lab/phy).

Two-photon imaging. Imaging data were acquired using a custom-built two-photon microscope. A resonant scanning mirror and galvanometric mirror (CRS 8 KHz and 6210H, Cambridge Technology) separated by a scan lens-based relay on the scan head (MM201, Thorlabs) allowed fast scanning through a dichroic beamsplitter (757-nm long pass, Semrock) and 20×/0.5 NA air immersion objective lens (Plan Fluor, Nikon). Green and red emission lights were separated by a dichroic beamsplitter (568-nm long pass, Semrock) and bandpass filters (525/50 nm and 641/75 nm, Semrock) and collected by GaAsP photomultiplier tubes (H7422PA-40, Hamamatsu) coupled to transimpedance amplifiers (TIA60, Thorlabs). A diode-pumped, mode-locked Ti:sapphire laser (Spectra-Physics) delivered excitation light at 920 nm with an average power of approximately 60 mW at the top face of the GRIN lens[90], modulated by a Pockels cell (350-80, Conoptics). The microscope was controlled by ScanImage (v4; Vidrio Technologies). The behavioural platform was mounted on an XYZ translation stage (LTS150 and MLJ050, Thorlabs) to position the mouse under the objective, and the top face of the GRIN lens was first located using a 470-nm LED (M470L2, Thorlabs).

Owing to the limited axial resolution of the implanted GRIN lens, we acquired only a single imaging plane at 15.2 Hz unidirectionally with 1.4× digital zoom and a resolution of 512 × 512 pixels (approximately 0.7 µm per pixel isotropic). Imaging was either continuous or triggered 2.6 s before odour or unexpected reward onset, depending on the session. Bleaching of GCaMP7s was negligible over this time. TTL pulses were sent from the microscope to Bpod to synchronize imaging and behavioural data. Imaging typically began approximately 4 weeks after GRIN implantation, to allow sufficient time for the virus to express and for inflammation to clear.

Two-photon preprocessing. We used the Suite2p toolbox[91] (v0.10.3) to register frames, detect cells, extract Ca^{2+} signals and deconvolve these traces. We used parameter values of tau = 2.0 (to approximately match the decay constant of GCaMP7s[82]), sparse_mode=False, diameter=20, high_pass=75, neucoeff=0.58; fs was set to the measured frame rate for that session (approximately 15.2 Hz), and all other parameters were set to their defaults. In brief, non-rigid motion correction was used in blocks of 128 × 128 pixels to register all frames to a common reference image using phase correlation. Cell detection consisted of finding and smoothing spatial principal components and then extending region of interests (ROIs) spatially around the peaks in these principal components. Next, Ca^{2+} traces were extracted from each ROI after discarding any pixels belonging to multiple ROIs. Finally, neuropil contamination and deconvolved spikes were estimated in a single step from Ca^{2+} fluorescence in each ROI using the OASIS algorithm[92] with a non-negativity constraint. This deconvolved activity was used for all subsequent analysis. ROIs were manually curated on the basis of anatomical and functional criteria using the Suite2p GUI to exclude neuropil and ROIs with few or ill-formed transients.

Face and body imaging. In addition to the lick port, we monitored behaviour using two cameras at 30 Hz, one pointed at the face (FL3-U3-13Y3M, PointGrey) and one pointed at the body (CM3-U3-13S2C, PointGrey) under both visible and infrared LED illumination. Cameras were synchronized from Bpod once per trial using general-purpose input/output (GPIO) inputs, and data were written to disk via Bonsai[93]. Behavioural features were extracted using custom code alongside Facemap[49] (v0.2.0). Face motion energy was computed as the absolute value of the difference between consecutive frames and summed across all pixels to yield the 'whisking' signal. In addition, we performed singular value decomposition on the motion energy video (in chunks, following ref. 49) and projected the movie onto the top 50 components to obtain their activity patterns over time. Pupil area was estimated simply as the mean (inverse) pixel value within a mask, after interpolating over blink events. Running was computed using the phase correlation of the cropped body video, to take into account limb and tail movements.

Optogenetic manipulation. Laser light (473 nm; LRS-0473-GFM-00100-03, Laserglow Technologies) was delivered to the implanted tapered fibres using a custom-built rig (modelled after refs. 94,95) coupled to a high-performance patch cord (0.66 NA, OPT/PC-FC-LCF-200/230-HP-2.2L KIT, Plexon). In brief, light was split into two identical paths using a 50/50 beamsplitter cube (CCM1-BS013, Thorlabs). Each path was then focused onto a galvanometric mirror (Novanta 6210K) and re-collimated using an achromatic doublet (AC508-100-A-ML, Thorlabs), before being focused onto the back of the patch cord using an aspheric condenser lens (ACL50832U, Thorlabs). This setup allowed us to modulate the angle at which light entered the patch cord, and thus the distance at which it exited the tapered fibre. We delivered light at two different angles (three in some experiments), but here we analysed only ventral manipulation trials, in which the incident angle of light was approximately 0°, light exited near the tip of the fibre, and coupling between the patch cord and fibre was approximately 50%[94].

The laser output (and the angle of the galvanometric mirrors) was controlled by Bpod via PulsePal[96] (v2; 1102, Sanworks). Stimulation was delivered bilaterally during the 2-s-long trace period, immediately before the reward. For CoChR excitation experiments, we used 10-ms pulses at 20 Hz with an output power at the tapered fibre of 100 µW. For GtACR1 inhibition, we used a constant, 1-mW pulse for the full 2 s. In both cases, stimulation was delivered on 45.5% of trials, uniformly at random across manipulation locations and trial types.

Histology and immunohistochemistry. Mice were deeply anaesthetized with ketamine–dexmedetomidine (80 and 1.1 mg kg^{-1}, respectively) and then transcardially perfused using 4% paraformaldehyde. The brains were sliced at 100 µm into coronal sections using a vibratome (Leica) and stored in PBS. If performing immunostaining, slice thickness was 75 µm. These slices were then permeabilized with 0.5% Triton X-100, blocked with 10% FBS and stained with rabbit anti-TH antibody (AB152, EMD Millipore; RRID: AB_390204) at 1:750 dilution at 4 °C for 24 h to reveal dopamine axons in the striatum. Next, slices were stained with fluorescent secondary antibodies (Alexa Fluor 488 goat anti-rabbit secondary antibody; A-11008, Invitrogen; RRID: AB_143165) and DAPI at 1:500 dilution at 4 °C for 24 h. Slices were then mounted on glass slides (Vectashield antifade mounting medium, H-1000, or with DAPI for non-stained slices, H-1200, Vector Laboratories) and imaged using a Zeiss Axio Scan Z1 slide scanner fluorescence microscope. We visually verified the placement of all GRIN lenses and fibres to be within the lAcbSh.

Data analysis

Atlas registration. For electrophysiology experiments, we registered slices to the Allen Mouse Brain Atlas with SHARP-Track[97] and used it to trace dyed probe trajectories in the anteroposterior and mediolateral directions, as well as visualize the registered trajectories as a coronal stack. We also used this registration to define the unique dorsoventral extent of the lateral ventral striatal 6-OHDA lesion of each mouse, and we considered only neurons that fell within this range to have been lesioned. To more accurately ascertain the depth of recordings, we used the Ephys Atlas GUI by the International Brain Lab (https://github.com/int-brain-lab/iblapps/tree/master/atlaselectrophysiology), focusing on the boundary between the ventral pallidum and nucleus accumbens due to the abrupt change in electrophysiological characteristics at this interface. When necessary, we also adopted their convention that in Allen Common Coordinate Framework[98] coordinates, bregma = 5400 anteroposterior, 332 dorsoventral and 5739 mediolateral. For plotting probe trajectories in 3D, we used the Brainrender library[99].

For more fine-grained analysis of subregions, we used the Kim Lab atlas[100] accessed through the BrainGlobe Atlas API[101]. This atlas applies the Franklin and Paxinos[74] labels to the Allen Common Coordinate Framework[98], with additional striatal subregions defined by Hintiryan et al.[102]. For some subregions, the parcellation was finer than we needed, so we pooled subregions (as defined by refs. 100,102) as follows:

- Olfactory tubercle: Tu1; Tu2; Tu3
- Ventral pallidum: VP
- Medial nucleus accumbens shell: AcbSh
- Lateral nucleus accumbens shell: lAcbSh; CB; IPACL
- Nucleus accumbens core: AcbC
- Ventromedial striatum: CPr, imv; CPi, vm, vm; CPi, vm, v; CPi, vm, cvm
- Ventrolateral striatum: CPr, l, vm; CPi, vl, imv; CPi, vl, v; CPi, vl, vt; CPi, vl, cvl
- Dorsomedial striatum: CPr, m; CPr, imd; CPi, dm, dl; CPi, dm, im; CPi, dm, cd; CPi, dm, dt
- Dorsolateral striatum: CPr, l, ls; CPi, dl, d; CPi, dl, imd.

Unit inclusion criteria. To be included for analysis, units from Neuropixels recordings had to have a minimum firing rate of 0.1 Hz and to have been stable, defined as a coefficient of variation of firing rate (computed in 10 equally sized, contiguous, disjoint blocks during the session) less than 1. 13,997 single units survived these inclusion criteria in the main dataset. In the lesion dataset, we additionally filtered neurons by their dorsoventral position: only those that fell within the dorsoventral range of the lesion were included in the matched control dataset for that mouse. Of the 9,081 neurons that survived the electrophysiological criteria, 4,879 were in the correct anatomical location, of which 2,283 came from the control and 2,596 came from the lesioned hemisphere.

Putative cell-type identification. We assigned units to putative cell types using previously established criteria[103]. In brief, to be considered MSNs, units were required to have broad waveforms (Kilosort template trough-to-peak waveform duration of more than 400 µs) and post-spike suppression of 40 ms or slower. For the latter, we used the autocorrelation function with a bin width of 1 ms. Post-spike suppression was quantified as the duration for which the autocorrelation function was less than its average during lags between 600 ms and 900 ms.

Statistical software. All statistical analysis, except where explicitly stated, was performed in Python using the NumPy (v1.22.3), SciPy (v1.7.3), pandas (v1.1.4), scikit-learn (v1.0.2), statsmodels (v0.14.0), Matplotlib (v3.5.1) and seaborn (v0.12.2) packages[104–110]. All reported *P* values are two-tailed. We did not perform tests for normality or correct for multiple comparisons. If not otherwise specified, statistical tests used linear mixed effects models (LMEs) with a random intercept for each mouse, and, if applicable, a random slope for each mouse as a function of grouping (for example, across versus within distribution), implemented in statsmodels. Full model specifications for every LME can be found in Supplementary Table 1.

Units of analysis. For the behaviour, control and manipulation datasets (Figs. 1, 2 and 5), each observation was an individual session—that is, we used simultaneously recorded neurons and behaviour and computed effects (PCA, RDA, parallelism score and classification) on a session-by-session basis. This was also the case for parallelism score in the lesioned dataset (Fig. 3g), as this analysis already requires subsampling to 100 neurons (see below). However, given the limited spatial extent of our lesion and our lower number of simultaneously recorded neurons, for the remainder of the lesion dataset (Fig. 3) we used pseudo-populations. More specifically, we created pseudo-populations by splitting the dataset into disjoint sets of trials[111], which were stitched across sessions, but not across animals. Within each session, we used simultaneously recorded trials across neurons to preserve noise correlations where possible. For these LMEs, pseudo-populations provided the observations, and 'mouse' was again the grouping variable for random effects. The same procedure was used for all subregion-specific analyses (Extended Data Figs. 2e,l, 4a–d and 8g–j) and artificial neural network (ANN)-based decoding (Extended Data Fig. 4g–j), again due to the lower number of simultaneously recorded neurons available for these analyses.

For the imaging dataset (Fig. 4) and ANN-based transfer (Extended Data Fig. 4k,l), we did not have enough neurons in all animals to assess distributional coding. We therefore pooled neurons not only across sessions but also across animals within genotype. Pseudo-populations were otherwise constructed exactly as in the lesion case. To be consistent with the parametric nature of LMEs while recognizing that observations were no longer specific to individual mice, we used one-sample Student's *t*-tests to assess statistical significance relative to chance levels and LMEs (with just one observation per group) to assess differences between groupings.

Neuron-level analyses (for example, Extended Data Figs. 2d,g,n,o,q,r, 5d,g,h and 7n) treated neurons as individual observations, with random effects of 'session' nested within mouse.

Plotting conventions. In the figures, the asterisks over lines connecting different groupings indicate significant differences between groups, whereas asterisks without corresponding lines indicate that the group is significantly different from chance. Chance levels are indicated by grey dashed lines. The shaded regions from 0 to 1 s represent the interval of odour delivery, and the vertical lines at 3 s indicate reward timing. Except where otherwise noted, vertical bars or shading around data points indicate the mean ± 95% confidence interval (c.i.) of the relevant units of analysis, be they mice or pseudo-populations.

Time periods for analysis. In general, we analysed behavioural and neural data during the late trace period, 1–0 s before reward delivery. However, for licking comparisons before and after odour onset, we also used the baseline period (1–0 s before odour onset); for odour decoding, we used the odour period (0–1 s after odour onset); and for reward or RPE coding, we used the outcome period (0–1 s after reward delivery). Analysis of variability across trials (Extended Data Fig. 6) examined changes across all of these time periods, as well as the early trace period, 2–1 s before reward delivery. Neural and behavioural data were averaged within these 1-s periods before analysis, with the exception of plots of classification or regression time courses, in which averages within non-overlapping 250-ms bins were used for increased granularity.

Visualization of neural time courses. For smoothed plots of neural time courses (Figs. 1f,g, 2a and 5d,g and Extended Data Figs. 2b and 10a,b), we smoothed neural activity (spike trains or deconvolved activity traces) with a Gaussian kernel (s.d. of 100 ms) before plotting or reducing dimensionality. *Z*-scored firing rates were computed using the mean and standard deviation of this smoothed trace. PCA time courses (Fig. 1g) were extracted by computing the average normalized, smoothed firing rate for each trial type and concatenating these into a 2D matrix of shape $N \times (T \times 6)$, where N is the number of neurons, T is the number of time points per trial, and 6 corresponds to the six possible odours. PCA was then performed and the time courses were reconstructed separately for each of the six odours. All other analyses used unsmoothed data to remain uncontaminated by later time points.

PCA and RDA. For two-dimensional principal component plots, normalized activity during the late trace period was averaged across trials within a given type to produce a matrix of shape $N \times 6$. We then applied PCA to reduce this matrix to shape 2×6, having retained only the top 2 principal components. Results were qualitatively identical when using all neurons or only putative MSNs for the main dataset (Fig. 2). We report Euclidean distances between projected trial types, measured separately along each principal component. RDA was similar, except that we computed cosine distances in the native (pseudo-) population normalized firing rate space, rather than a lower-dimensional projection.

For the Bernoulli, Diverse Distributions and Fourth Moments tasks, we computed Euclidean distance matrices between trial types separately for PC1 and PC2. We then computed the Pearson correlation between the (flattened) empirical distance matrix and the distance matrix for each model to get a single estimate of model fit (Extended Data Fig. 8d,e).

We note that PCA and RDA (as well as parallelism score, below) all rely on trial averaging. Therefore, the small amount of trial-by-trial updating that we observed in our GLM cannot account for these signatures of distributional coding.

Parallelism score. Following ref. 53, we computed the normalized mean firing rate in response to each of the Fixed and Variable odours. There are two possible ways to meaningfully pair up these four odours: (1) Fixed 1 versus Variable 1 and Fixed 2 versus Variable 2, or (2) Fixed 1 versus Variable 2 and Fixed 2 versus Variable 1. In both cases, we can compute difference vectors pointing from Variable to Fixed (Fig. 2b) and then take the cosine similarity between them. The parallelism score that we report is simply this cosine similarity, averaged over the two possible divisions. Because this statistic will be affected by the dimensionality of the vectors in question, we subsampled all populations to 100 neurons, averaging over 100 random subsamples for each split and session. Note that in the case of isotropic noise, the vectors that we define are equivalent to those defined by a maximum-margin linear classifier between the two conditions. However, the high parallelism score does not necessarily imply high CCGP, for example, if the test conditions are much closer together than the training conditions, or the noise is high and/or anisotropic.

Classification. For both behavioural and neural binary classification, we used a support vector classifier with a linear kernel, hinge loss function, L2 penalty, balanced accuracy scoring across classes and regularization parameter 5×10^{-3}, implemented in scikit-learn. The linear kernel allows for easy interpretation of the learned weights. Input data (unnormalized spike counts, lick counts or mean Facemap predictors) were transformed using StandardScaler (computed on training data) before being fed to the classifier.

We ran six different classification analyses: CCGP[53], pairwise decoding, congruency, mean, odour and variable reward amount, as described in the main text and figure legends. Across-distribution and within-distribution results were just the average over the relevant dichotomies (for example, the four possible ways to set up CCGP). For all simultaneous decoding analyses except for CCGP, fivefold cross-validation was used, and reported classification accuracy was the average over these five folds. For CCGP, cross-validation was unnecessary because training and test sets were fully disjoint already. Similarly, for pseudo-population-based decoding (Figs. 3 and 4), five training sets and one disjoint test set were used in all cases. For six-way odour classification, we used multinomial logistic regression rather than a support vector classifier, again with a regularization parameter of 5×10^{-3} and balanced accuracy scoring across classes.

Cross-temporal decoding (Extended Data Figs. 2k and 3h–j) settings were identical to the above. For the odour, pairwise and congruency analyses, we ensured that the same trial never appeared in both the training and the testing sets, despite the different time windows used, to avoid leakage due to temporal autocorrelation. For CCGP, train and test trials were always different, so this was not a concern.

Cosine similarity to classification boundary. Both linear classification and regression find a high-dimensional weight vector in neural state space; computing the cosine similarity between these vectors can identify whether two analyses are homing in on the same or different features. For each session, in addition to performing classification as described above, we regressed input data (unnormalized spike counts, lick counts or mean Facemap predictors) during the same time period against per-trial mean or variance (using StandardScaler followed by RidgeCV with default scikit-learn parameters). Note that the regression uses all six trial types, whereas the classification is limited to looking at only two (pairwise or CCGP) or four (congruency or mean) odours at a time. We then took the weights learned by each regression and computed the cosine similarity with the classification weights (separately for each of the five classification cross-validation folds for non-CCGP decoders; each session was summarized as the average of these five measurements). We report the results of an LME testing either the difference from a chance value of 0, indicating orthogonality (CCGP) or the difference between the absolute cosine similarities for across-distribution and within-distribution decoders (pairwise and congruency; Extended Data Fig. 3f,g).

Distribution-coding subpopulation. To identify neurons that contributed significantly to distribution decoding, we extracted the coefficients from the CCGP, pairwise and congruency decoders of each session and averaged them across dichotomies (and across cross-validation folds if necessary). For the pairwise and congruency analyses, we additionally took the difference between across-distribution and within-distribution coefficients. For each quantile level (computed on each set of coefficients individually for each mouse and each decoder), we then calculated the fraction of neurons above this quantile level for all three decoders compared with null decoders in which trial types had been shuffled before being run through the decoder. We chose a cut-off such that only 2.5% of these cells from the null decoders survived; for the actual data, this corresponded to 1,600 significant distribution-coding neurons, or 11.43% of the total. We refer to these neurons as the 'distribution-coding subpopulation' (Extended Data Fig. 4e–l).

Percentage of significant cells. To compute correlations with different variables of interest, we calculated the trial-wise Pearson correlation between unsmoothed activity in each bin and the value of the variable of interest on that trial. We then repeated this procedure, except that for each neuron independently, we shuffled the mappings between odour and distribution. For example, when considering correlations with mean value, a Fixed 1 trial would correspond to a mean of 4 (μl). If upon shuffling, Fixed 1 odours were mapped to Nothing 2, then the corresponding mean in the shuffled dataset would be 0. Percentages of cells significantly correlating with variables of interest (positively, negatively or without restriction) were averaged over the four 250-ms bins corresponding to the late trace period. We then subtracted the shuffled from the unshuffled fraction to account for odour coding.

When plotted (for example, Extended Data Figs. 2o,r and 7o), each point denotes the per-mouse difference in fraction of significant cells (that is, cells with uncorrected $P < 0.05$) for the unshuffled and shuffled data, separately for cells that correlated positively or negatively with mean reward.

For conjunctive coding (Extended Data Fig. 2h), we compared the actual number of cells with significant correlations for both mean and RPE to the null hypothesis of independent coding, for which individual probabilities would be expected to multiply. In the electrophysiology datasets, for which there were sufficient neurons per session, we computed these fractions separately for each session and fit an LME using fractions in each session as the observations (Extended Data Fig. 2).

Changes in neural activity relative to baseline. To assess changes in neural activity relative to the baseline period (Extended Data Fig. 10d), we first grouped all unrewarded (Nothing) and rewarded (Fixed and Variable) trials for each neuron. We then ran a rank-sum test between late trace activity and baseline activity, separately on each neuron and trial-type grouping. Finally, we computed the fraction of cells per mouse that increased or decreased significantly ($\alpha = 0.05$), and then ran paired samples Student's t-tests on the respective fractions for rewarded versus unrewarded trial types for each group of mice.

Comparisons across subregions, hemispheres and genotypes. Whenever subregions, hemispheres or genotypes were directly compared, we randomly subsampled the number of neurons so that population sizes were identical across this comparison. For subregion and hemisphere (lesioned versus control), this matching was done within-animal; therefore, changes in valuation or continual learning cannot explain these differences. When comparing subregions, we excluded a subregion from an animal if it did not contain at least 40 neurons, hence the differing number of dots (animals) per subregion (Extended Data Figs. 2e,l, 4a–d,f and 8g–j). For genotype (D1 versus D2 MSNs), matching was done across-animals for the entire population of D1 or D2 neurons. To allow for higher neuron counts, all of these imaging-based decoding analyses were performed on pseudo-populations.

ANN-based distribution decoding. To determine whether neural populations contained sufficient information to reconstruct the complete reward distribution, rather than simply perform binary classification based on reward variance, we constructed an ANN-based distribution decoder. Pseudo-population activity from the distribution-coding subpopulation $\mathbf{a}$ was first mapped into 16 dimensions by a trainable, unregularized decoding matrix W. The network takes $W\mathbf{a}$ as input and outputs the predicted distribution. It has one input layer, two hidden layers and one output layer. Each of the two hidden layers had 32 neurons and used the non-linear activation function $f(x) = \ln(1 + \exp(x + 1)) - 1$, which is close to the identity function for $x \gg 0$ and to -1 for $x \ll 0$. The output layer had size 4, with each dimension corresponding to a possible reward size (0, 2, 4 or 6 μl). After linear combination, we also applied the nonlinear function $f(x)$ as specified above, followed by the softmax function to turn the output into a normalized probability distribution.

We applied stochastic gradient descent (SGD) to minimize the following loss function based on the 1-Wasserstein distance (D):

$$L(W, \text{network weights}) = \langle D(\text{decoded\_dist}, \text{groundtruth\_dist})\rangle + \lambda\ \|\text{network weights}\|_2^2,$$

where D is defined as $D(P, Q) = \sum_n |P(r_n) - Q(r_n)|$ for discrete cumulative distribution functions (CDFs) P and Q, where the sum is over all used reward magnitudes, and where r_n is the respective reward magnitude. In other words, the 1-Wasserstein distance measures the unsigned area between two CDFs. For plotting, we normalized this metric by dividing by the minimum achievable Wasserstein distance that would result from predicting the same distribution for every trial type across the training and test sets ('Wasserstein distance relative to reference').

For all experiments, λ was set to 0.02 and the learning rate was 0.002. All the trainable weights were randomly initialized with a mean of 0 and standard deviation of 1, and then divided by 15. For each disjoint pseudo-population, we trained 5 randomly and differently initialized candidate ANNs each for 1,200 iterations, and picked the best-performing ANN to further train for 10,000 iterations. The ANN was implemented in Julia (v1.6.7) and trained on a GPU (GeForce RTX 2070, NVIDIA).

In the standard decoding setting, all six trial types were included in the training and testing sets (with different trials in each). For decoding restricted to trial types with the same mean, only Fixed and Variable trial types were used, but split according to the same logic. In both cases, we performed decoding independently from each mouse, and we compared our results to what happened when we randomly shuffled the odour-distribution mappings before training. If merely odour identity (or, in the restricted case, mean) is encoded, then the ordered and shuffled networks should attain similar performance.

Finally, in the transfer analysis, in a similar spirit to CCGP, we trained on only four trial types and then tested on the held-out two trial types. 'Matched' transfers used one Fixed and one Variable odour in the training set, assigned to the proper distribution, and evaluated performance on the corresponding test odour. 'Mismatched' transfers used either two Fixed or two Variable odours in the training set, assigning one to each distribution, and evaluated performance on the held-out odours, again assigning one to each distribution. Nothing trial types were always assigned to Nothing distributions. To gain statistical power, we pooled neurons across mice for these analyses.

Generalized linear model. To assess the contributions of trial history, reward, reward prediction, sensory and motor-related variables to neural activity, we constructed a Poisson GLM with a bin width of 20 ms. This models the logarithm of the firing rate $\boldsymbol{\mu}_t$ within time bin t (in units of bin^{-1}, not s^{-1}) as a linear combination of predictor variables in row vector X_t, weighted by fitted coefficients in the column vector $\boldsymbol{\beta}$. The observed spike counts $\mathbf{y}_t$ are then treated as Poisson-distributed random variables that are independent across time points, conditional on the values of X_t. In matrix notation (Extended Data Fig. 5a):

$$\boldsymbol{\mu} = \exp(X\boldsymbol{\beta})$$
$$\mathbf{y}|\boldsymbol{\mu} \sim \text{Poisson}(\boldsymbol{\mu})$$

In constructing the design matrix X (with shape $B \times P$, where B is the number of time bins and P is the number of predictor variables; see Extended Data Fig. 5a,b), trial-length regressors (time in trial and trial history) were broken up into seven raised cosine basis functions, with 1-s spacing and 4-s width, tiling the 6 s of each (odour-cued) trial. Trial history consisted of reward magnitude, expected reward and RPE for up to two trials back. The height of each of these basis functions during the applicable time bins was directly proportional to the value of the corresponding regressor; for example, the height of the 1-back reward magnitude regressor following a 6 μl reward was three times higher than following a 2 μl reward. Reward, reward prediction and sensory regressors were scaled in the same manner, time-locked to reward or odour onset, and then convolved with a raised cosine basis that had been logarithmically scaled along the time axis[112].

Because the REDRL model implies that individual neuronal firing rates encode (a linear combination of) expectile values, we used five different expectile levels to scale the reward prediction regressors, corresponding to $\tau = 0.1, 0.3, 0.5, 0.7$ and 0.9. A unique family of sensory regressors was also included for each conditioned stimulus to capture potentially idiosyncratic odour responses. Licking, whisking and running regressors were convolved with the same basis but in a manner that allowed neural activity to be predictive as well as reactive; that is, they were also time-reversed around zero lag. Pupil area and face motion singular value decompositions from Facemap were input directly to the model without convolution. Finally, we included 20 nuisance regressors, which were evenly spaced raised cosine bases spanning the entire session, with a width equal to four times the spacing. These were included to flexibly capture the effects of electrode drift and avoid misattributing it to other variables that may have happened to be correlated. The contribution of these nuisance regressors to fractional deviance explained was eliminated by zeroing their coefficients before computing deviance. Unexpected reward trials (up to 20 per session) were only 3 s long; we simply removed the last 3 s of the trial-length regressors in these cases. The entire regressor matrix was *z*-scored before fitting.

We split the data trial-wise into a training set (85%) and testing set (15%). Within the training set, we performed fivefold cross-validation to select the regularization strength (λ), with splitting again performed trial-wise. We used group lasso regularization[113], which encourages sparsity between groups of variables but uses non-sparse L2 regularization on the within-variable bases, with groups given in Extended Data Fig. 5a. Thus, the loss *L* consisted of the negative log Poisson likelihood of the observed spike counts *l*, plus this regularization term:

$$\ell(\boldsymbol{\beta}; X, \mathbf{y}) = \sum_t (\mathbf{y}_t X_t \boldsymbol{\beta} - \exp(X_t \boldsymbol{\beta}) - \log(\mathbf{y}_t!))$$

$$\mathcal{L}(\boldsymbol{\beta}; X, \mathbf{y}) = -\ell(\boldsymbol{\beta}; X, \mathbf{y}) + \lambda \sum_i \sqrt{g_i}\, \|\boldsymbol{\beta}_i\|_2,$$

where g_i and $\boldsymbol{\beta}_i$ are the length and weight vector for variable group *i*, and $\|\cdot\|_2$ is the L2 norm.

Models were fit with GPU acceleration on the FAS Research Computing cluster using the GLM_Tensorflow_2 toolbox[114], which allowed us to fit all neurons from a given session in parallel. We minimized the loss with Adam optimization using a learning rate of 0.005. We fit models for eight logarithmically spaced values of λ between $10^{-4.5}$ and 10^{-1} and used an se_fraction of 0.75 for model selection. This corresponds to choosing the largest λ with model deviance within 0.75 standard errors (across cross-validation folds) of the deviance for the minimizing λ. All hyperparameters were chosen by examining speed and accuracy of fits on a small handful of pilot sessions. After selecting λ, we refit the model on the entire training set and then evaluated it on the test set. We followed the same procedure when fitting models in which trial history, expectile and motor regressors were held out before refitting. Because sensory regressors could in principle recapitulate all the information in the expectile regressors, we held these out alongside the expectiles.

GLM analysis. Once GLMs had been fit to each session and for each set of included variables, we computed deviance and fraction deviance explained on the held-out test set (Extended Data Fig. 5c,d) as:

$$\mathrm{Dev}(\mathbf{y}, \boldsymbol{\mu}) = 2 \sum_t \left(\mathbf{y}_t \log \frac{\mathbf{y}_t}{\boldsymbol{\mu}_t} - \mathbf{y}_t + \boldsymbol{\mu}_t \right)$$

$$\text{fraction deviance explained} = 1 - \frac{\mathrm{Dev}_{\mathrm{model}}}{\mathrm{Dev}_{\mathrm{null}}}$$

To conservatively estimate feature-specific contributions to encoding, we computed the difference between the full and reduced models for each family of regressors (Extended Data Fig. 5e,g). To complement this approach with a less conservative estimate, and to isolate the contribution of expectiles specifically (as opposed to odour-related responses), we also computed a 'kernel strength' by multiplying the history, expectile or motor coefficients from the full model by their respective basis functions, summing over the basis functions to get the complete kernel, and then integrating over time. To combine over groups of regressors within a family (for example, 0.1 through 0.9 expectiles), we summed these individual kernel strengths (or their absolute values in the case of history and motor) across family members (Extended Data Fig. 5f,h). When assessing correlations between differences in fraction deviance explained for the various reduced models, we excluded neurons whose fraction deviance explained was 0.01 or less for the full model to ensure that we were only considering neurons that were reasonably well-fit.

Fano factor analysis. Across-trial variability was quantified using the Fano factor, using published MATLAB code[115]. In brief, spike counts were computed in 100-ms bins for each trial, with a sliding step size of 50 ms. We then calculated the across-trial variance and mean of the spike count. For each combination of trial type and time bin, we computed the regression slope of the variance (*y*) as a function of the mean (*x*), weighted by the estimated sampling error for the variance.

To estimate the mean-matched Fano factor, we found the greatest common distribution of spike counts across all time points, binned at a resolution of 0.5 spikes. Then for each time point, we matched the analysed distribution of mean rates to this common distribution and repeated this procedure 10 times using different random seeds (Extended Data Fig. 6).

Computational modelling

In this section, we briefly review the theory behind various distributional RL algorithms before specifying the details of our implementation, for the purpose of comparing the learned code to neural activity and generating predictions for optogenetic perturbations. All models were trained for 2,000 trials per distribution.

REDRL. EDRL was first put forwards as a novel machine learning algorithm[39] and later used to explain dopamine neuron diversity in the mammalian midbrain[3,116]. EDRL approximately minimizes the expectile regression (ER) loss function:

$$\mathrm{ER}(V; \mathcal{D}, \tau) = \mathbb{E}_{Z \sim \mathcal{D}}[[\tau \mathbb{1}_{Z > V} + (1 - \tau)\mathbb{1}_{Z \leq V}](Z - V)^2],$$

where *V* is the value predictor, $\mathcal{D}$ is the target distribution, *Z* is a random sample from $\mathcal{D}$, τ is the asymmetry, and $\mathbb{1}$ is the indicator function, which is 1 when the subscript is satisfied and 0 when it is violated. It is an asymmetrically weighted squared error loss function; in this sense, it generalizes the mean (squared error loss, equivalent to the 0.5th expectile) just as quantiles generalize the median[54]. Note that *V* here is a scalar, and capitalization is merely for consistency with the RL literature.

EDRL and REDRL minimize this expectile regression loss function simultaneously for many values of τ, indexed by *i*, generally using SGD with respect to the value predictors (or their parameters). This formulation is sufficiently general that it can be combined with nonlinear function approximation and temporal difference learning methods, and its effectiveness has been demonstrated on the suite of Atari video games[39]. However, for simplicity, here we present the Rescorla–Wagner[117] version of the update rule for tabular states, so the random sample from $\mathcal{D}$ reduces to simply the experienced reward, *r*. This is the learning rule depicted in Fig. 2d:

$$\delta_i = r - V_i$$

$$V_i \leftarrow V_i + \alpha_i^- \cdot \delta_i, \text{ if } \delta_i \leq 0$$

$$V_i \leftarrow V_i + \alpha_i^+ \cdot \delta_i, \text{ if } \delta_i > 0$$

For the learning simulations (Fig. 2d), we used learning rates $\alpha = \alpha_i^+ + \alpha_i^- = 0.03$ and vectorized the updates. We initialized all value predictors to 2 to show their variable rates of convergence to their respective expectile values, but the algorithm is insensitive to this choice of initialization.

In the biological implementation of the REDRL algorithm (Fig. 2e–g), we decompose this update into two piecewise linear functions.

The first function models dopamine RPEs. As RPE is defined as actual minus predicted reward, the reward amount that elicits no change in dopamine firing relative to baseline—the 'zero-crossing point'[3]—is equivalent to the learned value prediction for that neuron. Pessimistic dopamine neurons have steeper slopes for rewards below their associated value prediction ($\alpha_i'^{-}$) and shallower slopes above it ($\alpha_i'^{+}$), reflecting relatively low learning rates from positive RPEs. The converse is true of optimistic dopamine neurons.

The second function defines the effects of dopamine on plasticity at corticostriatal synapses, and it differs between D1 and D2 MSNs (indexed by m) by a reflection over the y axis. D1 MSNs increase synaptic weights more from positive RPEs (ρ_m^{+}), whereas D2 MSNs increase synaptic weights more from negative RPEs[12–14] (ρ_m^{-}). In Fig. 2, $\rho_m^{-/+}$ are set to 0.75/3 for D1 and 3/0.75 for D2 MSNs, respectively. Although asymmetric, these synaptic weight updates are not fully dichotomous; D1 and D2 MSNs still learn slightly from dopamine changes in their non-preferred directions[66,118], in line with the shallower but non-zero slope of D1 and D2 receptor occupancy curves at baseline dopamine concentrations[66,119,120].

Composing these functions gives rise to the following update rules:

$$\begin{gathered}
\mathrm{D1}_i \leftarrow \mathrm{D1}_i + \alpha_i'^{-} \cdot \rho_{\mathrm{D1}}^{-} \cdot \delta_i,\ \text{if } \delta_i \leq 0\\
\mathrm{D1}_i \leftarrow \mathrm{D1}_i + \alpha_i'^{+} \cdot \rho_{\mathrm{D1}}^{+} \cdot \delta_i,\ \text{if } \delta_i > 0\\
\mathrm{D2}_i \leftarrow \mathrm{D2}_i - \alpha_i'^{-} \cdot \rho_{\mathrm{D2}}^{-} \cdot \delta_i,\ \text{if } \delta_i \leq 0\\
\mathrm{D2}_i \leftarrow \mathrm{D2}_i - \alpha_i'^{+} \cdot \rho_{\mathrm{D2}}^{+} \cdot \delta_i,\ \text{if } \delta_i > 0
\end{gathered}$$

Note that D1 and D2 neurons receive unique indices i, so there is no overlap in the idealized case. As a consequence of the opponent plasticity rule, changes in synaptic weights in D1 and D2 MSNs have opposing effects on the encoded value predictor, modelled simply by the identity function (for D1 MSNs) or its negation (for D2 MSNs):

$$\begin{gathered}
V_i = \mathrm{D1}_i\\
V_i = \max(\text{rewards}) - \mathrm{D2}_i
\end{gathered}$$

Therefore, this update rule becomes equivalent to the algorithmic rule from EDRL if we let $\alpha_i^{-} = \alpha_i'^{-} \cdot \rho_m^{-}$ and $\alpha_i^{+} = \alpha_i'^{+} \cdot \rho_m^{+}$.

The degree of optimism or pessimism is parameterized by the dimensionless quantity $\tau_i = \frac{\alpha_i^{+}}{\alpha_i^{+} + \alpha_i^{-}}$, which ranges from 0 to 1. Importantly, τ_i uses the net asymmetries learned by the MSNs as opposed to the asymmetries of the dopamine neurons. Both the expectile that is learned in the striatum and the zero-crossing point of the corresponding dopamine neuron are dictated by τ_i, which can give rise to multiple dopamine neurons with the same apparent asymmetry but different zero-crossing points, depending on whether they communicate with D1 or D2 MSNs. This stands in contrast to the EDRL model, in which the dopamine neuron asymmetries alone fully determine the zero-crossing point. Nonetheless, REDRL also predicts a positive correlation between zero-crossing points and asymmetries, as previously observed[3].

For D1 MSNs, $\rho_m^{+} > \rho_m^{-}$ and so τ_i skews optimistic; analogously, for D2 MSNs, $\rho_m^{+} < \rho_m^{-}$, so τ_i skews pessimistic. The precise distribution of τ's will depend on the distribution of dopamine neuron asymmetries ($\alpha_i'^{+}$ and $\alpha_i'^{-}$) as well as the ratio of ρ_m^{+} to ρ_m^{-}, neither of which has been measured precisely. To avoid making too many assumptions and to simplify interpretation, we plotted all REDRL results based on a simulation of 10 predictors with uniform spacing of τ_i between 0.05 and 0.95, with all $\tau_i > 0.5$ assigned to D1 MSNs and all $\tau_i < 0.5$ assigned to D2 MSNs. Furthermore, we directly computed the expectiles of the relevant reward distributions (rather than obtaining them incrementally from samples and updates) to eliminate noise. We confirmed that all of our main results were robust to these choices of τ and simulation approach.

Finally, we emphasize that differential plasticity in D1 and D2 MSNs in response to positive and negative dopamine transients is a known empirical feature of this system[12–14]; our novel theoretical contribution is to show how a piecewise linear plasticity rule fulfils the precise mathematical requirements for D1 and D2 MSNs to converge preferentially to optimistic and pessimistic expectiles, respectively.

Quantile distributional RL. Quantile distributional RL (QDRL) is exactly akin to EDRL, except that we minimize the quantile regression (QR) loss[121]:

$$\mathrm{QR}(V; \mathcal{D}, \tau) = \mathbb{E}_{Z\sim\mathcal{D}}[[\tau \mathbb{1}_{Z>V} + (1-\tau)\mathbb{1}_{Z<V}]|Z-V|],$$

This is an asymmetrically weighted absolute value loss function, which would return the median when positive and negative errors are balanced ($\tau = 0.5$). The update rule, derived by SGD, utilizes only the sign of the prediction error, not its magnitude[54]:

$$\begin{gathered}
V_i \leftarrow V_i - \alpha_i^{-},\ \text{if } \delta_i < 0\\
V_i \leftarrow V_i + \alpha_i^{+},\ \text{if } \delta_i > 0
\end{gathered}$$

Unlike expectiles, quantiles have an intuitive interpretation: the τ-th quantile is the number such that τ fraction of samples from the distribution fall below that value and $1-\tau$ fall above it. It is therefore the inverse of the CDF. We additionally implemented a 'reflected' version of QDRL by applying the same transformation to D2 MSNs, those predictors with $\tau_i < 0.5$.

We also note that it is possible to interpolate between EDRL and QDRL using Huber quantiles[121,122]. This is simply an asymmetric squared loss within a certain interval (controlled by a hyperparameter κ), and a standard quantile loss outside this interval. The update rule is likewise a combination of EDRL and QDRL: piecewise linear within some range before saturating. This rule would obtain if, for example, plasticity could only change some maximum amount in either direction at any given time, as is likely the case in the brain. Of note, the Huber quantile loss is frequently used in machine learning applications[121].

Categorical distributional RL. Categorical distributional RL (CDRL)[33] adopts a very different approach to learning the reward distribution. Rather than a quantile or expectile function, CDRL imagines a set of 'atoms', which function similarly to bins of a histogram. For that reason, we modelled these 'categorical codes' using one hypothetical neuron per reward size (0–8 μl), in increments of 2 μl. The height of that bin was then assumed to be linearly (and positively) related to the firing rate of that neuron. Generalizing this scheme to use basis functions over bin values does not qualitatively alter the predictions.

Laplace and cumulative code. The Laplace code[40] grew out of an effort to devise a fully local temporal difference learning rule for distributional RL. Its teaching signal is simply a sigmoidal function of reward: if reward exceeds some threshold, the neuron fires, and thresholds are heterogeneous across neurons. In the limit of infinitely steep sigmoids (Heaviside step functions), the value predictors converge to the probability that the reward exceeds the given threshold (discounted and summed over future time steps, in the case of temporal difference learning). This exceedance probability is equal to 1 − CDF of the reward distribution, for our simplified Rescorla–Wagner setting. Analogous to CDRL, we chose to model neural activity as linearly and positively related to this value of 1 − CDF at each of the reward bins. For completeness, we also investigated a 'cumulative' code, which was just the CDF at each reward bin, or 1 − the Laplace code. The spatial derivative of this cumulative code is then equivalent to the categorical code, assuming sufficient support.

Actor uncertainty model. The actor uncertainty model[66] manages to learn about reward uncertainty using biologically plausible learning rules in D1 and D2 MSNs. We therefore wanted to test its predictions against these other models. The actor uncertainty model makes use of two value predictors: one D1 and one D2 MSN, which learn as follows:

$$V = \mathrm{D1} - \mathrm{D2}$$
$$\mathrm{D1} \leftarrow \mathrm{D1} + \alpha|r - V|_{+} - \eta \cdot \mathrm{D1}$$
$$\mathrm{D2} \leftarrow \mathrm{D2} + \alpha|r - V|_{-} - \eta \cdot \mathrm{D2}$$

Here, $|x|_{+} = \max(x, 0)$ and $|x|_{-} = \max(-x, 0)$, and $0 < \eta < 1$ scales the decay term to ensure stability. Using this model, it can be shown[66] that D1 − D2 encodes an estimate of mean reward, and D1 + D2 encodes an estimate of reward spread. For our implementation, we set $\alpha = 0.1$ and $\eta = 0.01$.

Distributed actor uncertainty model. The distributed actor uncertainty model[123] works similarly, except that we allowed there to be different learning rates α_i^{+} and α_i^{-} for D1 and D2 MSNs, respectively, just as in the distributional RL setting. The difference $V_i = \mathrm{D1}_i - \mathrm{D2}_i$ approximates the τ_i-th expectile, biased by η. For our simulations, we chose $\alpha = \alpha_i^{+} + \alpha_i^{-} = 0.2$ and $\eta = 0.01$.

Comparing models with recording data. For each hypothetical unit (representing, for example, a single expectile level or reward bin) and trial type, we simulated 50 trials and 100 cells by adding independent Gaussian noise (s.d. = five times the standard deviation across all predictors for that code) to the converged value predictors, to generate some jitter for odours associated with the same exact distribution. Just as in the neural data, we computed trial-wise correlations with expected value and compared this with a baseline in which trial-type labels were randomly shuffled independently for each neuron. We averaged across (simulated) trials before applying dimensionality reduction or computing cosine distances. To generate predictions for optimistic or pessimistic neurons alone, we took appropriate subsets of the simulated data (for example, only neurons with $\tau < 0.5$ for pessimistic expectiles) before applying these analyses.

Modelling perturbations. Simulating optogenetic inhibition and excitation in these models (Extended Data Fig. 11) required slightly different choices, depending on the type of code. For expectile, quantile and actor uncertainty-based models, we clamped the relevant simulated neuron (or neurons) to either 0 or 8, the maximum reward value across all distributions, to simulate model inhibition and excitation, respectively. Note that it was the neural activity ($\mathrm{D1}_i$ or $\mathrm{D2}_i$) that we were directly clamping when applicable, not the value prediction that it encoded (V_i). For the expectile and quantile models, optimistic and pessimistic perturbations meant clamping the value of predictors with $\tau_i > 0.5$ and $\tau_i < 0.5$, respectively. For the actor uncertainty model, they were identified with the D1 and D2 MSN, respectively. Finally, for the distributed actor uncertainty model, we implemented two versions of the perturbation: one in which all D1 (optimistic) or all D2 (pessimistic) neurons were manipulated, and one in which only those with $\tau_i > 0.5$ or $\tau_i < 0.5$, respectively, were manipulated. We call the latter the 'partial distributed actor uncertainty' model, for the purposes of model comparison. For the actor uncertainty models, it is only the difference $\mathrm{D1}_i - \mathrm{D2}_i$ that is bounded within the range of reward sizes, not the activities individually. We therefore added or subtracted a fixed amount (the maximum reward size across all trial types, 8 µl, bounded below by zero) across reward predictors to simulate excitation or inhibition, respectively, in these models, rather than clamping their value to a constant.

For categorical, cumulative and Laplace codes, the semantics of each simulated neuron are different: their activations range from 0 to 1 and encode a (cumulative) probability, rather than a value. Thus, inhibiting or exciting them meant changing the relevant probability to 0 or 1, respectively. Pessimistic neurons were those that corresponded to the 0-µl or 2-µl bins, and optimistic neurons corresponded to 6 µl and 8 µl. To reconstitute a properly normalized probability distribution after the perturbation, in the case of the categorical code, we divided by the sum of the predictors (or made it a uniform distribution if the sum was zero). For the categorical and Laplace codes, we took the spatial derivative of the implied CDF, subtracted off the minimum if any value was negative, and then divided by the sum (or made it uniform if the sum was zero).

In all cases, we found the mean of the (imputed) perturbed probability distribution and then compared it with the mean without any perturbation, separately for inhibition and excitation, to model the effect of optogenetic manipulation on lick rate.

Comparing models with optogenetic perturbation data. We used the predicted manipulation to no manipulation differences from each model as a regressor with which to predict the difference in licking during the last half second of the trace period across trial types, averaged across mice, using linear regression (with no intercept term). Separate regressions were fit for inhibition and excitation to allow for potentially different scaling in each case, and their coefficients of determination were averaged to produce a single summary measure of goodness of fit.

Reporting summary

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

Data availability

Preprocessed data are documented and available for download on Dryad[124].

Code availability

The code used for analysis and generation of all figures in this paper is available on GitHub[125] (https://github.com/alowet/distributionalRL).

75. Gong, S. et al. A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. *Nature* **425**, 917–925 (2003).
76. Gong, S. et al. Targeting Cre recombinase to specific neuron populations with bacterial artificial chromosome constructs. *J. Neurosci.* **27**, 9817–9823 (2007).
77. Gerfen, C. R., Paletzki, R. & Heintz, N. GENSAT BAC cre-recombinase driver lines to study the functional organization of cerebral cortical and basal ganglia circuits. *Neuron* **80**, 1368–1383 (2013).
78. Govorunova, E. G., Sineshchekov, O. A., Janz, R., Liu, X. & Spudich, J. L. Natural light-gated anion channels: a family of microbial rhodopsins for advanced optogenetics. *Science* **349**, 647–650 (2015).
79. Li, N. et al. Spatiotemporal constraints on optogenetic inactivation in cortical circuits. *eLife* **8**, e48622 (2019).
80. Cohen, J. Y., Haesler, S., Vong, L., Lowell, B. B. & Uchida, N. Neuron-type-specific signals for reward and punishment in the ventral tegmental area. *Nature* **482**, 85–88 (2012).
81. Thiele, S. L., Warre, R. & Nash, J. E. Development of a unilaterally-lesioned 6-OHDA mouse model of Parkinson's disease. *J. Vis. Exp.* **60**, e3234 (2012).
82. Dana, H. et al. High-performance calcium sensors for imaging activity in neuronal populations and microcompartments. *Nat. Methods* **16**, 649–657 (2019).
83. Klapoetke, N. C. et al. Independent optical excitation of distinct neural populations. *Nat. Methods* **11**, 338–346 (2014).
84. Lee, J. & Sabatini, B. L. Striatal indirect pathway mediates exploration via collicular competition. *Nature* **599**, 645–649 (2021).
85. Uchida, N. & Mainen, Z. F. Speed and accuracy of olfactory discrimination in the rat. *Nat. Neurosci.* **6**, 1224–1229 (2003).
86. Pavlov, I. P. *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex* (Oxford Univ. Press, 1927).
87. Jun, J. J. et al. Fully integrated silicon probes for high-density recording of neural activity. *Nature* **551**, 232–236 (2017).
88. Steinmetz, N. A. et al. Neuropixels 2.0: a miniaturized high-density probe for stable, long-term brain recordings. *Science* **372**, eabf4588 (2021).
89. Pachitariu, M., Sridhar, S., Pennington, J. & Stringer, C. Spike sorting with Kilosort4. *Nat. Methods* **21**, 914–921 (2024).
90. Zhou, Z. C. et al. Deep-brain optical recording of neural dynamics during behavior. *Neuron* **111**, 3716–3738 (2023).
91. Pachitariu, M. et al. Suite2p: beyond 10,000 neurons with standard two-photon microscopy. Preprint at *bioRxiv* https://doi.org/10.1101/061507 (2017).

92. Friedrich, J., Zhou, P. & Paninski, L. Fast online deconvolution of calcium imaging data. *PLoS Comput. Biol.* **13**, e1005423 (2017).
93. Lopes, G. et al. Bonsai: an event-based framework for processing and controlling data streams. *Front. Neuroinform.* **9**, 7 (2015).
94. Pisanello, M. et al. Tailoring light delivery for optogenetics by modal demultiplexing in tapered optical fibers. *Sci. Rep.* **8**, 4467 (2018).
95. Lee, J., Wang, W. & Sabatini, B. L. Anatomically segregated basal ganglia pathways allow parallel behavioral modulation. *Nat. Neurosci.* **23**, 1388–1398 (2020).
96. Sanders, J. I. & Kepecs, A. A low-cost programmable pulse generator for physiology and behavior. *Front. Neuroeng.* **7**, 43 (2014).
97. Shamash, P., Carandini, M., Harris, K. & Steinmetz, N. A tool for analyzing electrode tracks from slice histology. Preprint at *bioRxiv* https://doi.org/10.1101/447995 (2018).
98. Wang, Q. et al. The Allen Mouse Brain Common Coordinate Framework: a 3D reference atlas. *Cell* **181**, 936–953.e20 (2020).
99. Claudi, F. et al. Visualizing anatomically registered data with brainrender. *eLife* **10**, e65751 (2021).
100. Chon, U., Vanselow, D. J., Cheng, K. C. & Kim, Y. Enhanced and unified anatomical labeling for a common mouse brain atlas. *Nat. Commun.* **10**, 5067 (2019).
101. Claudi, F. et al. BrainGlobe Atlas API: a common interface for neuroanatomical atlases. *J. Open Source Softw.* **5**, 2668 (2020).
102. Hintiryan, H. et al. The mouse cortico-striatal projectome. *Nat. Neurosci.* **19**, 1100–1114 (2016).
103. Peters, A. J., Fabre, J. M. J., Steinmetz, N. A., Harris, K. D. & Carandini, M. Striatal activity topographically reflects cortical activity. *Nature* **591**, 420–425 (2021).
104. Harris, C. R. et al. Array programming with NumPy. *Nature* **585**, 357–362 (2020).
105. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
106. McKinney, W. Data structures for statistical computing in Python. In *Proc. 9th Python in Science Conference* (eds van der Walt, S. & Millman, J.) 56–61 (SciPy, 2010).
107. Buitinck, L. et al. API design for machine learning software: experiences from the scikit-learn project. In *Proc. ECML PKDD Workshop: Languages for Data Mining and Machine Learning* (eds Crémilleux, B. et al.) 108–122 (ECML PKDD, 2013).
108. Seabold, S. & Perktold, J. Statsmodels: econometric and statistical modeling with Python. In *Proc. 9th Python in Science Conference* (eds van der Walt, S. & Millman, J.) 92–96 (SciPy, 2010).
109. Hunter, J. D. Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
110. Waskom, M. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).
111. Dietterich, T. G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.* **10**, 1895–1923 (1998).
112. Pillow, J. W. et al. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature* **454**, 995–999 (2008).
113. Yuan, M. & Lin, Y. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Series B Stat. Methodol.* **68**, 49–67 (2006).
114. Tseng, S.-Y., Chettih, S. N., Arlt, C., Barroso-Luque, R. & Harvey, C. D. Shared and specialized coding across posterior cortical areas for dynamic navigation decisions. *Neuron* **110**, 2484–2502.e16 (2022).
115. Churchland, M. M. et al. Stimulus onset quenches neural variability: a widespread cortical phenomenon. *Nat. Neurosci.* **13**, 369–378 (2010).
116. Eshel, N., Tian, J., Bukwich, M. & Uchida, N. Dopamine neurons share common response function for reward prediction error. *Nat. Neurosci.* **19**, 479–486 (2016).
117. Rescorla, R. A. & Wagner, A. R. in *Classical Conditioning II: Current Research and Theory* (eds Black, A. H. & Prokasy, W. F.) 64–99 (Appleton-Century-Crofts, 1972).
118. Gurney, K. N., Humphries, M. D. & Redgrave, P. A new framework for cortico-striatal plasticity: behavioural theory meets in vitro data at the reinforcement–action interface. *PLoS Biol.* **13**, e1002034 (2015).
119. Rice, M. E. & Cragg, S. J. Dopamine spillover after quantal release: rethinking dopamine transmission in the nigrostriatal pathway. *Brain Res. Rev.* **58**, 303–313 (2008).
120. Dreyer, J. K., Herrik, K. F., Berg, R. W. & Hounsgaard, J. D. Influence of phasic and tonic dopamine release on receptor activation. *J. Neurosci.* **30**, 14273–14283 (2010).
121. Dabney, W., Rowland, M., Bellemare, M. & Munos, R. Distributional reinforcement learning with quantile regression. In *Proc. 32nd AAAI Conference on Artificial Intelligence* (eds McIlraith, S. A. & Weinberger, K. Q.) 2892–2901 (AAAI Press, 2018).
122. Huber, P. J. Robust estimation of a location parameter. *Ann. Math. Stat.* **35**, 73–101 (1964).
123. Romero Pinto, S. & Uchida, N. Tonic dopamine and biases in value learning linked through a biologically inspired reinforcement learning model. Preprint at *bioRxiv* https://doi.org/10.1101/2023.11.10.566580 (2023).
124. Lowet, A. S. et al. Data from: an opponent striatal circuit for distributional reinforcement learning. *Dryad* https://doi.org/10.5061/dryad.80gb5mm0m (2024).
125. Lowet, A. S. alowet/distributionalRL: Publication-ready version (v1.0.2). *Zenodo* https://doi.org/10.5281/zenodo.14554845 (2024).
126. Chandak, Y. et al. Universal off-policy evaluation. In *Proc. Advances in Neural Information Processing Systems 34* (eds Ranzato, M. et al.) 27475–27490 (NeurIPS, 2021).
127. Gagne, C. & Dayan, P. Peril, prudence and planning as risk, avoidance and worry. *J. Math. Psychol.* **106**, 102617 (2022).
128. Rockafellar, R. T. & Uryasev, S. Optimization of conditional value-at-risk. *J. Risk* **2**, 21–41 (2000).
129. Fiser, J., Berkes, P., Orbán, G. & Lengyel, M. Statistically optimal perception and learning: from behavior to neural representations. *Trends Cogn. Sci.* **14**, 119–130 (2010).

Acknowledgements We thank members of the Uchida laboratory for valuable discussions and comments on the manuscript; E. Soucy and B. Graham for providing critical assistance with instrumentation; A. Girasole and B. Sabatini for sharing the GtACR1 mouse line; X. Cai, B. Sabatini, C. Harvey and S. J. Gershman for helpful conversations; and M. Carandini, K. D. Harris, A. J. Peters and other members of the Cortex laboratory for their advice on Neuropixels recording. This work was supported by grants from the US National Institutes of Health (R01NS116753 to N.U. and J.D., and F31NS124095 to A.S.L.), the Human Frontier Science Program (LT000801/2018 to S.M.), the Harvard Brain Science Initiative, and the Brain and Behavior Research Foundation (NARSAD Young Investigator no. 30035 to S.M.). We thank the Harvard Center for Biological Imaging (RRID: SCR_018673) for infrastructure and support for ex vivo imaging, which was funded in part by the Simmons Award (to A.S.L.). The computations in this paper were run in part on the FASRC Cannon cluster supported by the FAS Division of Science Research Computing Group at Harvard University.

Author contributions A.S.L. and N.U. designed the experiments. A.S.L. and M.M. performed the experiments, with initial help from S.M. A.S.L. and M.M. preprocessed the data. A.S.L. analysed the data and devised and implemented the computational models with input from J.D. and N.U. Q.Z. implemented ANN-based distributional decoding under the supervision of J.D. A.S.L. wrote the first draft of the manuscript and created the figures. N.U., J.D., S.M. and A.S.L. edited the manuscript.

Competing interests The authors declare no competing interest.

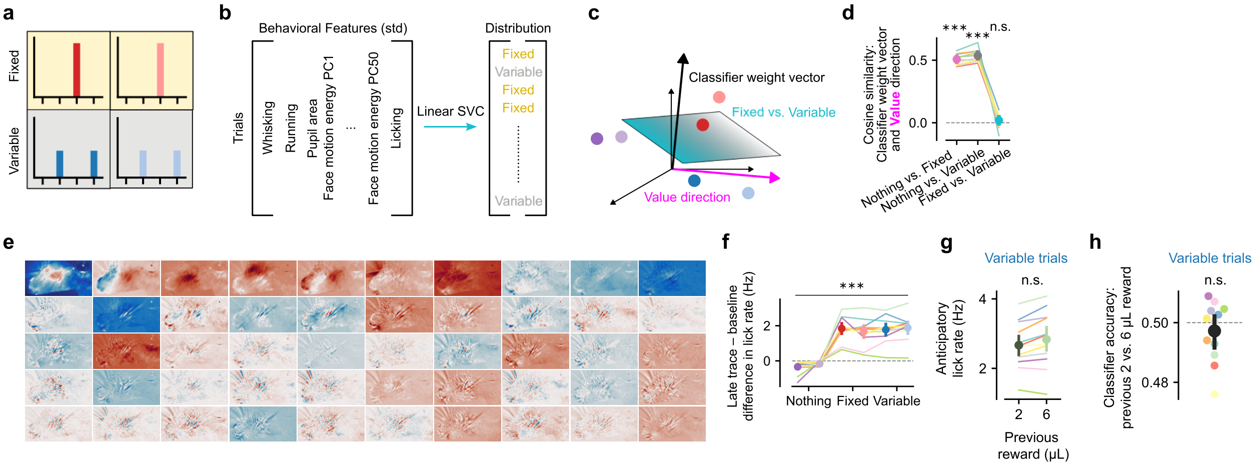
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-024-08488-5.

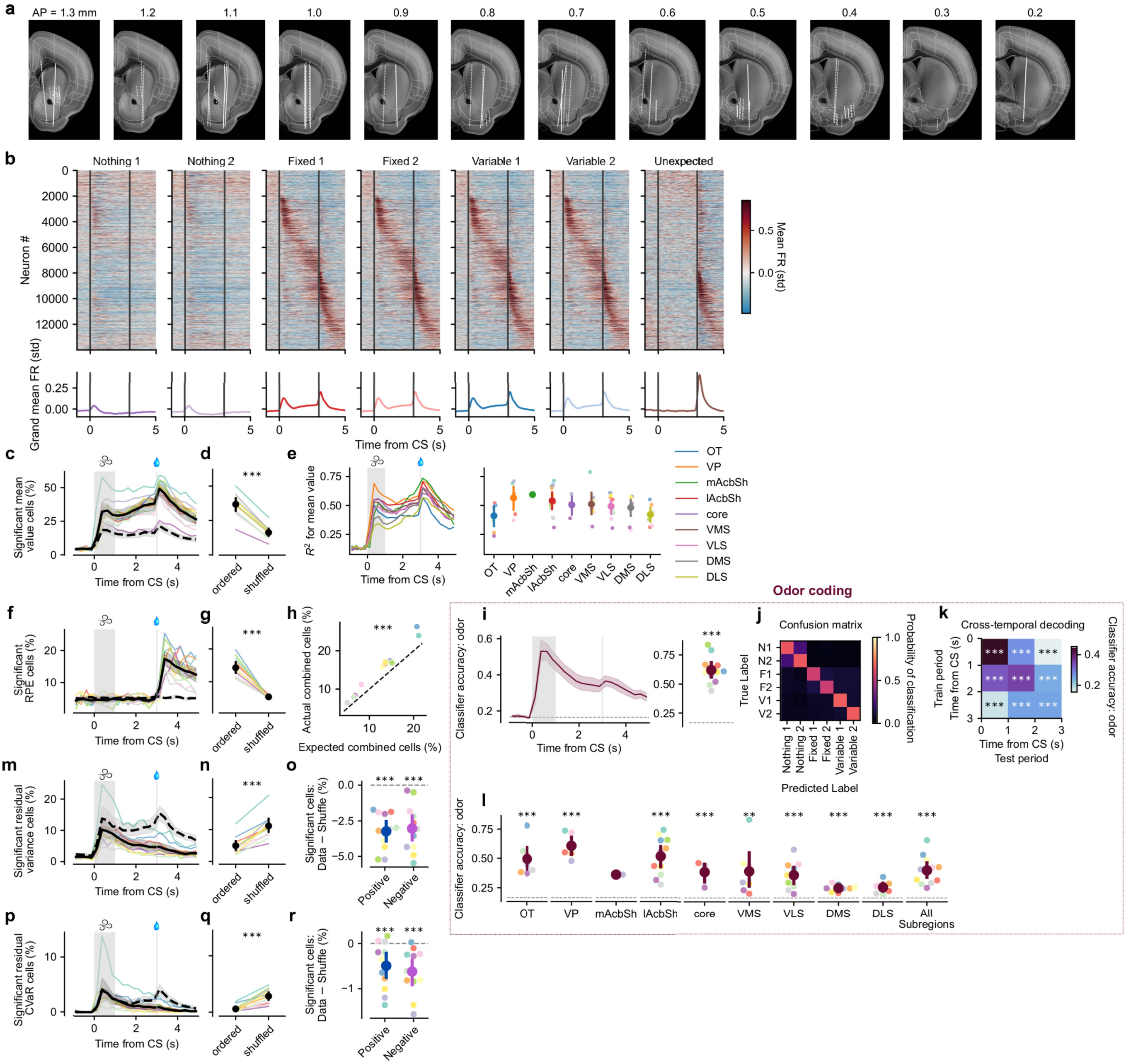
Correspondence and requests for materials should be addressed to Jan Drugowitsch or Naoshige Uchida.

Peer review information *Nature* thanks Ilya Monosov, Blake Richards and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at http://www.nature.com/reprints.

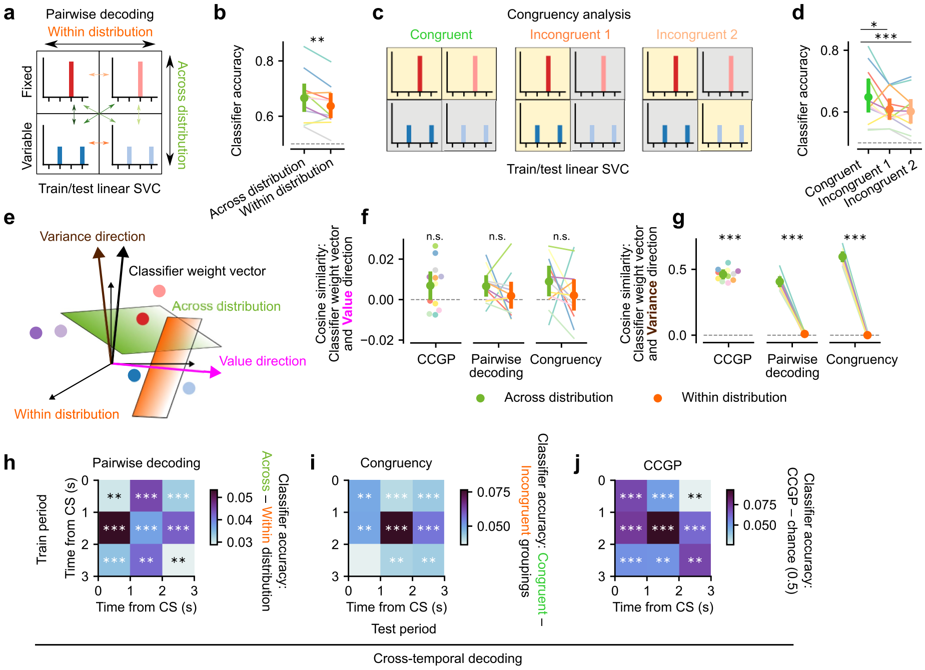


Extended Data Fig. 1 | Additional behavioral analysis. a, Schematic for behavioral classification analysis in panels **b**–**e**. Odours corresponding to the same distribution were treated as the same class. This is illustrated for the case of Fixed vs. Variable odour classification, with the background shading (yellow vs. grey) indicating the target for the classifier. **b**, Schematic of behavioural classification. On each validation fold, whisking, running, pupil area, licking, and the top 50 face motion energy PCs in the training set were z-scored and then passed to a support vector classifier (SVC) with a linear kernel, which predicts the associated distribution. **c**, Schematic of orthogonality analysis. The weights learned by the SVC define a vector orthogonal to the hyperplane that best separates distributions. A separate vector can be defined by regressing the mean reward ("Value direction") of each trial against their corresponding behavioural regressors. While the SVC hyperplane considers only four odours at a time, the regression direction takes into account all six odours. **d**, Cosine similarity between the classifier weight vector and the Value direction. Any differences in behavior between Fixed and Variable trials are orthogonal to Value (relative to chance level of 0: $p < 0.001$ for Nothing vs. Fixed, $p < 0.001$ for Nothing vs. Variable, $p = 0.154$ for Fixed vs. Variable). **e**, Spatial masks corresponding to face motion energy PCs in an example session, sorted by variance explained. Successive PCs emphasize finer and finer aspects of mouse whisking, sniffing, and licking behavior. **f**, The difference in lick rate between the late trace and baseline (1–0 s before odour onset) periods is significant for all trial types, including a decrease below baseline for both Nothing odours (all p's < 0.001). **g**, Anticipatory lick rate does not differ for Variable odours based on whether the previous trial with that odour led to 2 or 6 μL of reward ($p = 0.179$). **h**, A linear classifier trained to predict the amount of reward delivered on the previous Variable trial of a given odour performs at chance accuracy of 50% ($p = 0.326$).

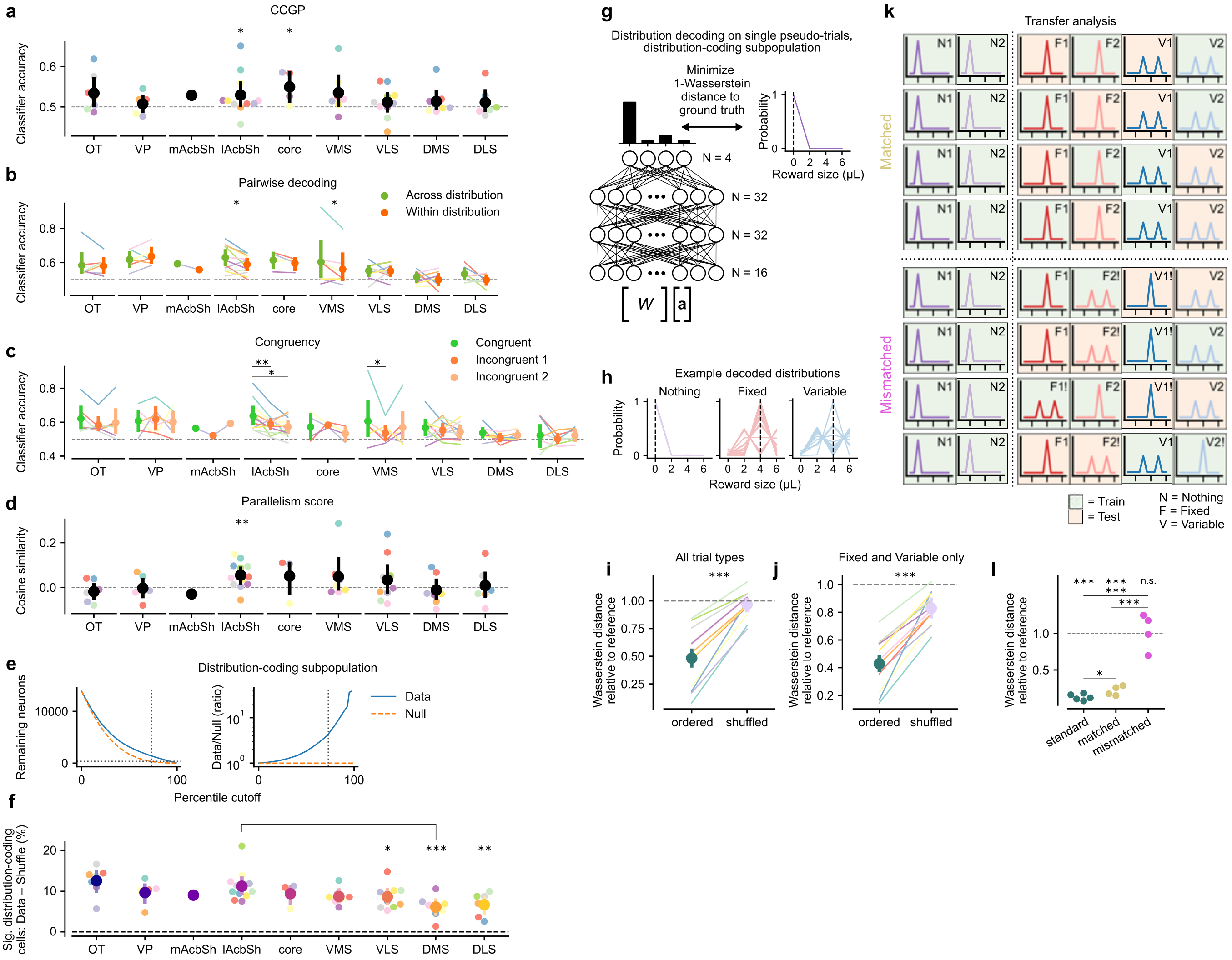


Extended Data Fig. 2 | See next page for caption.

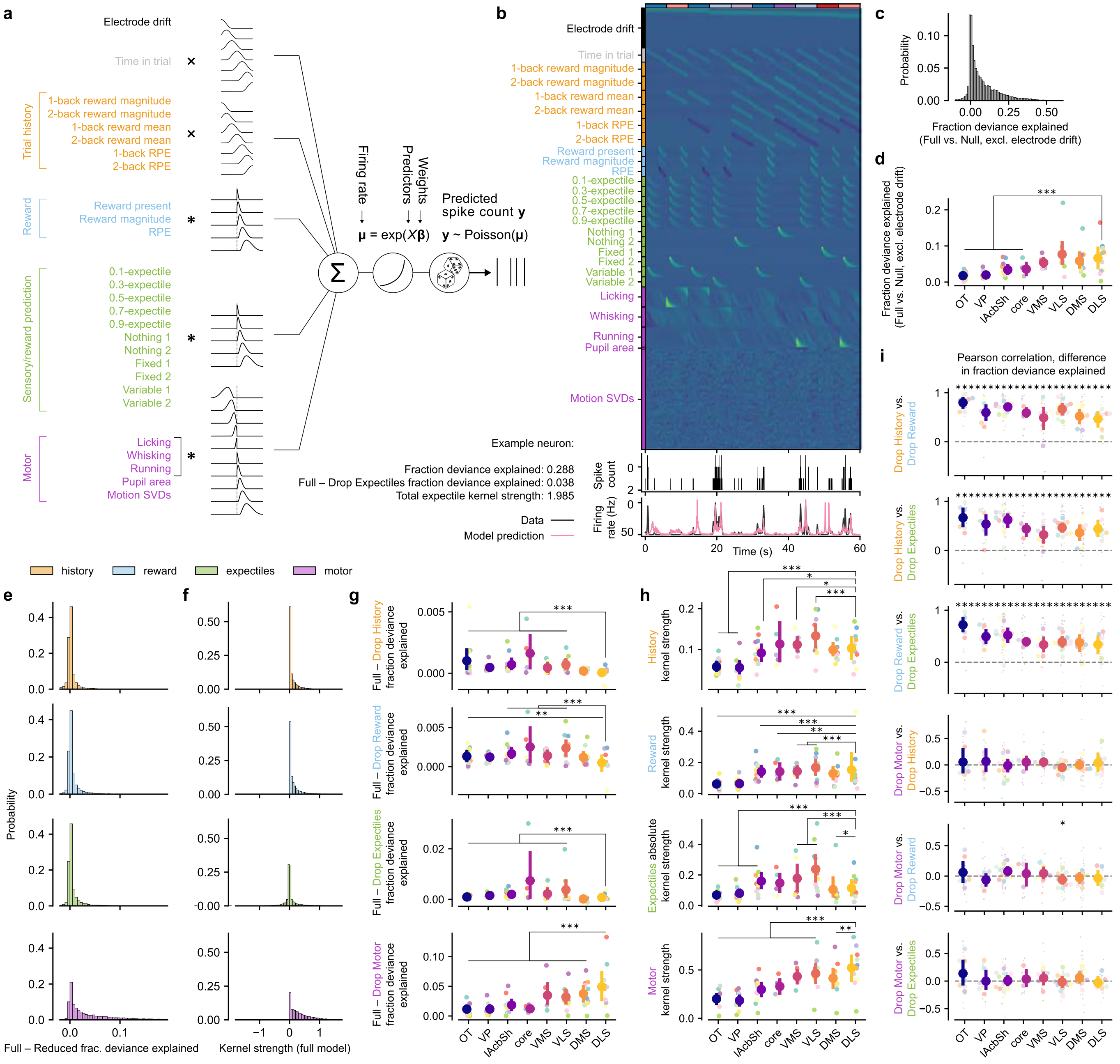
Extended Data Fig. 2 | Value, RPE, odour and risk coding across the striatum. **a**, Serial coronal sections showing recording sites of probe insertions (white dotted lines), registered to the Allen Common Coordinate Framework. **b**, *Top*, heatmaps showing average z-scored firing rate in response to each odour for each neuron. Neurons were sorted according to the time of peak activity when averaged on half of Variable 2 odour trials, and then plotted in this same order for the remainder of trials, grouped by trial type. The seventh and final trial type corresponds to Unexpected rewards, which were not preceded by an odour. *Bottom*, grand average z-scored firing rate across all neurons. **c**, Fraction of neurons that significantly correlate with mean reward, computed separately in non-overlapping 250 ms time bins. Each mouse is shown in a different colour, with the mean ± 95% c.i. across mice shown in solid black. Dashed line is the average across mice after shuffling the mapping between odours and distributions, thereby accounting for pure odour coding. **d**, Average percentage of significant cells during the late trace period ($p < 0.001$). **e**, *Left*, cross-validated R^2 predicting the mean reward on each trial as a function of striatal subregion, computed separately in non-overlapping 250 ms time bins. To ensure fair comparison across subregions, we for each animal generated multiple pseudo-populations of 40 neurons each by repeatedly sampling without replacement neural subpopulation across session boundaries until there were fewer than 40 neurons remaining. Animals with fewer than 40 neurons in the given region were excluded. Lines show averages across mice for each subregion. *Right*, average R^2 over the late trace period. Smaller dots show averages across pseudo-populations for each mouse with at least 40 neurons in that region. **f**, Same as **c**, except showing the fraction of neurons that significantly correlate with reward prediction error (RPE), defined as the difference between actual and expected reward. **g**, Same as **d**, except showing the average percentage of significant cells during the outcome period, 0–1 s after reward delivery ($p < 0.001$). **h**, The actual fraction of cells in each mouse that significantly correlated with both mean value and RPE was compared to the product of the individual fractions for mean and RPE-coding cells (the predicted fraction assuming independence; $p < 0.001$). **i**, *Left*, decoding accuracy across time of a multinomial logistic regression classifier decoding odour identity (dashed = chance level of 1/6). *Right*, quantification of odour classification accuracy during the odour period ($p < 0.001$ relative to chance level). **j**, Confusion matrix for odour decoding during the odour period shows high decoding accuracy for all odours, with relatively higher confusability for odours with the same mean. **k**, Cross-temporal decoding reveals that odour decoding is stable across time, allowing a classifier trained e.g. on late trace period activity to generalize well above chance to the odour period, and vice versa (all p's < 0.001 relative to chance level of 1/6). **l**, Pseudo-population odour decoding across subregions (see Methods section titled “Comparisons across subregions, hemispheres, and genotypes”). OT, olfactory tubercle; VP, ventral pallidum; mAcbSh, medial nucleus accumbens shell; lAcbSh, lateral nucleus accumbens shell; core, nucleus accumbens core; VMS, ventromedial striatum; VLS, ventrolateral striatum; DMS, dorsomedial striatum; DLS, dorsolateral striatum ($N = 1$ mouse for mAcbSh, $p = 0.006$ for VMS, all other p's < 0.001). **m**, Same as **c**, except showing the fraction of neurons that significantly correlate with variance, after regressing out the contribution of mean reward coding separately for each time bin. **n**, Average percentage of significant Residual Variance cells during the late trace period is *less* than would be predicted from odour coding alone ($p < 0.001$). **o**, Significantly fewer neurons encode residual variance positively and negatively than expected by chance (positive and negative p's < 0.001). **p-r**, Same as **m-o**, but for conditional value at risk (CVaR), a common risk measure used in finance and reinforcement learning[126–128], defined as the expected value within the lower α-quantile of a probability distribution. For our distributions, this will be equivalent to the mean for $\alpha > 0.5$ and equivalent to the minimum value for $\alpha < 0.5$, which differs only for the Variable distribution, where it is 2. The latter is what we plot here, after regressing out mean coding. Again, there are fewer residual CVaR cells than would be expected from odour coding alone ($p < 0.001$) and this is true for both positive- and negative-coding cells (both p's < 0.001).



Extended Data Fig. 3 | Distributional coding is robust, orthogonal to value, and consistent across time. a, Schematic of pairwise decoding analysis. Linear SVCs were trained on individual Fixed and Variable odours, two at a time. This resulted in six possible dichotomies, four of which encompassed one Fixed and one Variable odour (green arrows; "Across distribution") and two of which compared odours cuing the same exact distribution (orange arrows; "Within distribution"). **b**, Pairwise decoding during the late trace period was significantly better for across- than within-distribution pairs, consistent with distributional but not traditional RL ($p = 0.001$). **c**, Schematic of congruency analysis, which considered all four Fixed and Variable odours simultaneously. In the Congruent grouping, both Fixed odours were assigned to one class (yellow background) and both Variable odours were assigned to the other class (grey background), just as was done for behavioral decoding. By contrast, in the Incongruent groupings, class assignments cut across Fixed and Variable distributions. **d**, Classifier accuracy in the late trace period was higher for Congruent than Incongruent pairs, again consistent with distributional but not traditional RL (Congruent: $p = 0.028$ vs. Incongruent 1, $p < 0.001$ vs. Incongruent 2). **e**, Schematic illustrating the classifier weight vector (normal to the separating hyperplane for across- or within-distribution classifications) and the regression weight vector (for Value or Variance). **f**, Quantification of cosine similarity between the classifier weight vector and the Value direction shows that the vectors are not significantly different from orthogonal (CCGP: $p = 0.071$ cosine similarity relative to chance value of 0; Pairwise: $p = 0.797$ Across- vs. Within-distribution absolute cosine similarity; Congruency: $p = 0.493$ Across- vs. Within-distribution absolute cosine similarity). **g**, Same as **f**, but for Variance rather than Value direction ($p < 0.001$ for all comparisons). **h-j**, Cross-temporal decoding for the pairwise, congruency, and CCGP analyses. Distributional RL is favored during every time period between odour onset and reward delivery, and decoders trained during one period almost always generalize to other time periods.

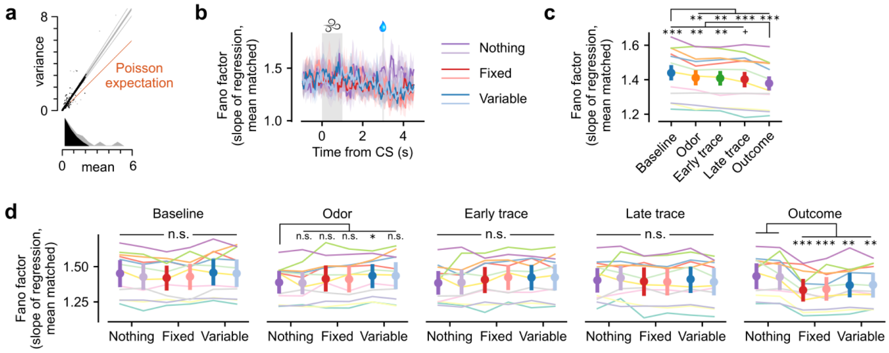


Extended Data Fig. 4 | A distribution-coding subpopulation is over-represented in the lAcbSh and permits ANN-based distribution decoding. **a**, Pseudo-population CCGP across subregions (relative to chance level of 0.5: $p = 0.059, 0.473, 0.044, 0.017, 0.088, 0.346, 0.257, 0.407$, and 0.133 for OT, VP, mAcbSh, lAcbSh, core, VMS, VLS, DMS, and DLS, respectively. Same order applies to all statistics in this figure). Pseudo-populations were constructed as in Extended Data Fig. 2l. **b**, Pseudo-population pairwise decoding across subregions (Across- vs. Within-distribution: $p = 0.861, 0.344, 0.883, 0.010, 0.409, 0.040, 0.882, 0.482, 0.106$). **c**, Pseudo-population congruency analysis across subregions (Congruent vs. Incongruent 1: $p = 0.097, 0.817, 0.744, 0.007, 0.832, 0.047, 0.523, 0.138, 0.523$; Congruent vs. Incongruent 2: $p = 0.306, 0.760, 0.815, 0.010, 0.473, 0.177, 0.316, 0.486, 0.985$). **d**, Parallelism score across subregions (relative to chance level of 0: $p = 0.300, 0.878, 1.00, 0.001, 0.229, 0.243, 0.273, 0.615, 0.764$). **e**, *Left*, fraction of neurons with classifier coefficients above the percentile cutoff for all three (CCGP, pairwise, and congruency) analyses. Horizontal dotted line indicates level at which 2.5% of null coefficients fell above the cutoff; this was the 73rd percentile (vertical dotted line), and retained 11.43% of neurons. *Right*, ratio of data to null coefficients falling above the cutoff (log scale). **f**, Fraction of distribution-coding cells in each subregion. This fraction is significantly higher in the lAcbSh than in more dorsal subregions (relative to lAcbSh: $p = 0.339, 0.285, 0.473, 0.274, 0.071, 0.038, 0.001$ for OT, VP, mAcbSh, core, VMS, VLS, and DLS, respectively; $p < 0.001$ for DMS). **g**, ANN schematic. Single-trial spike counts from the distribution-coding subpopulation **a** were linearly mapped into 16 dimensions by the trainable matrix *W* and then fed through the network (see Methods). After a final layer, a softmax function transformed activations into a properly-normalized probability distribution, whose 1-Wasserstein distance to ground truth was minimized with stochastic gradient descent. **h**, Example decoded distributions from the test set, shown as line plots to distinguish individual pseudo-trials. **i**, Wasserstein distance relative to reference for the ANN trained on all six trial types, with and without shuffling odour-distribution mappings ($p < 0.001$ ordered vs. shuffled; $p < 0.001$ ordered relative to chance value of 1; $p = 0.350$ shuffled relative to chance value of 1). **j**, Same as **i**, but for ANN trained on only the rewarded odours, which shared the same mean ($p < 0.001$ ordered vs. shuffled, ordered relative to chance value of 1, and shuffled relative to chance value of 1). **k**, Schematic depicting setup for transfer analysis. Four trial types, including both Nothing odours, were used for training (green background), and the other two were used for testing (orange background). Matched pairings veridically assigned odours to distributions, while mismatched pairings used either only Fixed or only Variable odours for training while assigning one member per training pair and one member per testing pair to the opposite distribution (indicated by the exclamation mark). There were four possible ways to draw the matched dichotomies, all of which are shown (rows). For the mismatched dichotomies, the distributions (Fixed or Variable) could be arbitrarily assigned to both pairs of red and blue odors, and then either red or blue could be assigned to the training versus test set, so only four of the eight total possibilities are shown. **l**, Wasserstein distance relative to reference for standard (mean ± s.e.m. = 0.128 ± 0.019), matched (0.217 ± 0.032), and mismatched (1.028 ± 0.123) settings. Standard is identical to analysis shown in **c**, except that for this decoder, neurons from all mice were pooled. Matched transfer yields distributions that are nearly as accurate as training with all six trial types ($p < 0.001$ for matched vs. mismatched and standard vs. mismatched, Student's *t*-test for independent samples; $p = 0.043$ for standard vs. matched, Student's *t*-test for independent samples; $p < 0.001$ for standard and matched relative to chance value of 1, one-sample Student's *t*-test; $p = 0.836$ for mismatched relative to chance value of 1, one-sample Student's *t*-test).

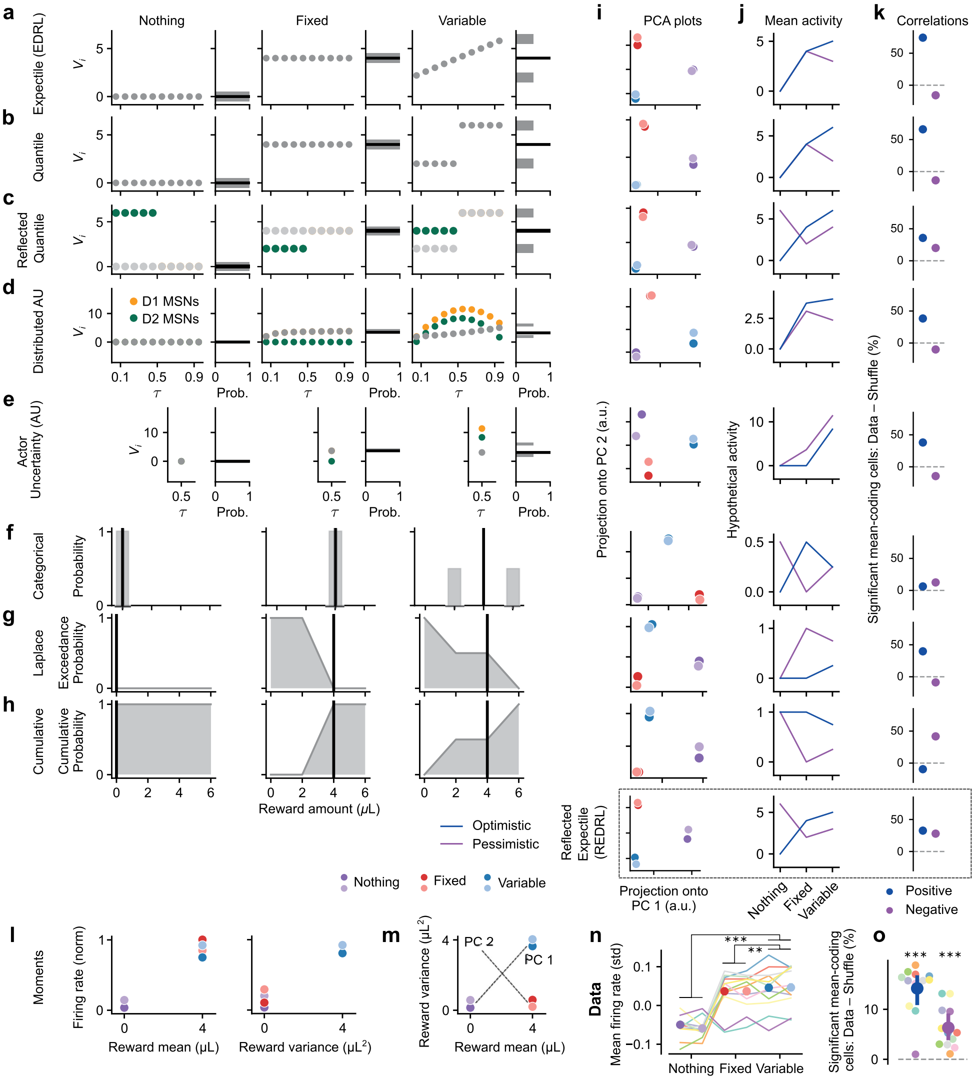


Extended Data Fig. 5 | See next page for caption.

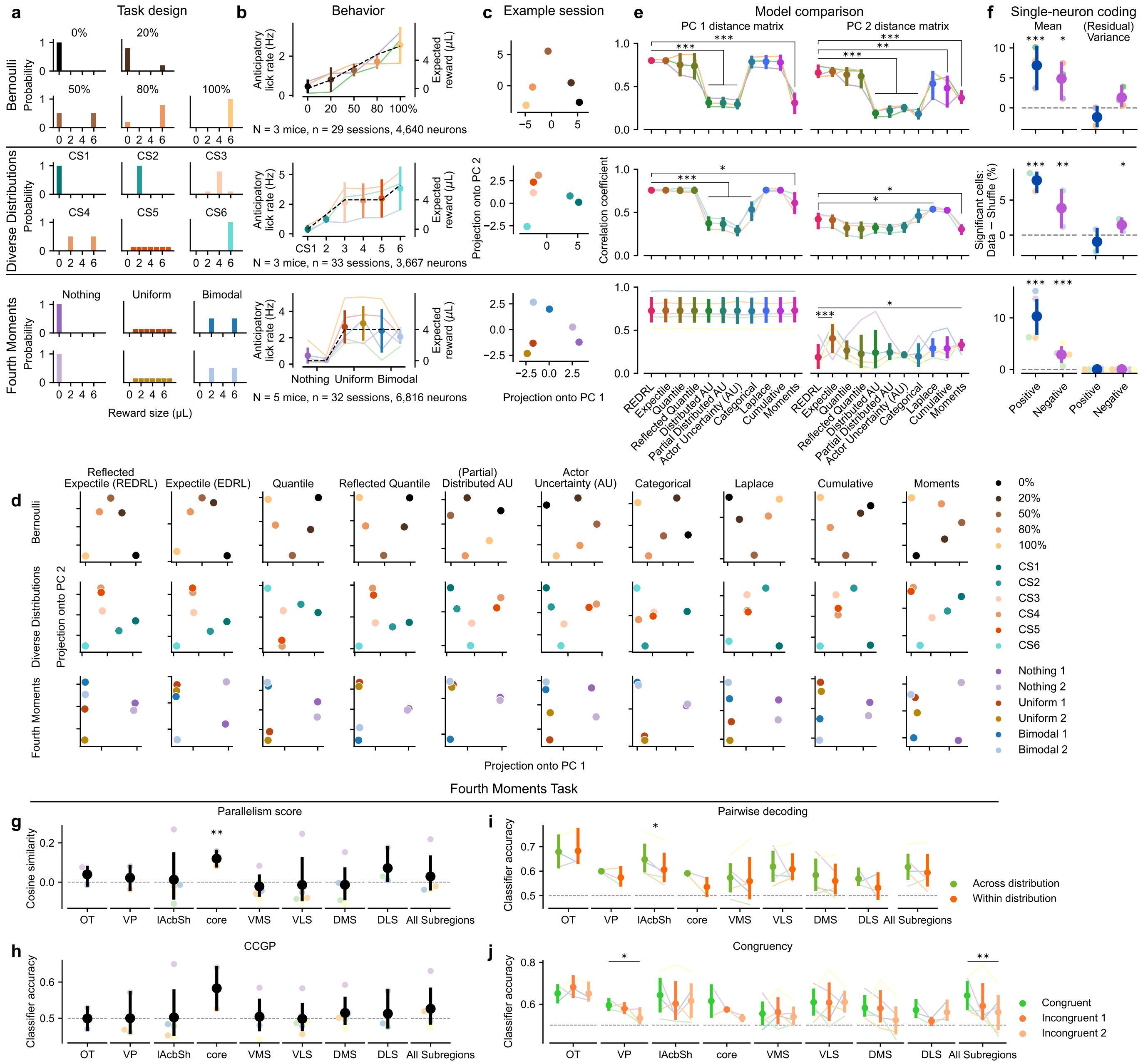
Extended Data Fig. 5 | A generalized linear model (GLM) to examine trial history, reward, reward prediction, and motor encoding in the striatum. **a**, Schematic illustrating the design of the GLM (see Methods). Briefly, trial-length regressors (time in trial and trial history) were broken up into 7 raised cosine basis functions tiling the 6 seconds of each (odour-cued) trial. Reward, reward prediction, and sensory regressors were time-locked to reward or odour onset and then convolved with a logarithmically-scaled raised cosine basis[112]. Licking, whisking, and running regressors were convolved with the same basis in both the forward and reverse directions. Pupil area and face motion SVDs from Facemap were input directly to the model without convolving. The Poisson GLM computes the sum of the regressors weighted by their fitted coefficients, passes this through an exponential nonlinearity, and uses this rate to predict spike counts in 20 ms bins. **b**, *Top*, example regressor matrix for 10 test trials. Each row corresponds to a different predictor, binned on the left by regressor type (rectangles) and group (colour). Rectangles on top demarcate different trials, coloured by trial type. *Middle*, empirical spike counts in each bin for an example neuron. *Bottom*, smoothed empirical firing rate (black) and model prediction (pink) for the trials shown. Deviance statistics in every panel of this figure rely on a held-out test set (never used during cross-validation), after zeroing out the contribution of electrode drift. **c**, Histogram of fraction deviance explained for all neurons. **d**, Fraction deviance explained as a function of striatal subregion (relative to DLS: $p < 0.001$ for OT, VP, lAcbSh, and core; $p = 0.490, 0.608, 0.054$ for VMS, VLS, and DMS, respectively). For these analyses, mAcbSh was omitted due to lack of neurons/animals. **e**, Difference in fraction deviance explained between the full model and reduced models in which trial history (*top row*), reward (*second row*), sensory and reward-prediction (*third row*), or motor (*bottom row*) regressors were excluded before re-fitting. **f**, Kernel strength (see Methods) of trial history (*top*), reward (*second*) expectile (*third*), and motor (*bottom*) regressors. **g**, As in **e**, but showing the difference in fraction deviance explained as a function of striatal subregion. (History, relative to DLS: $p = 0.124$ for DMS; $p < 0.001$ for all other subregions; Reward, relative to DLS: $p = 0.009, 0.141$, and 0.441 for OT, VP, and DMS, respectively; $p < 0.001$ for all other subregions; Expectiles, relative to DLS: $p = 0.234$ for DMS; $p < 0.001$ for all other subregions; Motor, relative to DLS: $p < 0.001$ for all subregions). **h**, As in **f**, but showing the kernel strength computed on the full model as a function of striatal subregion. (History, relative to DLS: $p < 0.001$ for OT, VP, and VLS; $p = 0.042, 0.288, 0.023$, and 0.926 for lAcbSh, core, VMS, and DMS, respectively; Reward, relative to DLS: 0.148, 0.004, 0.172 for VP, core, and DMS; $p < 0.001$ for all other subregions; Expectiles, relative to DLS: $p < 0.001$ for OT, VP, lAcbSh, VMS, and VLS; $p = 0.285$ and 0.014 for core and DMS, respectively; Motor, relative to DLS: $p = 0.004$ for DMS; $p < 0.001$ for all other subregions). **i**, Pearson correlation (across-neurons, within-sessions) of difference in deviance explained between reduced models. Holding out trial history, reward, or expectiles tends to similarly affect deviance for a given neuron, while being uncorrelated with motor behavior. Small dots, individual sessions; medium dots, mean across sessions within animals; large dots, mean ± 95% c.i. across mice. (Drop History vs. Drop Reward, Drop History vs. Drop Expectiles, and Drop Reward vs. Drop Expectiles, $p < 0.001$ for all subregions; Drop Motor vs. Drop History, $p = 0.644, 0.479, 0.993, 0.428, 0.133, 0.148, 0.674, 0.986$ for OT, VP, lAcbSh, core, VMS, VLS, DMS, and DLS respectively; Drop Motor vs. Drop Reward, $p = 0.626, 0.981, 0.134, 0.596, 0.473, 0.028, 0.745, 0.498$; Drop Motor vs. Drop Expectiles, $p = 0.331, 0.816, 0.796, 0.681, 0.193, 0.603, 0.148, 0.554$).



Extended Data Fig. 6 | Striatal activity patterns are inconsistent with sampling-based codes. a, Illustration of how the mean-matched Fano factor was computed[115]. The mean and variance (across trials) of the spike count for a single neuron contributed one data point to the scatter plot. Grey dots depict all neurons from an example session, time bin (here, centered 200 ms after odour onset), and odour (here, Variable 2). The grey line is the regression fit to all data, constrained to pass through zero and weighted according to the estimated s.e.m. of each variance measurement. Black dots are the data points preserved by mean matching at each time point, to eliminate the possibility that differences across time are driven by differences in firing rates, which could in principle violate the Poisson assumption. This transforms the distribution of mean counts from the grey to the black distribution. The regression slope for the mean matched data is plotted as the black line. Finally, the Poisson expectation of equal mean and variance is plotted in orange, with a slope of one. This procedure was performed independently on each session, time bin, and trial type. **b**, Time course of the computed mean-matched Fano factor (±95% c.i.) for the example session shown in **a**. That is, the slope of black line in **a** is the height of the light blue, Variable 2 line in **b** 200 ms after CS onset. **c**, Quantification of mean matched Fano factor across second-long time periods. Consistent with cortical observations[115], we see a quenching of variability upon CS onset (baseline: $p = 0.002$, 0.001, <0.001, <0.001 relative to odour, early trace, late trace, and outcome periods), and another one upon reward delivery (reward: $p < 0.001$, = 0.002, 0.006, 0.053 for baseline, odour, early, and late trace periods). **d**, Quantification of mean matched Fano factor across trial types, shown separately for each time period. In general, there is no tendency for Variable odours to elicit strong and sustained increases in variability, as would be predicted by sampling-based codes[129] (baseline, odour, early and late trace: all p's > 0.05, except Nothing 1 vs. Variable 1 for odour: $p = 0.032$ uncorrected). However, reward delivery specifically drives yet another decrease in variability during the outcome period (Nothing 1: $p = 0.570$ for Nothing 2; $p < 0.001$ for Fixed odours; $p = 0.002$ for Variable odours).

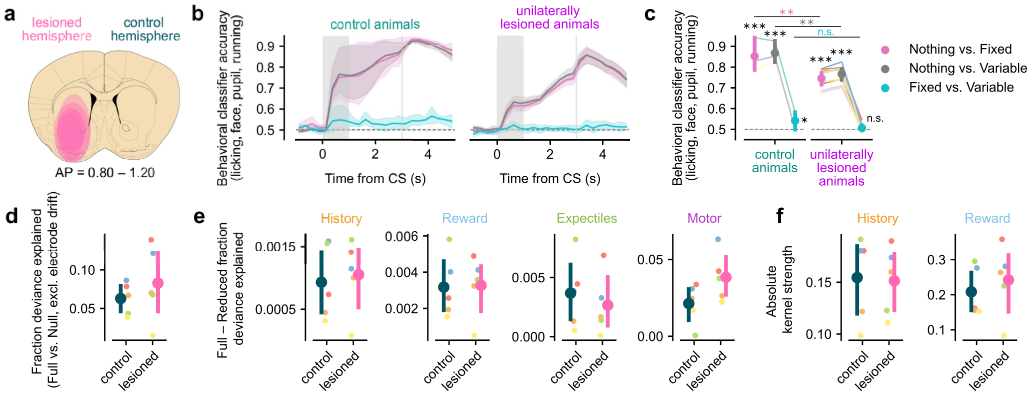


Extended Data Fig. 7 | Additional detail for distributional model comparisons. a, Schematic showing converged expectile code for each distribution (Nothing, Fixed, and Variable) learned by EDRL, as in Fig. 2d. The activation of each value predictor is shown as a function of τ, the level of pessimism or optimism. Together, they encompass the complete reward distribution. **b**, Same as **a**, but for quantiles rather than expectiles. **c**, Same as **b**, but for a reflected quantile code in which pessimistic (D2, green) neurons correlate negatively with V_i (grey). Optimistic (D1, yellow) neurons are identical to V_i, as in REDRL. **d**, Same as **a**, but showing the converged value predictors for the Distributed Actor Uncertainty model[123]. In it, D1 and D2 MSNs learn exclusively from positive and negative RPEs, respectively, such that their difference at each level of τ (grey dots) approximates each expectile, and their sum relates to the spread of the distribution. This drives maximal activity in response to Variable odours, which is why they separate out most clearly along PC 1. **e**, Same as **d**, but for a reduced version in which only a single pair of value predictors are learned with balanced positive and negative learning rates[66] ($\tau = 0.5$). **f**, Same as **a**, but for a categorical code in which distributions are encoded as a histogram[33]. Each neuron is imagined to correspond to a single reward bin, with its firing rate proportional to the height of that bin. **g**, Same as **f**, but for a Laplace code[40]. In the limit of infinitely steep reward sensitivities for the teaching signal, these value predictors converge to the probability that the reward delivered exceeds some threshold reward amount, the "exceedance probability". This is simply 1 minus the CDF of the probability distribution in question. Neural activities are taken to be proportional to this 1 − CDF value. **h**, Same as **g**, but for a population of neurons that flips the encoding, and so is directly proportional to the CDF. **i-k**, Qualitative features of each code in **a**–**h** plus random noise. REDRL predictions are included in the box on the last line, for comparison. **i**, PCA projection for each code. Only quantile-like codes give rise to the pattern observed in the data. **j**, Hypothetical activity in response to each distribution, averaged separately over optimistic (blue) and pessimistic (purple) predictors for each code type. Only the reflected codes and AU model predict a noticeable uptick in Variable relative to Fixed odours. **k**, Percentage of simulated predictors that significantly correlate with mean reward either positively (blue) or negatively (purple) for each code type. Only the reflected and categorical codes have a substantial fraction of both types of cells. In practice the positive-coding predictors are optimistic and the negative-coding predictors are pessimistic. **l**, A hypothetical "distributional" code in which each neuron's firing rate linearly correlates with either reward mean (*left*) or variance (*right*). **m**, Each trial type, replotted in mean–variance space. From this picture, it is clear that for this particular set of reward distributions, Fixed odours will be located at the midpoint between Nothing and Variable odours along PC 1, though altering the ratio of mean- to variance-coding neurons will move Fixed odours left or right along PC 1. Different sets of reward distributions could lead to different geometries. **n**, Mean z-scored firing rates for each neuron, in addition to being higher for rewarded than unrewarded odours ($p < 0.001$), were also higher for Variable than for Fixed odours ($p = 0.006$), as assessed by an LME with neuron level observations, averaged over trials, and session-level random effects nested within mouse. **o**, Same as Extended Data Fig. 2o, but for mean. Fraction is higher than chance for both positive- and negative-coding cells (both p's < 0.001).

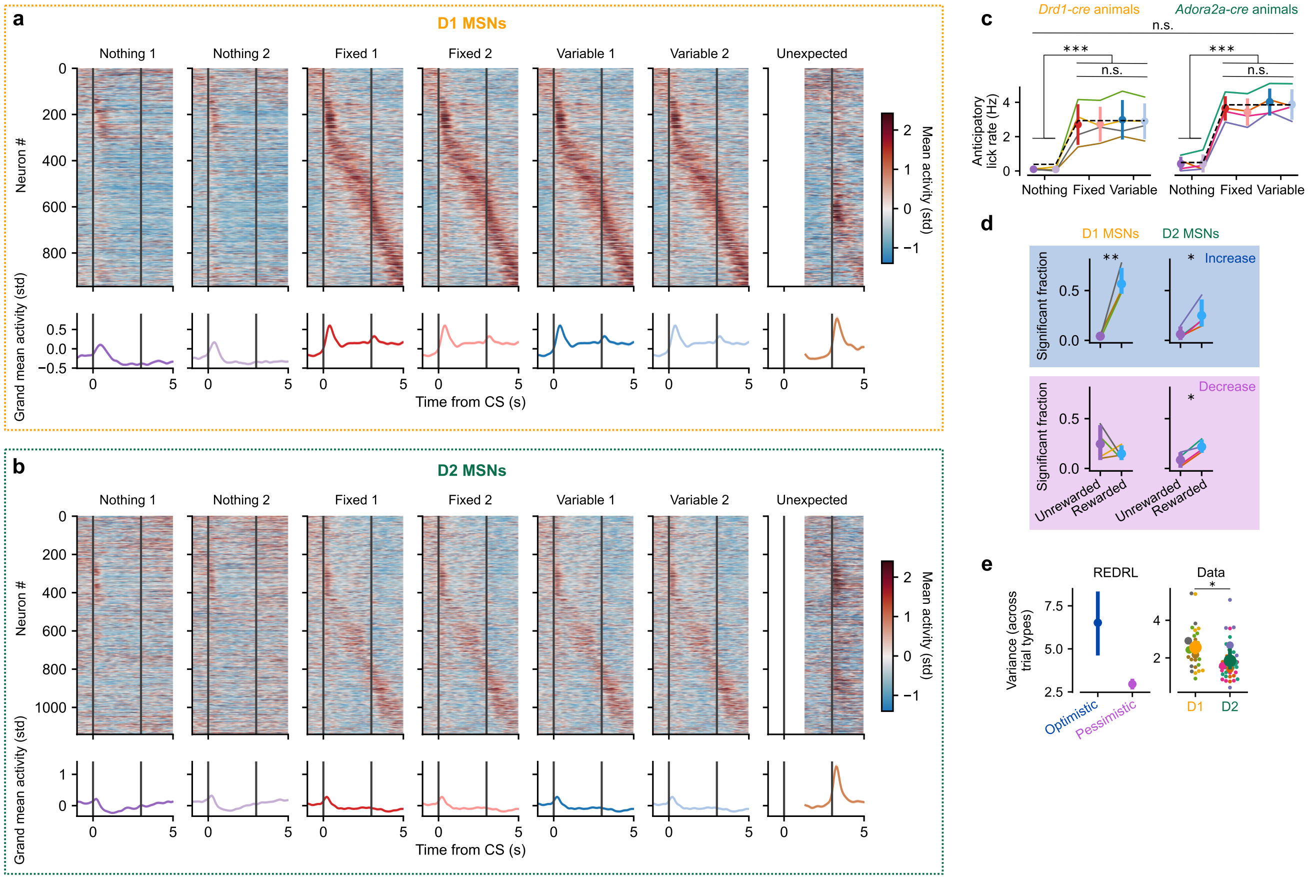


Extended Data Fig. 8 | See next page for caption.

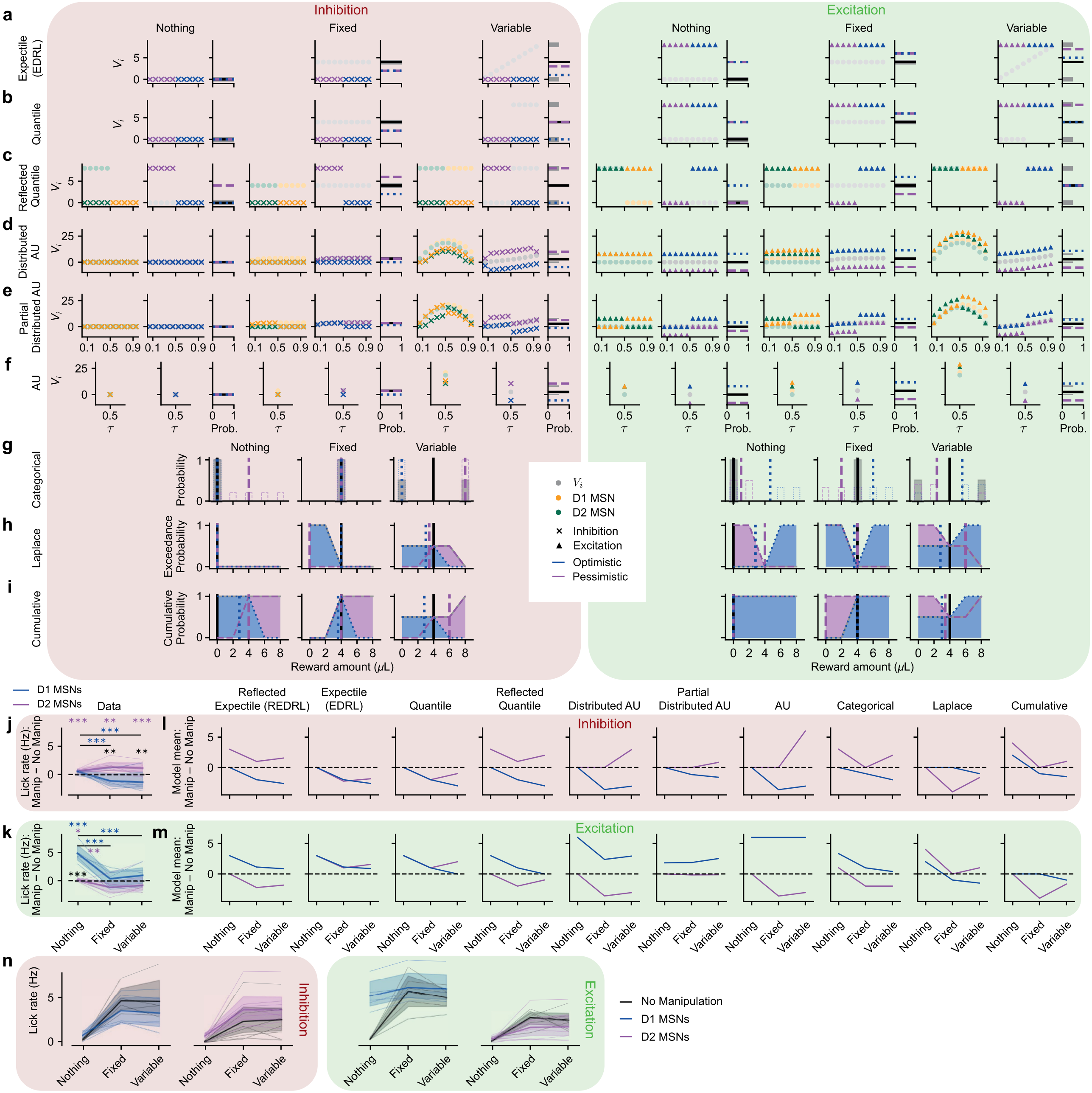
Extended Data Fig. 8 | REDRL consistently predicts population responses across three additional classical conditioning tasks. a, Reward distributions for the Bernoulli (*top*), Diverse Distributions (*middle*), and Fourth Moments (*bottom*) tasks. **b**, Anticipatory lick rate during the late trace period for each task and trial type. (Bernoulli task: 0%, $p < 0.001$ versus 50, 80, and 100%; 20%, $p < 0.001$ versus 80 and 100%; 50%, $p < 0.001$ versus 100%; 80%, $p = 0.008$ versus 100%. Diverse Distributions task: CS 1, $p = 0.008$ versus CS 2, $p < 0.001$ versus CS 3–6; CS 2, $p < 0.001$ versus CS 3–6; CS 3, $p = 0.560$, 0.243, <0.001 versus CS 4–6, respectively; CS 4, $p = 0.560$, 0.001 versus CS 5–6, respectively; CS 5, $p = 0.009$ versus CS 6. Fourth Moments task: Nothing 1 or Nothing 2, $p < 0.001$ versus Uniform 1, Uniform 2, Bimodal 1, and Bimodal 2; Uniform 1, $p = 0.570$, 0.336, <0.001 versus Uniform 2, Bimodal 1, and Bimodal 2, respectively; Uniform 2, $p = 0.126$, <0.001 versus Bimodal 1 and Bimodal 2, respectively; Bimodal 1, $p = 0.016$ versus Bimodal 2). Dashed line indicates mean reward for that trial, given on the secondary *y*-axis. **c**, 2D PC projections for example sessions in each task. **d**, 2D PC projections for each model on each of the three tasks. **e**, Quantification of Pearson correlation between the Euclidean distance matrices measured between each trial type along either PC 1 (*left*) or PC 2 (*right*). (Bernoulli task: PC 1 relative to REDRL, $p = 0.994$, 0.459, 0.284, <0.001, <0.001, <0.001, 0.861, 0.888, 0.772, <0.001 for Expectile, Quantile, Reflected Quantile, Distributed AU, Partial Distributed AU, AU, Categorical, Laplace, Cumulative, and Moments codes, respectively; PC 2 relative to REDRL, $p = 0.666$, 0.964, 0.653, <0.001, <0.001, <0.001, <0.001, 0.078, 0.002, <0.001. Diverse Distributions task: PC 1 relative to REDRL, $p = 0.999$, 0.963, 0.985, <0.001, <0.001, <0.001, <0.001, 0.993, 0.994, 0.011; PC 2 relative to REDRL, $p = 0.863$, 0.077, 0.050, 0.096, 0.054, 0.147, 0.428, 0.038, 0.065, 0.047. Fourth Moments task: PC 1 relative to REDRL, $p = 0.891$, 0.990, 0.997, 0.951, 0.928, 0.978, 0.828, 0.984, 0.927, 0.921; PC 2 relative to REDRL, $p < 0.001$, 0.127, 0.325, 0.167, 0.305, 0.891, 0.839, 0.075, 0.060, 0.021). **f**, Difference between observed and trial-type shuffled data in the percentage of cells significantly correlating positively or negatively during the late trace period with either mean (*left*) or residual variance (*right*). In the Bernoulli task, mean and variance are orthogonal by design, so residual variance is equivalent to variance. In the Fourth Moments task, mean and variance are fully colinear, so residual variance is always equal to zero. (Bernoulli task: $p < 0.001$, = 0.013, 0.112, 0.225 for Positive and Negative mean and residual variance differences relative to zero, respectively. Diverse Distributions task: $p < 0.001$, = 0.009, 0.312, 0.026. Fourth Moments task: both mean p's < 0.001). **g**, Pseudo-population parallelism score across subregions in the Fourth Moments task, comparing neural representations of Uniform and Bimodal distributions (relative to chance level of 0: $p = 0.291$, 0.150, 0.851, 0.002, 0.465, 0.832, 0.775, 0.175, 0.548 for OT, VP, lAcbSh, core, VMS, VLS, DMS, DLS, and All Subregions, respectively. Same order applies to remaining panels in this figure). Pseudo-populations were constructed as in Extended Data Fig. 2l, and mAcbSh was excluded because of too few neurons in all animals. **h**, Same as **g**, but for CCGP (relative to chance level of 0.5: $p = 0.975$, 0.997, 0.948, 0.150, 0.852, 0.945, 0.474, 0.693, 0.337). **i**, Same as **g**, but for pairwise decoding (Across- vs. Within-distribution: $p = 0.893$, 0.411, 0.012, 0.184, 0.590, 0.762, 0.256, 0.327, 0.311). **j**, Same as **g**, but for congruency analysis (Congruent vs. Incongruent 1: $p = 0.457$, 0.411, 0.333, 0.606, 0.833, 0.966, 0.956, 0.106, 0.225; Congruent vs. Incongruent 2: $p = 0.993$, 0.014, 0.265, 0.228, 0.602, 0.978, 0.073, 0.760, 0.007).



Extended Data Fig. 9 | Additional data for 6-OHDA experiments. **a**, Consensus heat map[74] of all five animals' lesion locations. 6-OHDA was injected in the lAcbSh but diffused into the VLS, so we considered both regions to be lesioned. We excluded OT, despite the fact that it was often lesioned, because it is not physically contiguous and showed weaker evidence of distributional coding in control animals. The illustration was adapted from ref. 74, Elsevier. **b**, Behavioral decoding analysis comparing fully intact animals ($N = 3$) and unilaterally lesioned ($N = 9$) animals across time. For this analysis, animals were considered lesioned if they had received any 6-OHDA injection, even if that hemisphere was never recorded or was mistargeted relative to Neuropixels recording location. **c**, Quantification of behavioral classifier accuracy during the late trace period. While across-mean behavioral decoding was stronger in the control than the lesioned animals (effect of lesion: $p = 0.006$, 0.001, 0.173 for Nothing vs. Fixed, Nothing vs. Variable, and Fixed vs. Variable, respectively), both groups of animals clearly learned the task and had above-chance across-mean decoding ($p < 0.001$ compared to chance level of 50% for both Nothing vs. Fixed and Nothing vs. Variable in control as well as lesioned animals). Interestingly, Fixed vs. Variable classification was also weakly significant ($p = 0.032$ relative to chance level of 50%) for fully intact control animals, providing behavioral evidence that they did in fact learn this distinction. **d**, Median fraction deviance explained by the GLM (Extended Data Fig. 5) for neurons in control vs. lesioned hemispheres ($p = 0.831$). **e**, Difference in fraction deviance explained between full model and models in which history (*left*; $p = 0.474$), reward (*second*; $p = 0.623$) sensory/reward prediction (*third*; $p = 0.861$) or motor (*right*; $p = 0.618$) regressors had been dropped out. **f**, Absolute kernel strength of history (*left*; $p = 0.634$), reward (*second*; $p = 0.089$), expectiles (*third*; $p = 0.448$) or motor (*right*; $p = 0.145$) regressors.



Extended Data Fig. 10 | Additional data for two-photon calcium imaging experiments. a, D1 MSN activity. *Top*, heatmaps showing average z-scored deconvolved calcium activity in response to each odour for each neuron, as in Extended Data Fig. 2b. *Bottom*, grand average z-scored deconvolved calcium activity across all neurons. **b**, Same as **a**, but for D2 MSN activity. **c**, Anticipatory lick rates for each trial type, computed during the late trace period separately for *Drd1-cre* and *Adora2a-cre* animals (in which we imaged D1 or D2 MSNs, respectively). *(Drd1-cre*, Nothing 1 or Nothing 2: $p < 0.001$ versus Fixed 1, Fixed 2, Variable 1, and Variable 2; *Drd1-cre*, Fixed 1: $p = 0.960$, 0.458, 0.642 versus Fixed 2, Variable 1, and Variable 2, respectively; $N = 4$ mice, 29 sessions. *Adora2a-cre*, Nothing 1 or Nothing 2: $p < 0.001$ versus Fixed 1, Fixed 2, Variable 1, and Variable 2; *Adora2a-cre*, Fixed 1: $p = 0.790$, 0.608, 0.686 versus Fixed 2, Variable 1, and Variable 2, respectively; $N = 4$ mice, 41 sessions. Main effect of genotype, relative to Nothing 1: $p = 0.785$; interaction of genotype and trial type: $p = 0.888$, 0.387, 0.525, 0.350, 0.331 for Nothing 2, Fixed 1, Fixed 2, Variable 1, and Variable 2, respectively; $N = 8$ mice, 70 sessions). As in Fig. 1c, dashed lines indicate mean reward for that trial type. **d**, Fraction of neurons whose late trace activity increased (*top*) or decreased (*bottom*) relative to baseline, shown separately for D1 (*left*) and D2 (*right*) MSNs and unrewarded (Nothing) versus rewarded (Fixed and Variable) odours (x-axis); these trial types were pooled before analysis. As expected, a larger fraction of D1 MSNs increases to rewarded rather than unrewarded odours ($p = 0.006$; mean ± s.e.m. = 0.524 ± 0.074), while there is no difference in the fractions that decrease ($p = 0.423$; mean ± s.e.m. = −0.098 ± 0.106). Meanwhile, for D2 MSNs, a significantly greater fraction of neurons change their activity on rewarded compared to unrewarded trials, by either increasing ($p = 0.022$; mean ± s.e.m. = 0.189 ± 0.043) or decreasing ($p = 0.016$; mean ± s.e.m. = 0.133 ± 0.027) their activity relative to baseline. Asterisks and p-values report the results of paired samples Student's t-tests on rewarded vs. unrewarded fractions across mice. **e**, REDRL predicts higher variance across trial types for optimistic than for pessimistic reward predictors on average (*left*), which is also true in the two-photon data for D1 and D2 MSNs, respectively (*right*). Small dots are averages within sessions, medium dots are averages within mice, and large dots with error bars show averages ± 95% c.i. across mice ($p = 0.017$ for effect of genotype).



Extended Data Fig. 11 | See next page for caption.

Extended Data Fig. 11 | Additional detail for distributional model manipulations. a, Schematic showing how optogenetic perturbations were simulated for an expectile code (from EDRL). Optimistic (blue) or pessimistic (purple) predictors were shifted from their original values (semi-transparent grey circles) and clamped to low or high values to mimic inhibition (*left*, "x"s) or excitation (*right*, triangles), respectively. Panels on the right depict the ground-truth reward distribution, its mean (black line), and the means of the manipulated sets of value predictors (blue or purple dashed lines). **b**, Same as **a**, but for a quantile rather than expectile code. **c**, Same as **b**, but for a reflected quantile code. The additional, leftmost panel for each distribution depicts the activity of D1 (yellow) and D2 (green) MSNs at baseline (semi-transparent circles) and after manipulations (opaque "x"s and triangles). These are what are directly clamped by the simulated optogenetic inhibition or excitation. As a result, the effect on the implied value predictors (middle panel) corresponding to D2 MSNs are of opposite sign, as is the change in predicted mean (right panel). **d**, Same as **c**, but for the Distributed Actor Uncertainty (AU) model. Since D1 and D2 MSN activities in this model can exceed the maximum reward value, the left panel shows that perturbations were simulated by adding or subtracting a fixed amount from each activity level (opaque "x"s and triangles) relative to baseline (semi-transparent circles). The middle panel plots the resulting value predictors, computed as the pointwise differences between D1 and D2 MSN activities, for pessimistic (purple) and optimistic (blue) manipulations in comparison to baseline (grey semi-transparent circles). **e**, Same as **d**, except that only the optimistic or pessimistic half of MSNs were manipulated to simulate perturbations of D1 or D2 MSNs, respectively. **f**, Same as **d**, except for the original Actor Uncertainty (AU) model in which there is only one pair of value predictors with balanced learning rates ($\tau = 0.5$). **g**, Schematic showing how optogenetic perturbations were simulated for a categorical code (from CDRL), which effectively represents the reward distribution using a histogram. Pessimistic (0, 2 μL; purple) or optimistic (6, 8 μL; blue) bins were clamped to 0 or 1 to simulate inhibition or excitation, respectively, relative to baseline (grey). The resulting distributions were normalized to sum to one (see Methods). Dashed vertical lines show the means of the ground-truth (black) and manipulated distributions. **h**, Same as **g**, except for a Laplace code[40] in which each neuron corresponds to the height of 1 − CDF at a particular point. While the baseline case is always monotonically decreasing, simulated excitation or inhibition can change this. Means were computed by differentiating and then normalizing (see Methods). **i**, Same as **h**, except for a cumulative code where each neuron corresponds to the height of the CDF at a particular point. **j**, Actual differences in lick rate during the last half second of the trace period in response to inhibition of D1 or D2 MSNs, copied from Fig. 5f. **k**, Same as **j**, but for excitation. **l**, Predicted difference in mean reward due to inhibition for REDRL and each of the alternative models in **a**–**i**. **m**, Same as **l**, but for excitation. **n**, Average lick rates in each group of animals, with (blue and purple) or without (black) manipulations, rarely exceeded 5 Hz.

nature portfolio

Corresponding author(s): Naoshige Uchida

Last updated by author(s): Nov 20, 2024

Reporting Summary

Nature Portfolio 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 Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

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

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	The experimental apparatus was controlled using Bpod software (https://sanworks.github.io/Bpod_Wiki/) written in MATLAB (MathWorks). Electrophysiological data were collected with SpikeGLX (https://billkarsh.github.io/SpikeGLX/), clustered offline using Kilosort3 (https://github.com/MouseLand/Kilosort), and manually curated in Phy (https://github.com/cortex-lab/phy). 2P data were collected with ScanImage (https://docs.scanimage.org/index.html) and processed using suite2p (https://github.com/MouseLand/suite2p). Behavioral videos were collected with bonsai (https://bonsai-rx.org/) and processed using Facemap v. 0.2 (https://github.com/MouseLand/facemap).
Data analysis	Analysis was performed using custom-written code in Python, which is available on GitHub (https://github.com/alowet/distributionalRL) and documented on Zenodo (https://doi.org/10.5281/zenodo.14183769).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Processed data (odor-onset-aligned spiking, deconvolved calcium activity, and behavioral variables) is available on Dryad (https://doi.org/10.5061/dryad.80gb5mm0m).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	This study does not involve human participants, their data or biological materials
Reporting on race, ethnicity, or other socially relevant groupings	This study does not involve human participants, their data or biological materials
Population characteristics	This study does not involve human participants, their data or biological materials
Recruitment	This study does not involve human participants, their data or biological materials
Ethics oversight	This study does not involve human participants, their data or biological materials

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

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	Experimental data collection was deigned to achieve maximal statistical power within the available time and financial resources to perform experimental data collection.
Data exclusions	Sessions with insignificant behavior ($p > 0.05$, Mann-Whitney U test between Rewarded and Unrewarded trials) were not analyzed. Unstable neurons and neurons with a firing rate < 0.1 Hz were excluded from further analysis, as described in the Methods. We treated neurons as "lesioned" only if we could confirm that they fell within the TH-depleted region in histology, otherwise all recording sessions from the lesioned hemisphere were excluded
Replication	Data was collected from many animals (23 for electrophysiology, 8 for two-photon imaging, and 25 for optogenetics). Statistical analyses took into account the hierarchical structure of the data (many sessions per animal, collected from many animals)
Randomization	Odor-distribution mappings were randomly assigned. We counterbalanced the order in which we recorded from control and lesioned hemispheres.
Blinding	Investigators were not blinded to the identity of the mice. This was required for animal care considerations and the study did not involve controls that required blinding. Analysis was performed after data collection so investigators were blind to results during data collection.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

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

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	A total of 56 adult C57BL/6J (Jackson Laboratory) male and female mice were used in these experiments. Mice were backcrossed for over 5 generations with C57/BL6J mice.
Wild animals	No wild animals were used in this study.
Reporting on sex	We report the sex ratios of all animals used in all experiments. We never analyzed sex as a variable due to lack of statistical power.
Field-collected samples	No field-collected samples were used in this study.
Ethics oversight	All procedures were performed in accordance with the National Institutes of Health Guide for the Care and Use of Laboratory Animals and approved by the Harvard Animal Care and Use Committee.

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

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*

nature portfolio

https://doi.org/10.1038/s41586-024-08488-5

Supplementary information

An opponent striatal circuit for distributional reinforcement learning

In the format provided by the
authors and unedited

Supplementary Discussion

We would like to offer three extensions and points of clarification to complement our discussion in the Main Text of the paper.

Expectiles versus quantiles

Given the high degree of similarity between REDRL and a reflected quantile code (RQDRL), it is natural to wonder whether some Pavlovian conditioning experiment might be able to disentangle these two code types. We believe the answer is yes, although this is most easily done at the level of dopamine neurons, rather than behavior or striatal neurons.

Beginning with behavior, the fundamental problem is that the method used to learn and represent a probability distribution in distributional RL is distinct from the mechanism of action selection. In other words, the underlying distributional RL algorithm does not put any hard constraints on the policy the agent uses for action selection. Most often, agents are assumed to simply take the action with the highest expected value (thereby ignoring distributional information at decision time). This is true regardless of whether the expected value is represented explicitly (as it would be in an expectile code with $\tau = 0.5$) or must be computed from the underlying probability distribution (as in a quantile code). However, other options are possible; for example, one can include an exploration bonus proportional to the left truncated variance of the return distribution and have this bonus decay with training time, a choice that actually improves performance at test time[130]. While the details of any particular proposal are not especially important here, the overarching principle is that the format of the distribution (and the algorithm used to learn it, like REDRL or RQDRL) leave the question of behavioral choice unresolved.

Furthermore, provided that the behavioral policy is computed on the basis of the encoded probability distribution, a particular policy will lead to the same choices, irrespective of whether REDRL or RQDRL is used to learn and represent this distribution. (By contrast, a policy that is computed directly on the basis of the encoded statistics — not the full distribution — could in principle differ for expectiles and quantiles.) This is because, with enough neurons (and so expectiles or quantiles), one can represent any distribution with high fidelity. From this perspective, even if the policy is known, behavior would only allow one to distinguish between different shapes of the encoded distribution, not between different formats with which this distribution is encoded in neural activity.

At the level of the striatum, the predictions for expectile-like and quantile-like codes are extremely similar and highly sensitive to the amount of noise used in our simulations. Intuitively, this is because the difference in "smoothness" between the code types (Fig. 3c; Extended Data Fig. 7c) is obscured by the different gains and scaling between different neurons. Nonetheless, we were able to identify some sets of highly skewed distributions for which REDRL and

RQDRL might be expected to produce different population geometries, depending on the amount of noise in the neural data and the range of τ values used. In particular, the Euclidean distance along PC 1 between Nothing odors and (mean-matched) Fixed or left-Skewed odors is predicted to differ, being greater for Fixed odors in the case of REDRL and Skewed odors in the case of RQDRL. It remains to be seen whether this could be observed in striatal data at physiological levels of noise.

Fortunately, previously-reported data from our laboratory[3] provide independent empirical reason for preferring REDRL over RQDRL. These algorithms aim to minimize an asymmetrically-weighted squared error and absolute error loss function, respectively[54]. Mathematically, stochastic gradient descent to minimize these loss functions requires that the expectile updates have a piecewise linear form, while quantile updates have a stepwise form[54]. Because dopamine responses are empirically quite smooth as a function of reward magnitude[121], we followed Dabney et al., 2020 in favoring the expectile formulation[3]. Even considering synaptic effects, the influence of dopamine on synaptic strength does not appear to be binary[13], as would be required by RQDRL.

However, we again note that a saturating version of these update rules, which converge to estimators known as Huber quantiles[126], would also be consistent with our data. Huber quantiles have an additional hyperparameter, κ, which controls the magnitude of prediction error at which the loss function switches from the expectile to the quantile version. In this way, they provide a seamless way to interpolate between REDRL and RQDRL. Estimating κ on the basis of neural data is outside the scope of the current work, but we see no reason why it should be impossible in principle, particularly by making use of the skewness manipulations described above. Thus, while we focus on REDRL in this paper for simplicity and consistency[3], this should not be construed as a rejection of a Huber quantile-based implementation of reflected distributional RL.

Alternatives to RPE accounts of dopamine

Second, the RPE theory of dopamine, of which distributional RL makes significant use, has recently come under pressure from several directions[26,27]. Most notably, Jeong et al., 2022 put forward the ANCCR model[26], which proposes that dopamine instead signals an "adjusted net contingency for causal relations." A full treatment of this paper is beyond the scope of this brief Discussion[131,132] and would be benefited by an examination of MSN (and dopamine) activity over the course of learning, which we did not undertake in the present paper. (We do, however, note that REDRL predicts that D1 MSNs will acquire positive associations faster than D2 MSNs, while D2 MSNs may be preferentially involved in later discrimination or extinction, as has been previously observed[13,14]). Instead, we would like to highlight one aspect of the ANCCR model that has received relatively little attention in the published literature, and which we perceive as a core shortcoming with respect to the results presented here.

The principal insight of ANCCR is that an agent might be interested in "retrospective associations" (roughly, the probability of cue conditional on reward) in addition to "prospective associations" (roughly, the probability of reward conditional on cue). However, the firing of dopamine neurons is well-known to depend not only on reward probability but also on reward magnitude. To account for this magnitude dependence within their ANCCR framework, Jeong et al. "adjust" their net contingencies using a "causal weight." Critically, this adjustment fully combines information about reward probability and magnitude such that they cannot be later disentangled (in this respect, it is quite similar to traditional RL). Thus, the published ANCCR model cannot straightforwardly account for our current data or previously-published results from dopamine neurons[3] in which probability and magnitude information is partially separable (in the form of expectiles), allowing one to distinguish between probability distributions with the same mean but different higher-order moments.

Probabilistic coding in the brain

Third, REDRL lends a new perspective to the coding of uncertainty in the brain. Typical treatments of this topic focus on *perceptual* uncertainty, where the observer's role is to infer the distribution of world states consistent with a pattern of neural activity[32]. While the problem is usually formulated as one of Bayesian inference[31], the associated uncertainty is frequently attributed to noisy inputs rather than ones that are genuinely ambiguous (as in the case of the Necker cube[133]). Moreover, the "causes" that the brain needs to infer are in general high-dimensional, leading many researchers to prefer sampling-based codes in these settings[51,80].

In RL, by contrast, uncertainty generally arises from a combination of state ambiguity, insufficient exploration, and intrinsic stochasticity[134], all of which complicate the problem of learning from limited experience. Fortunately, these various sources of uncertainty ultimately collapse onto a single dimension, that of reward (or more generally, the discounted future return), simplifying the representation of the probability distribution and suggesting the use of other kinds of probabilistic codes. Distributional RL excels in partitioning out intrinsic environmental uncertainty (sometimes called "aleatoric uncertainty") from other sources ("epistemic uncertainty")[135], potentially allowing for improvements in state representation[60,136], exploration[130,137–139], value estimation[140], model-based learning[141], off-policy learning[75], and risk sensitivity[142–145]. It remains to be determined whether animals make a distinction between aleatoric and epistemic uncertainty, facilitated by distributional RL, to improve exploration or offline learning in a manner similar to artificial agents.

Supplementary References

130. Mavrin, B. *et al.* Distributional reinforcement learning for efficient exploration. *arXiv [cs.LG]* (2019).

131. Qian, L. *et al.* The role of prospective contingency in the control of behavior and dopamine signals during associative learning. Preprint at *bioRxiv* https://doi.org/10.1101/2024.02.05.578961 (2024).

132. Garr, E. *et al.* Mesostriatal dopamine is sensitive to changes in specific cue-reward contingencies. *Sci Adv* **10**, eadn4203 (2024).

133. Necker, L. A. LXI. Observations on some remarkable optical phænomena seen in Switzerland; and on an optical phænomenon which occurs on viewing a figure of a crystal or geometrical solid. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **1**, 329–337 (1832).

134. Gershman, S. J. & Uchida, N. Believing in dopamine. *Nat. Rev. Neurosci.* **20**, 703–714 (2019).

135. Lockwood, O. & Si, M. A review of uncertainty for deep reinforcement learning. *AIIDE* **18**, 155–162 (2022).

136. Lyle, C., Castro, P. S. & Bellemare, M. G. A comparative analysis of expected and distributional reinforcement learning. *arXiv [cs.LG]* (2019).

137. Nikolov, N., Kirschner, J., Berkenkamp, F. & Krause, A. Information-directed exploration for deep reinforcement learning. *arXiv [cs.LG]* (2018).

138. Clements, W. R., Van Delft, B., Robaglia, B.-M., Slaoui, R. B. & Toth, S. Estimating risk and uncertainty in deep reinforcement learning. *arXiv [cs.LG]* (2019).

139. Zhang, S. & Yao, H. QUOTA: the quantile option architecture for reinforcement learning. *AAAI* **33**, 5797–5804 (2019).

140. Wang, K., Zhou, K., Wu, R., Kallus, N. & Sun, W. The benefits of being distributional: small-loss bounds for reinforcement learning. *arXiv [cs.LG]* (2023).

141. Luis, C. E., Bottero, A. G., Vinogradska, J., Berkenkamp, F. & Peters, J. Value-distributional model-based reinforcement learning. *arXiv [cs.LG]* (2023).

142. Kim, D., Lee, K. & Oh, S. Trust region-based safe distributional reinforcement learning for multiple constraints. In *Advances in Neural Information Processing Systems* (eds. Oh, A. et al.) **36**, 19908–19939 (2023).

143. Kastner, T., Erdogdu, M. A. & Farahmand, A.-M. Distributional model equivalence for risk-sensitive reinforcement learning. *arXiv [cs.LG]* (2023).

144. Cai, X.-Q. *et al.* Distributional pareto-optimal multi-objective reinforcement learning. In *Advances in Neural Information Processing Systems* (eds. Oh, A et al.) **36**, 15593–15613 (2023).

145. Rigter, M., Lacerda, B. & Hawes, N. One risk to rule them all: a risk-sensitive perspective on model-based offline reinforcement learning. *arXiv [cs.LG]* (2022).