Distributional Reinforcement Learning in the Mammalian Brain

Adam S. Lowet[*]
Molecular and Cellular Biology
Harvard University
Cambridge, MA 02138

Qiao Zheng
Neurobiology
Harvard Medical School
Boston, MA 02115

Melissa Meng
Molecular and Cellular Biology
Harvard University
Cambridge, MA 02138

Sara Matias
Molecular and Cellular Biology
Harvard University
Cambridge, MA 02138

Jan Drugowitsch
Neurobiology
Harvard Medical School
Boston, MA 02115

Naoshige Uchida
Molecular and Cellular Biology
Harvard University
Cambridge, MA 02138

Abstract

Distributional reinforcement learning (dRL) — learning to predict not just the average return but the entire probability distribution of returns — has achieved impressive performance across a wide range of benchmark machine learning tasks. In vertebrates, the basal ganglia strongly encodes mean value and has long been thought to implement RL, but little is known about whether, where, and how populations of neurons in this circuit encode information about *higher-order moments* of reward distributions. To fill this gap, we used Neuropixels probes to acutely record striatal activity from well-trained, water-restricted mice performing a classical conditioning task. Across several measures of representational distance, odors associated with the same reward distribution were encoded more similarly to one another than to odors associated with the same mean reward but different reward variance, as predicted by dRL but not traditional RL. Optogenetic manipulations and computational modeling suggested that genetically distinct populations of neurons encoded the left and right tails of these distributions. Together, these results reveal a remarkable degree of convergence between dRL and the mammalian brain and hint at further biological specializations of the same overarching algorithm.

1 Introduction

Since the firing of dopamine neurons was first suggested to resemble the reward prediction errors (RPEs) of reinforcement learning (RL) algorithms almost thirty years ago[1, 2], RL has provided a powerful theoretical framework with which to understand the basal ganglia. However, neuroscientists have struggled to connect more recent developments in machine learning, most notably the rise of deep RL, to these brain circuits. Although deep RL encompasses a wide range of approaches and insights, a major step forward came from the realization that expanding the objective function from simply the *value* — defined as the expected sum of discounted future reward, or *return* — to the *entire return distribution*, greatly improves performance across a wide range of tasks [3–5]. This technique, called "distributional reinforcement learning" (dRL), is an attractive candidate to consider in the context of the mammalian brain because (1) it can be implemented using only minor, biologically-plausible modifications to classic learning rules [6], (2) it is consistent with the observed structure of dopamine population activity [7, 8], and (3) it provides a natural mechanism to implement risk-sensitive policies, which are observed across a wide range of animal species [9–11].

[*] Corresponding author: `alowet@g.harvard.edu`

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Models of the brain's RL circuitry identify the striatum, the main input nucleus of the basal ganglia, as the site of coding for mean value [12] — or, more generally, return distributions — since it is the primary recipient of dopamine reward prediction errors (RPEs) which can modify the strength of corticostriatal synapses in a manner consistent with TD updates [13]. It is well-known that the basal ganglia circuitry is intimately involved in risk-sensitive decision-making in both healthy [14, 15] and diseased [16, 17] states, some of which has been captured by biologically-grounded computational models [18]. Nonetheless, it has proven remarkably difficult to identify the representational format and underlying algorithms by which the basal ganglia learn about reward distributions beyond the mean, with virtually all striatal recordings limited to finding strong correlations with mean value, selected actions, or reward delivery itself [19–24].

1.1 Experimental setup

Here, we harnessed the theory of dRL to approach this question in a novel way. We designed a classical conditioning task in which water-restricted mice were trained to associate neutral odors with different reward distributions, with odor assignments randomized across mice (Fig. 1a). We used three separate reward distributions: Nothing (100% chance of 0 μL reward), Fixed (100% chance of 4 μL reward) and Variable (50/50% chance of 2/6 μL reward; Fig. 1b). Because Fixed and Variable odors have the same mean but different variance, traditional RL does not distinguish between them on average, whereas dRL predicts that their representations should systematically differ. To get at whether any differences in odor representations were truly systematic, we paired each distribution with *two unique odors*, for a total of six odors. That way, we could ask whether odors associated with the same distribution were represented more similarly to one another than to odors associated with a different distribution of the same mean, as predicted by distributional but not traditional RL.

2 Results

2.1 Mice learn the task and value Fixed and Variable rewards equally

To ensure that the mice understood the task, we quantified anticipatory licking in the second that preceded reward delivery. Unsurprisingly, animals licked more to the Fixed and Variable odors than to the Nothing odors, showing that they learned the associations (Fig. 1c). Importantly, though, individual mice did not show a preference between the Fixed and Variable odors, which suggests that they valued them equally. To more rigorously rule out behavioral confounds, we analyzed not only licking but also the mice's face motion, pupil area, and running [25] and built classifiers to distinguish trial types from one another using only these behavioral observables. While we could easily decode Nothing odors from Fixed or Variable odors, we could not significantly distinguish between Fixed and Variable trials using behavior alone (Fig. 1d). This implies that any systematic neural differences between these trial types must be due to the learned associations with probability distributions and not to low-level behavior.

2.2 The first principal component of striatal activity reflects the mean

To interrogate the neural basis of a possibly distributional code, we recorded activity across a broad swath of the anterior striatum using Neuropixels probes (N=12 mice, 71 sessions, 13,997 neurons; Fig. 1e). We first verified that we could replicate previous findings of mean value coding in these regions [19–24]. Indeed, simply taking the grand average across the entire dataset (Fig. 1f) or projection onto the first principal component (PC; Fig. 1g) of z-scored firing rates revealed a strong tendency for neurons to fire more to rewarded (Fixed and Variable) than to unrewarded (Nothing) trial types.

2.3 Neurons represent information about higher-order moments of the return distribution

Unlike the observed behavior and population averages, not all neurons responded identically to Fixed and Variable odors; some neurons preferred Fixed while other preferred Variable odors (Fig. 1h). Importantly, these did not reflect idiosyncratic odor or risk preferences, as responses were consistent for both examples of the Nothing, Fixed, and Variable odors yet could differ for simultaneously-recorded neurons. To see if this was true of the population as a whole, we took activity during the

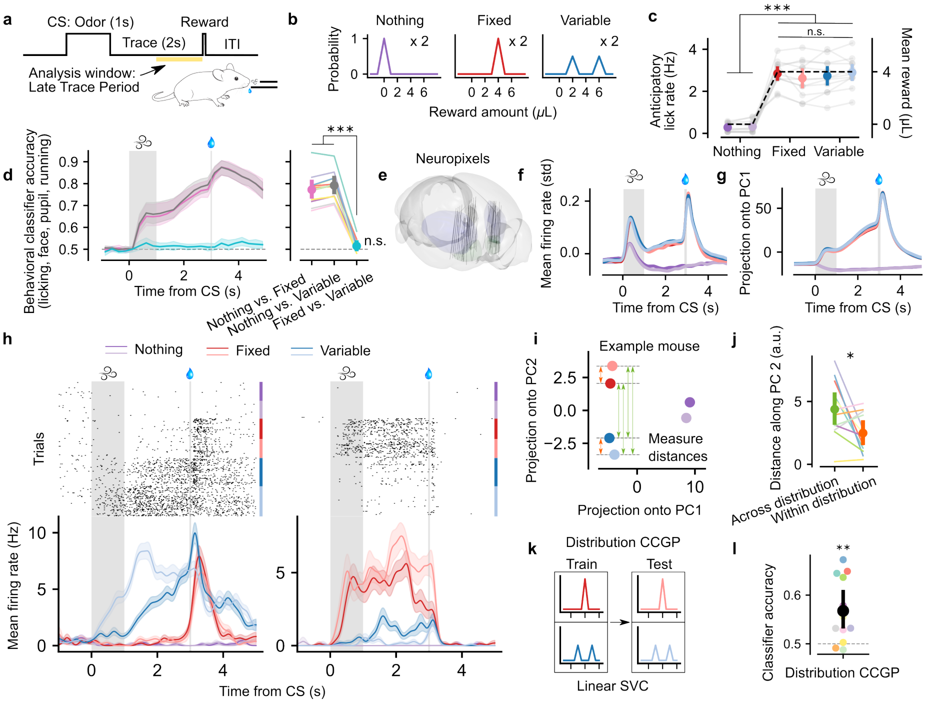

Figure 1: dRL in the striatum. a, Water-restricted, head-fixed mice were trained to associate odors with rewards. b, Probability distributions over reward amounts that were paired with odors. c, Anticipatory lick rates for each trial type. Gray lines denote individual mice. d, Accuracy of a linear classifier trained on licking, pupil area, whisking and running. Left, behavioral classifier accuracy across time. Right, quantification of behavioral classifier accuracy when trained on the entire Late Trace period. e, Reconstructed Neuropixels probe trajectories, aligned to the Allen Mouse Brain Common Coordinate Framework [27]. f, Grand average timecourse of z-scored firing rates, computed across all recorded neurons. g, Projection onto the first PC of neural activity, computed from the concatenated timecourse of average responses to each trial type. h, Raster plots (*top*) and PSTHs (*bottom*) for two simultaneously-recorded example neurons that prefer either Variable (*left*) or Fixed (*right*) odors. i, Projection of Late Trace activity into first two PCs for an example mouse. j, Distances along PC2 were greater for across distribution pairs (green arrows) than within-distribution pairs (orange arrows). k, Schematic showing an example dichotomy used for cross-condition generalization performance (CCGP) [26]. l, Average CCGP for simultaneously recorded populations. Each colored dot is the average across sessions for an individual mouse; black dot is the mean across mice. In all panels, error bars denote mean and 95% confidence intervals across mice.

Late Trace period and projected it into two-dimensional PC space independently for each mouse. PC1 again corresponded to mean value, but interestingly, PC2 seemed to separate out Fixed and Variable odors (Fig. 1i). To quantify this, we measured distances along PC2 between pairs of Rewarded odors. Across-distribution (one Fixed and one Variable) pairs were better-separated along PC2 than are within-distribution pairs, as predicted by distributional but not traditional RL (Fig. 1j).

To determine whether this distributional signature is detectable on a single-trial basis, we quantified the cross-condition generalization performance (CCGP) between different distributions with the same mean [26]. A linear decoder trained to discriminate one Fixed and one Variable odor reliably generalized to the other Fixed and Variable odors not seen during training (Fig. 1k-l). Thus, distributional coding in the striatum is factorized, allowing the same representation to be shared across multiple sensory inputs.

2.4 Opponency within the striatum may support distributional RL

The striatum consists of two principal populations of cells: dopamine receptor D1 and D2-expressing medium spiny neurons (MSNs) [28]. One challenge for biological implementations of RL has been

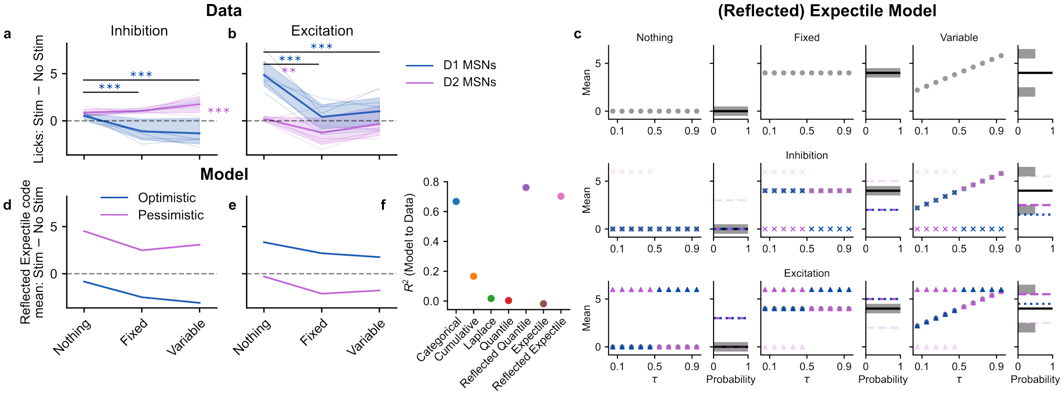

Figure 2: Opponency between D1 and D2 MSNs may support distributional RL. a, Differences in licking during the Late Trace period produced by optogenetic inhibition of D1 (blue) or D2 (purple) MSNs, relative to no stimulation. b. Same as (a), but for optogenetic activation. Asterisks with lines indicate significant differences between trial types for the given color, and asterisks on the right side indicate that all trial types of that color differed from zero. (c) Learned value predictors (*left*) and their corresponding reward distributions (*right*) in an expectile distributional RL simulation [7, 32] of the optogenetic manipulation task. Blue markers and lines show the results of optimistic ($\tau > 0.5$) perturbations and purple show pessimistic perturbations ($\tau < 0.5$). Faded markers and lines represent the "reflected" model, in which the activity of pessimistic predictors is inversely correlated with the value they convey. (d-e) Predicted mean differences in response to inhibition (a) or excitation (b) for the Reflected Expectile model in (c). (f) Opponent models (Categorical, Reflected Quantile, and Reflected Expectile) vastly outperform other models in their predictions of D1 and D2 manipulations.

how to harness these two separate populations because of their opposite plasticity rules and activity patterns. Synaptic weights onto D1 MSNs increase in response to *increases* in dopamine, while those onto D2 MSNs increase in response to *decreases* in dopamine [13, 29, 30]; analogously, D1 MSNs tend to correlate positively with expected value, while D2 MSNs correlate negatively [23, 31]. However, rather than being a bug in the RL architecture, such diversity could in principle be a feature, amplifying responses to positive or negative prediction errors and thereby biasing convergence to optimistic or pessimistic value predictors, respectively.

We therefore selectively inhibited [33, 34] or activated [35] D1 or D2 MSNs [36] in the ventral striatum using optogenetics, a technique that allows targeted delivery of light-sensitive ion channels to genetically-identified neurons [37]. In general, inhibiting D1 or activating D2 MSNs decreased licking, while activating D1 or inhibiting D2 MSNs increased licking (Fig. 2a-b). However, changes were not uniform across trial types; for example, activating D1 MSNs caused a much greater increase in licking for Nothing odors than for Fixed and Variable odors. We then compared these trends to a variety of dRL models, in which inhibition and and excitation were simulated by clamping value predictors to low or high levels, respectively, separately for optimistic and pessimistic neurons. To account for the inverse coding of D2 MSNs, we also fit "reflected" variants of these models in which inhibition increased and excitation decreased the associated values specifically for pessimistic predictors (Fig. 2c). Only models with inherent opponency could capture our data (Fig. 2d-f). The Reflected Expectile model is particularly interesting in this regard, since midbrain dopamine neurons have been previously suggested to form an expectile RPE code [7], and using D2 MSNs to encode expectiles below the mean would allow the striatum to learn from negative RPEs.

3 Discussion

Together, these findings highlight an impressive correspondence between dRL and the mammalian basal ganglia, with both learning the distribution of returns. However, only the brain instantiates value predictors in two distinct yet complementary populations. Part of the explanation for this difference is the simple fact that biological neurons, unlike artificial ones, are restricted to non-negative firing rates. Yet given the widespread observation of opponency throughout the brain, there might be a more fundamental reason for this division. One possibility is that neurons specialized in positive or negative outcomes can speed learning in rich or lean environments, respectively [38], or guide exploration [39]. In addition, just as is the case with ON/OFF pathways in vision [40], optimistic and pessimistic predictors may sometimes operate independently (as when assessing best or worst-case outcomes) and other times must be combined (as when computing expected value). Separate pathways would thereby enable maximum flexibility and speed of downstream computations. It remains to be seen whether such a division might also be of some benefit in machine learning, closing the loop between our algorithmic understanding of the basal ganglia and reinforcement learning [41].

References

1. Montague, P. R. *et al.* A Framework for Mesencephalic Dopamine Systems Based on Predictive Hebbian Learning. *The Journal of Neuroscience* **16,** 1936–1947 (1996).
2. Schultz, W. *et al.* A neural substrate of prediction and reward. *Science* **275,** 1593–1599 (1997).
3. Bellemare, M. G. *et al. A Distributional Perspective on Reinforcement Learning Proceedings of the 34th International Conference on Machine Learning* **70** (2017), 449–458.
4. Dabney, W. *et al. Distributional Reinforcement Learning With Quantile Regression Proceedings of the AAAI Conference on Artificial Intelligence* **32** (2018).
5. Wurman, P. R. *et al.* Outracing champion Gran Turismo drivers with deep reinforcement learning. *Nature* **602,** 223–228 (2022).
6. Lowet, A. S. *et al.* Distributional Reinforcement Learning in the Brain. *Trends Neurosci.* **43,** 980–997 (2020).
7. Dabney, W. *et al.* A distributional code for value in dopamine-based reinforcement learning. *Nature* **577,** 671–675 (2020).
8. Tano, P. *et al.* A local temporal difference code for distributional reinforcement learning. *Adv. Neural Inf. Process. Syst.* **33,** 13662–13673 (2020).
9. Caraco, T. *et al.* An empirical demonstration of risk-sensitive foraging preferences. *Anim. Behav.* **28,** 820–830 (1980).
10. Hertwig, R. & Erev, I. The description-experience gap in risky choice. *Trends Cogn. Sci.* **13,** 517–523 (2009).
11. Constantinople, C. M. *et al.* An Analysis of Decision under Risk in Rats. *Curr. Biol.* **29,** 2066–2074.e5 (2019).
12. Doya, K. Reinforcement learning: Computational theory and biological mechanisms. *HFSP J.* **1,** 30–40 (2007).
13. Yagishita, S. *et al.* A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science* **345,** 1616–1620 (2014).
14. Stopper, C. M. & Floresco, S. B. Contributions of the nucleus accumbens and its subregions to different aspects of risk-based decision making. *Cogn. Affect. Behav. Neurosci.* **11,** 97–112 (2011).
15. Zalocusky, K. A. *et al.* Nucleus accumbens D2R cells signal prior outcomes and control risky decision-making. *Nature* **531,** 642–646 (2016).
16. Kalkhoven, C. *et al.* Risk-taking and pathological gambling behavior in Huntington's disease. *Front. Behav. Neurosci.* **8,** 103 (2014).
17. Gatto, E. M. & Aldinio, V. Impulse Control Disorders in Parkinson's Disease. A Brief and Comprehensive Review. *Front. Neurol.* **10,** 351 (2019).
18. Mikhael, J. G. & Bogacz, R. Learning Reward Uncertainty in the Basal Ganglia. *PLoS Comput. Biol.* **12,** e1005062 (2016).
19. Schultz, W. *et al.* Neuronal activity in monkey ventral striatum related to the expectation of reward. *J. Neurosci.* **12,** 4595–4610 (1992).
20. Taha, S. A. & Fields, H. L. Encoding of palatability and appetitive behaviors by distinct neuronal populations in the nucleus accumbens. *J. Neurosci.* **25,** 1193–1202 (2005).
21. Strait, C. E. *et al.* Signatures of Value Comparison in Ventral Striatum Neurons. *PLoS Biol.* **13,** e1002173 (2015).
22. 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).

23. Shin, J. H. *et al.* Differential coding of reward and movement information in the dorsomedial striatal direct and indirect pathways. *Nat. Commun.* **9,** 404 (2018).
24. Shin, E. J. *et al.* Robust and distributed neural representation of action values. *Elife* **10** (2021).
25. Stringer, C. *et al.* Spontaneous behaviors drive multidimensional, brainwide activity. *Science* **364,** 255 (2019).
26. Bernardi, S. *et al.* The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. *Cell* **183,** 954–967.e21 (2020).
27. Claudi, F. *et al.* Visualizing anatomically registered data with brainrender. *Elife* **10** (2021).
28. Gerfen, C. R. *et al.* D_1 and D_2 Dopamine Receptor-Regulated Gene Expression of Striatonigral and Striatopallidal Neurons. *Science* **250,** 1429–1432 (1990).
29. Iino, Y. *et al.* Dopamine D2 receptors in discrimination learning and spine enlargement. *Nature* **579,** 555–560 (2020).
30. Lee, S. J. *et al.* Cell-type-specific asynchronous modulation of PKA by dopamine in learning. *Nature* (2020).
31. 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).
32. Rowland, M. *et al. Statistics and Samples in Distributional Reinforcement Learning Proceedings of the 36th International Conference on Machine Learning* **97** (2019), 5528–5536.
33. Govorunova, E. G. *et al.* NEUROSCIENCE. Natural light-gated anion channels: A family of microbial rhodopsins for advanced optogenetics. *Science* **349,** 647–650 (2015).
34. Li, N. *et al.* Spatiotemporal constraints on optogenetic inactivation in cortical circuits. *Elife* **8** (2019).
35. Klapoetke, N. C. *et al.* Independent optical excitation of distinct neural populations. *Nat. Methods* **11,** 338–346 (2014).
36. Gong, S. *et al.* A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. *Nature* **425,** 917–925 (2003).
37. Deisseroth, K. Optogenetics: 10 years of microbial opsins in neuroscience. *Nat. Neurosci.* **18,** 1213–1225 (2015).
38. 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).
39. Mavrin, B. *et al.* Distributional Reinforcement Learning for Efficient Exploration (2019).
40. Ichinose, T. & Habib, S. ON and OFF Signaling Pathways in the Retina and the Visual System. *Front Ophthalmol (Lausanne)* **2** (2022).
41. Hassabis, D. *et al.* Neuroscience-Inspired Artificial Intelligence. *Neuron* **95,** 245–258 (2017).