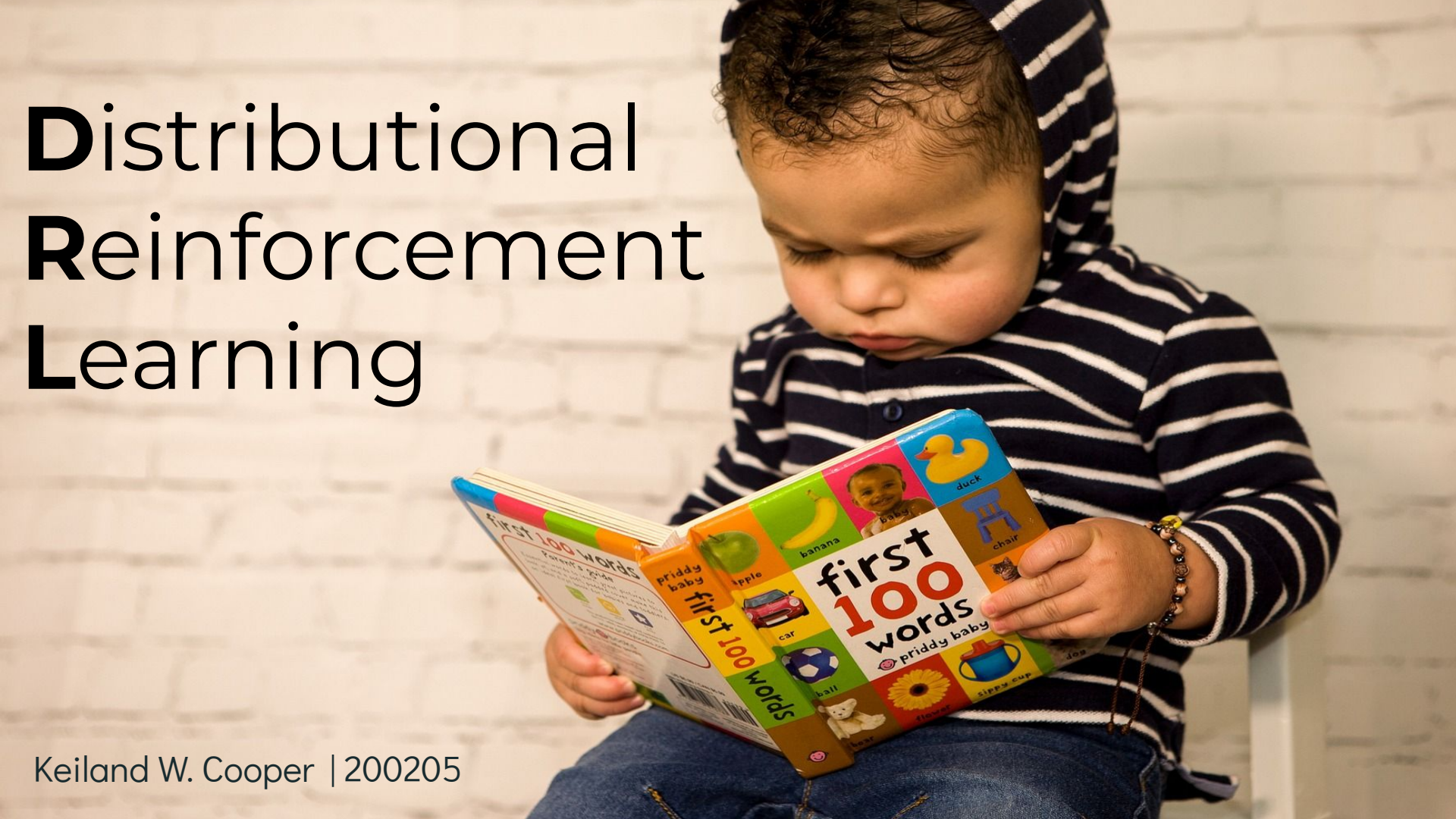
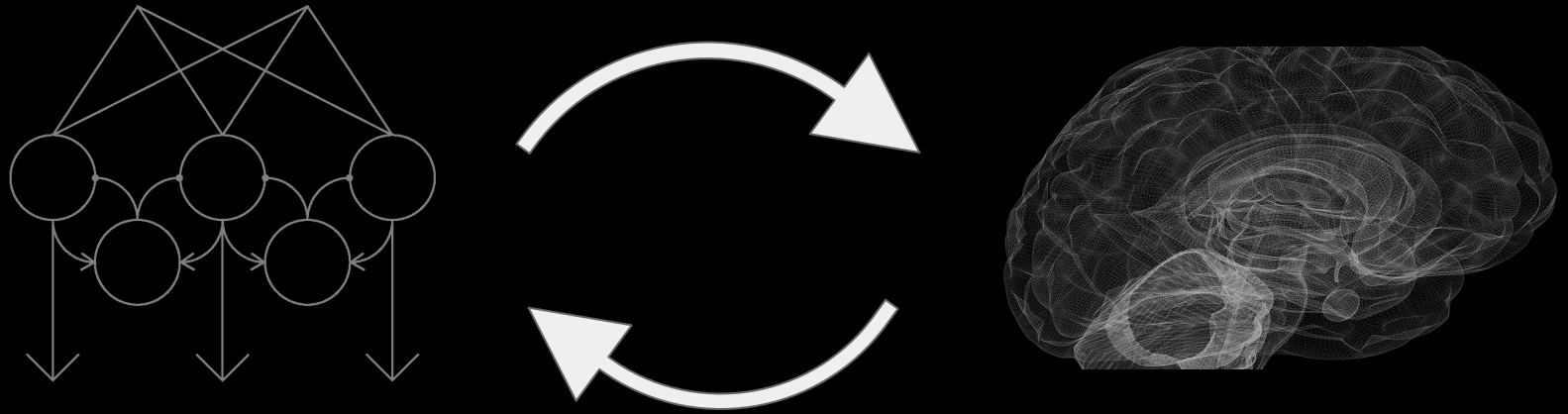
Distributional
Reinforcement
Learning



Keiland W. Cooper | 200205

AI & Neuroscience Feedback Loop



Ex. Catastrophic Forgetting: Cichon et al 2015 -> Kirkpatrick et al 2017

Ex. Hippocampal Decoding: LeCun et al 1998 -> Shahbaba et al 2019

Article

A distributional code for value in dopamine-based reinforcement learning

https://doi.org/10.1038/s41586-019-1924-6
Received: 3 January 2019
Accepted: 19 November 2019
Published online: 15 January 2020

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

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

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

Similar to the traditional form of temporal-difference RL—on which the dopamine theory was based—distributional RL assumes that reward-based learning is driven by a RPE, which signals the difference between received and anticipated reward. (For simplicity, we introduce the theory in terms of a single-step transition model, but the same principles hold for the general multi-step (discounted return) case; see Supplementary Information.) The key difference in distributional RL lies in how 'anticipated reward' is defined. In traditional RL, the representation learning (see Extended Data Figs. 2, 3 and Supplementary Information). This prompts the question of whether RL in the brain might leverage the benefits of distributional coding. This question is encouraged both by the fact that the brain utilizes distributional codes in numerous other domains[16], and by the fact that the mechanism of distributional RL is biologically plausible[6,17]. Here we tested several predictions of distributional RL using single-unit recordings in the ventral tegmental area (VTA) of mice performing tasks with probabilistic rewards.

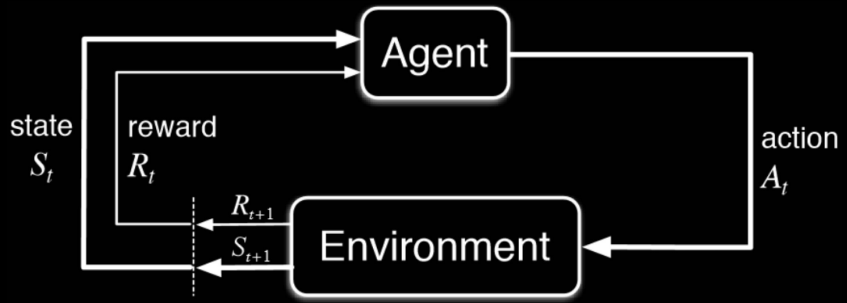
Value predictions vary among dopamine neurons

In contrast to classical temporal-difference (TD) learning, distributional RL posits a diverse set of RPE channels, each of which carries a different value prediction, with varying degrees of optimism across channels. (Value is formally defined in RL as the mean of future outcomes, but here we relax this definition to include predictions about future outcomes that are not necessarily the mean.) These value predictions in turn provide the reference points for different RPE signals, causing the latter to

“One may even say, strictly speaking, that almost all our knowledge is only probable; and in the small number of things that we are able to know with certainty, the principle means of arriving at the truth—induction and analogy—are based on probabilities”

- Pierre Simon Laplace

Traditional RL (tRL)



Sutton & Barto 1998

Temporal Difference theory
Reward prediction error (RPE) drives learning

Agent learns reward expectation:
Rewards below this expectation elicit -RPE
Rewards above this expectation elicit +RPE

$$Q(s,a) = r + \gamma max_{a'} Q(s',a')$$

Reversal point: RPE flips from - to +

See also Hessel et al 2019; Glimcher 2011; Montague et al 1996

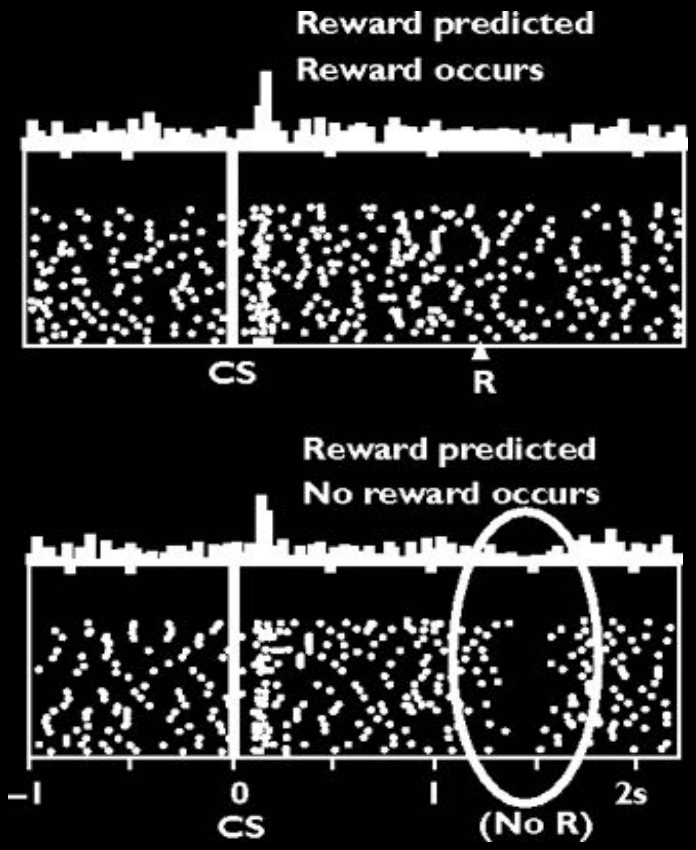
Traditional RL (tRL)

Reward prediction error (RPE) drives learning

Theory of dopamine and RL

VTA firing rate increase post CS

Each neuron codes for the same value?
Mean of all rewards, weighted by P(R)



See also Hessel et al 2019; Glimcher 2011; Montague et al 1996

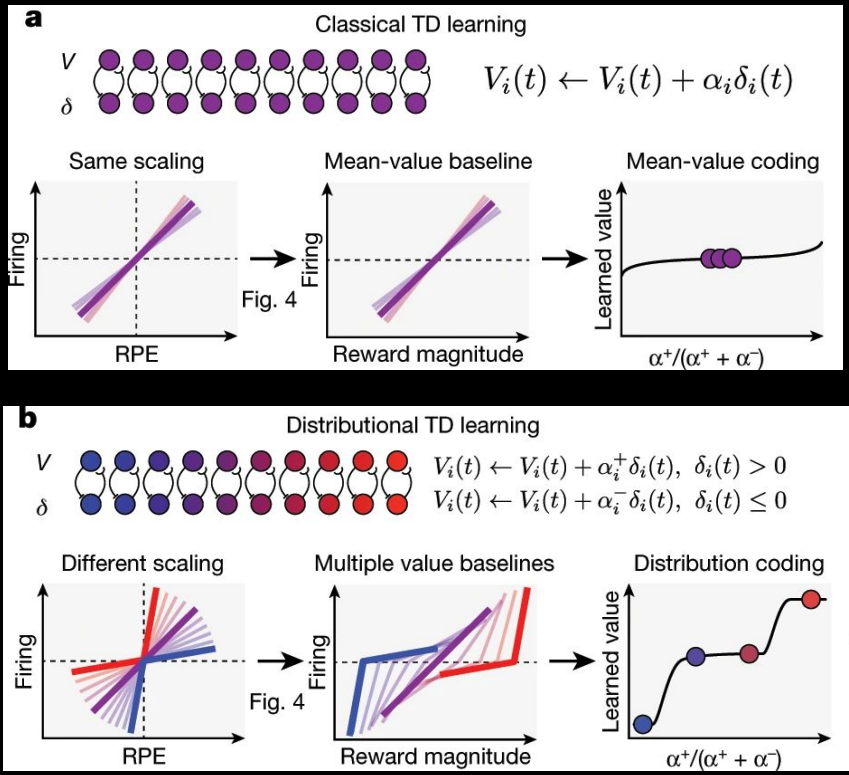
Distributional RL (dRL)

But does each neuron code for one value?

dRL posits, rather each predictor codes for P(V)

Prediction:
tRL: population firing rate correlates with RPE
dRL: ‘ ‘ scales individually

Scaled by alpha +/-
imbalance over predictors

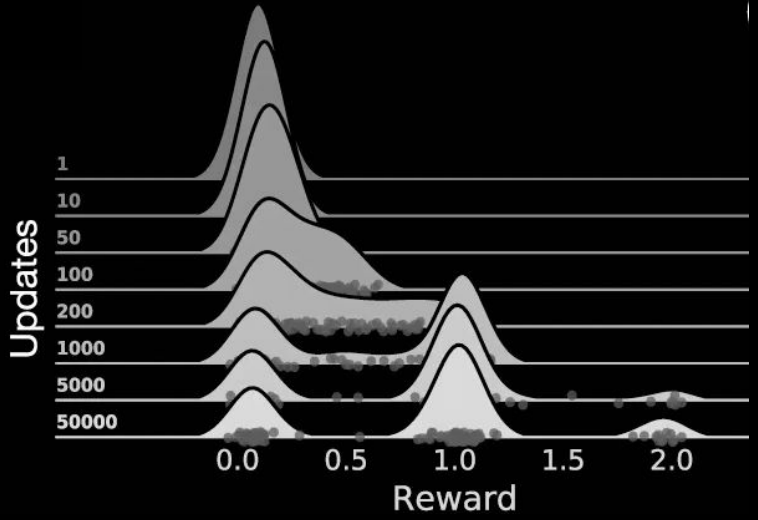


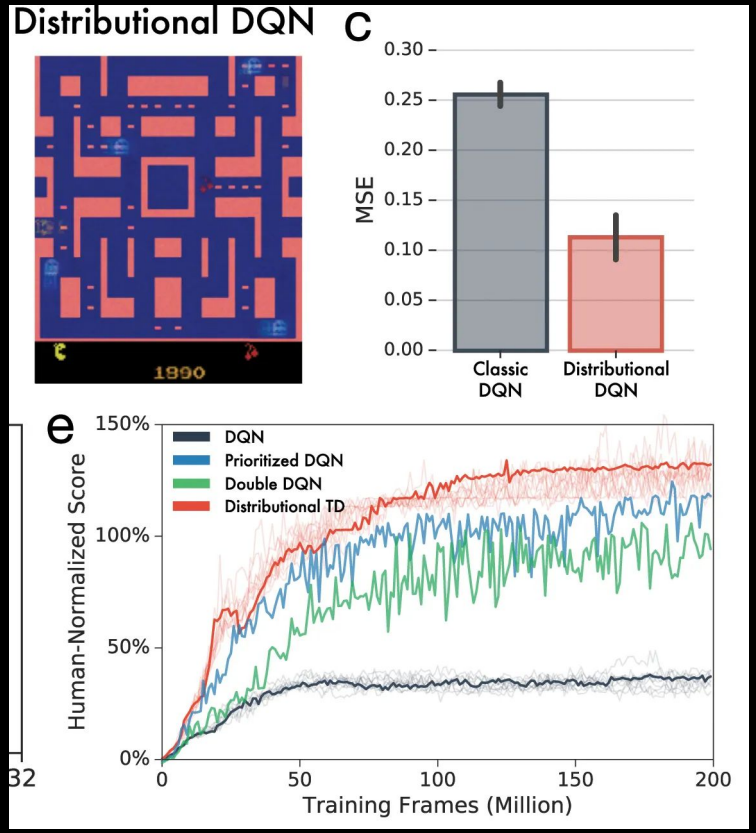
Morimura et al 2010; Pouget et al 2013; Lammel 2014; Bellemare et al 2017; Dabney 2018

Distributional RL (dRL)

Distribution is created over time

Can increase performance of RL Agent





Schaul 2016; Bellemare et al 2017; Dabney 2018

Tasks

Variable Magnitude

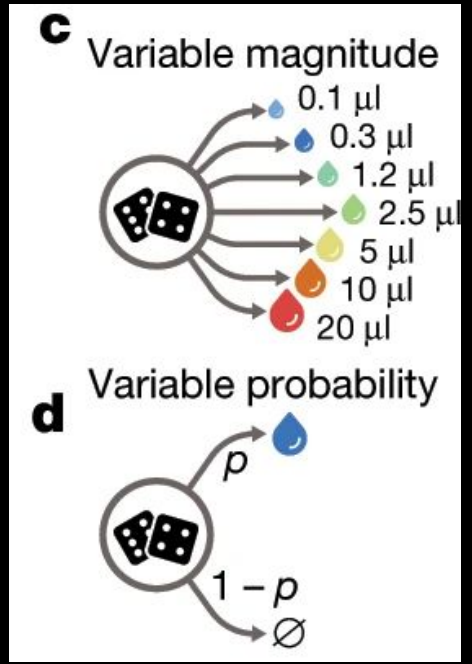
Cue -> variably scaled liquid reward

Unpredictable

Variable Probability Task

3 Cues -> variable P(Reward)

Reward magnitude fixed



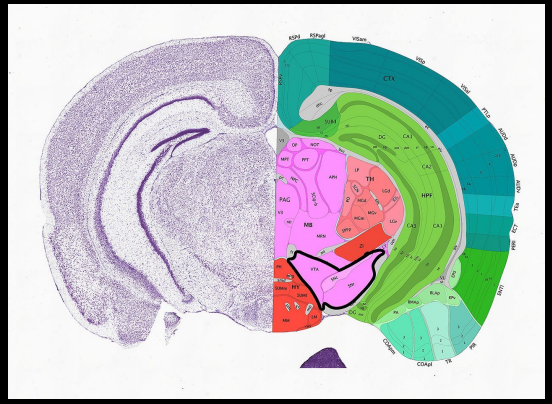
VTA Single Unit Recordings

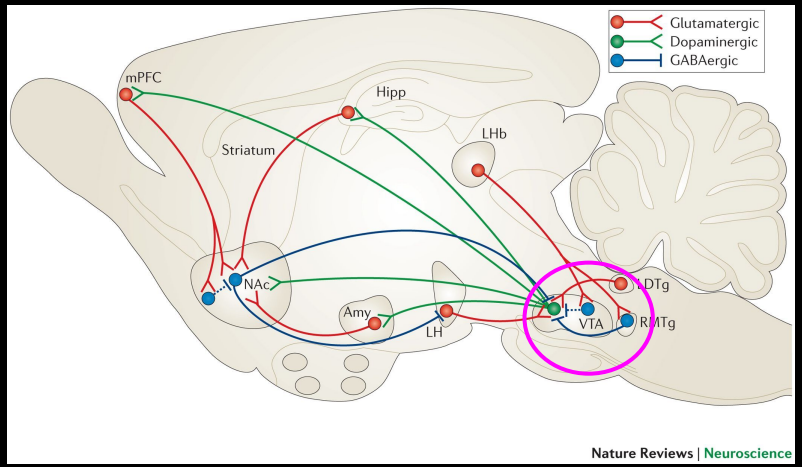
11 Mice implanted -> **Eshel et al 2015**
5 on VP; 6 on VM

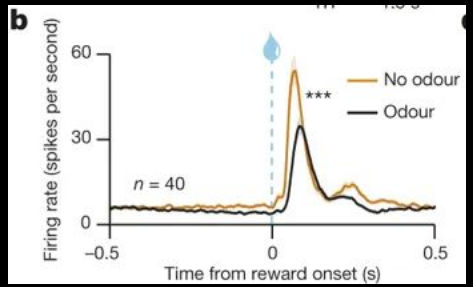
How do dopamine neurons calculate RPE?

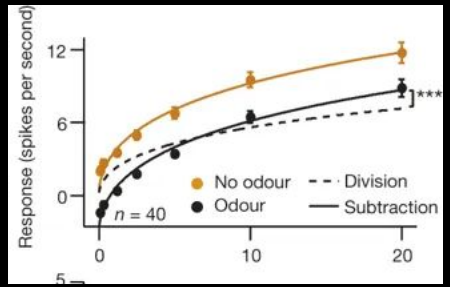
Selectively labeled dopamine neurons via opto

(Found subtraction via GABA was best fit)



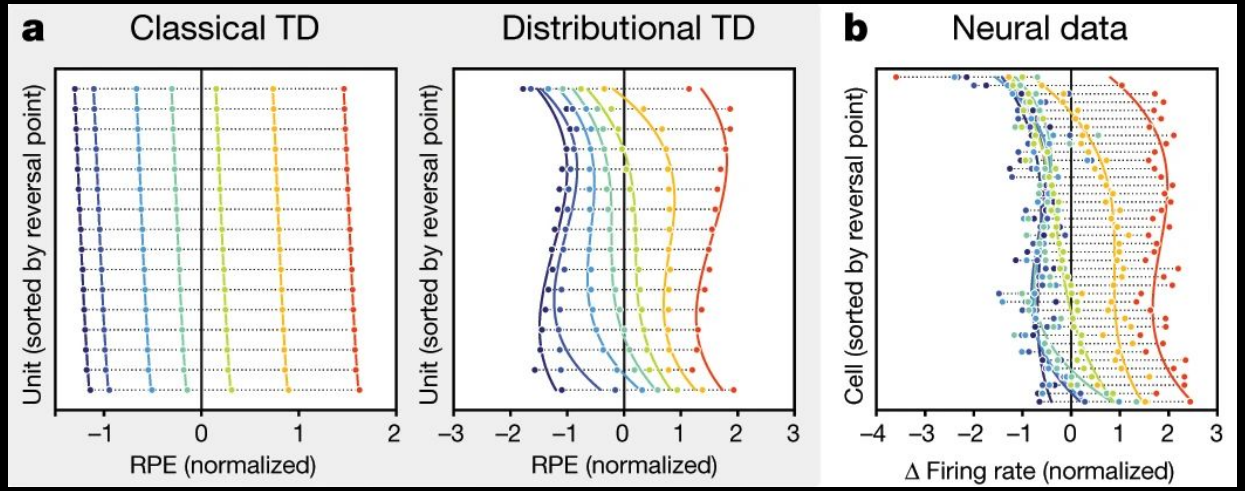


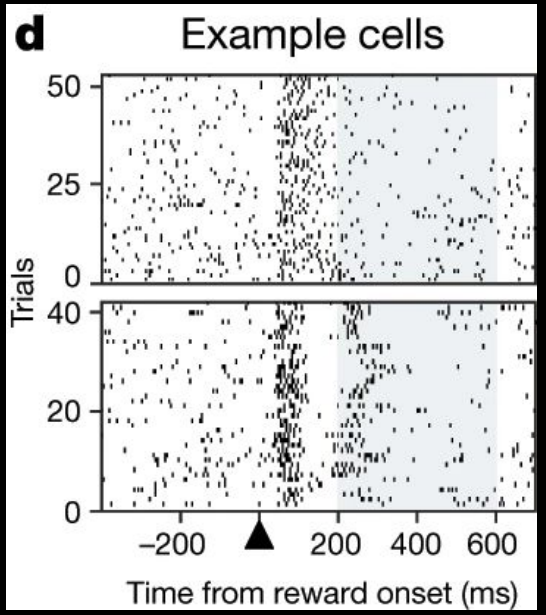




Russo & Nestler 2013

P(V) varies among VTA neurons

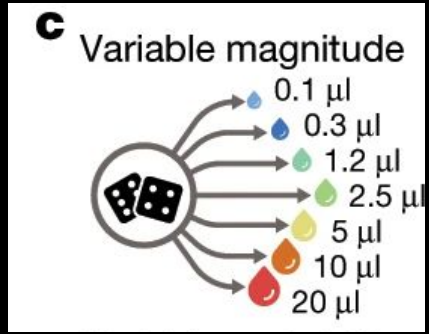




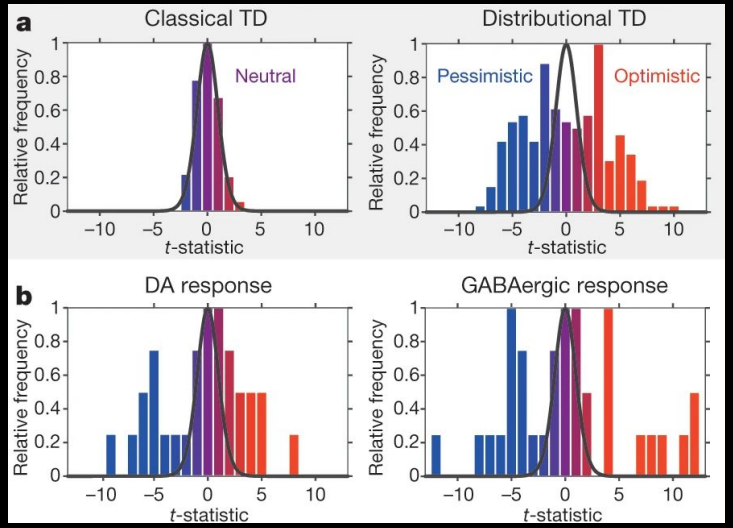
In dRL, predictors show graded optimism

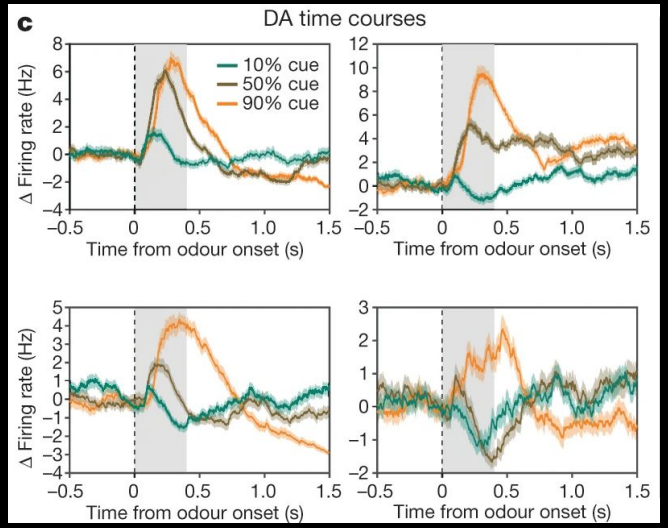
Hoz bar = 1 neuron; xax = delta FR; consistent across data

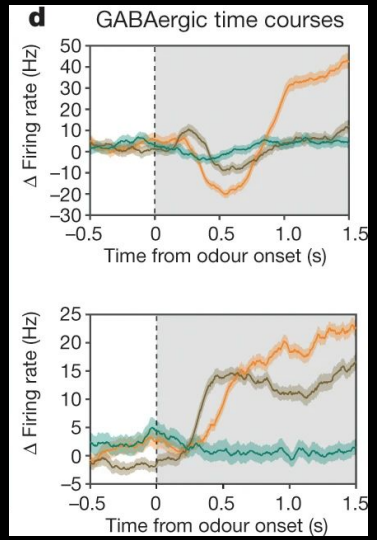
Variability found in firing rates to reward magnitude



P(V) varies among VTA neurons





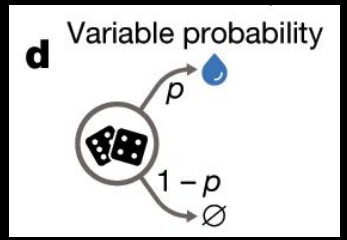


In dRL, predictors show graded optimism

50% cue response; 4 example dopamine neurons.

Variability found in firing rates to reward probability

Also shown in GABAergic neurons FR



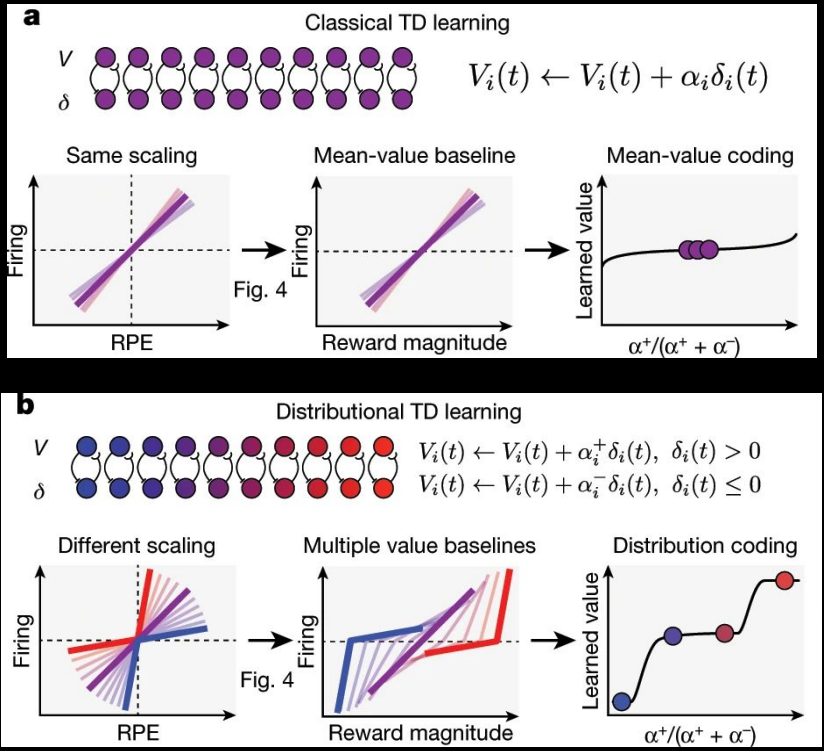
dRL from asymmetric RPE scaling

In tRL, + & - RPE's given equal weight
In dRL, *a* scales the predictor optimism

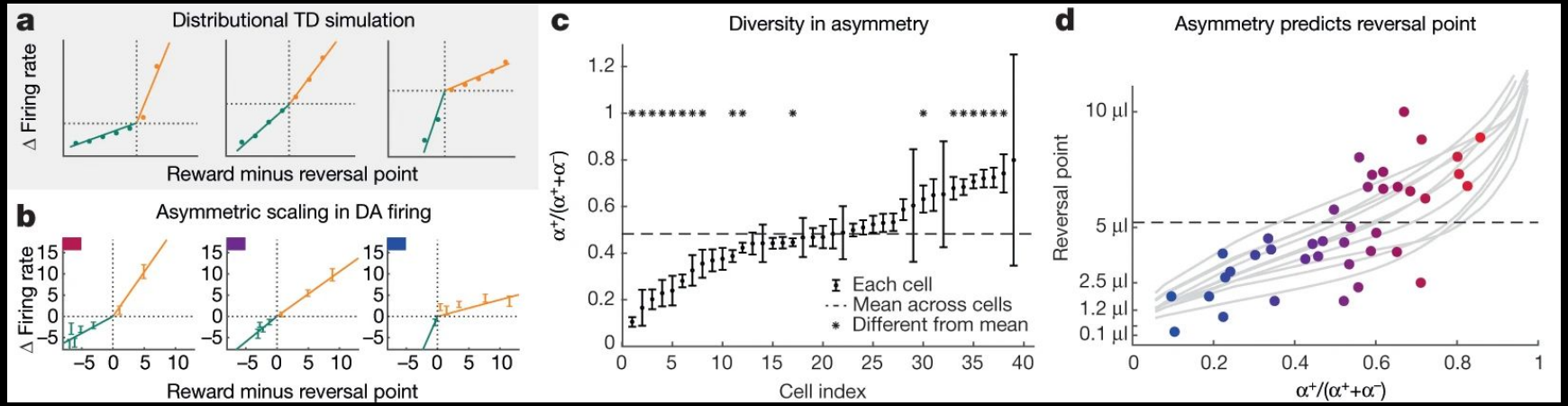
Predictions:

+ & - RPE's should scale differently

RPE asymmetry should correlate
with reversal point



dRL from asymmetric RPE scaling

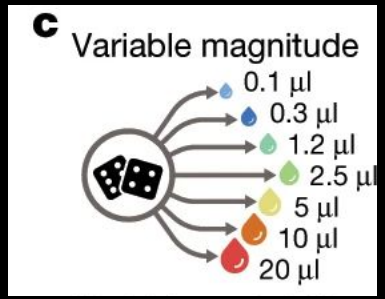


Estimated cell reversal point -> compute +/-*α*

Revealed scaling differences in dopamine neurons

Great deal of variability reported in scaling

Asymmetry correlates with reversal point
(in a single animal even)

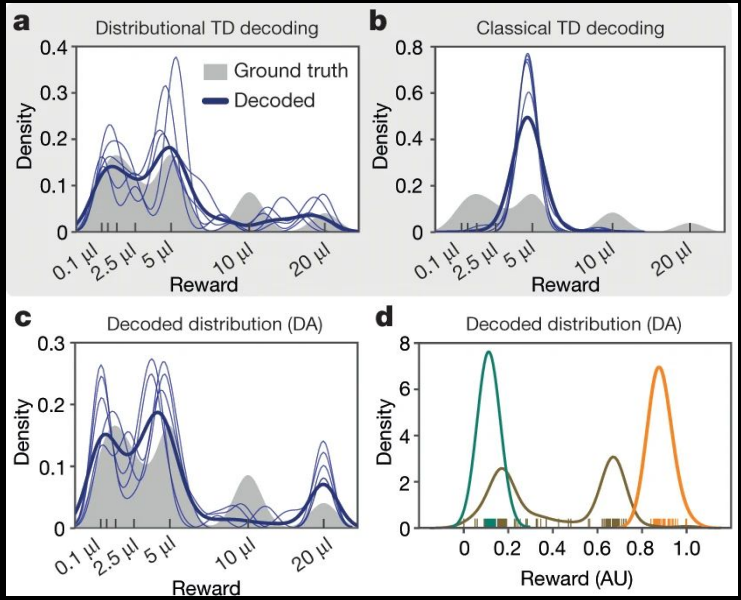


Decoding reward distributions

learned values -> expectiles -> P(R)
Expectiles: reversal points and +/-*α*'s

Can decode the ground truth by dRL in both tasks

Also could use GABA neurons to decode VM task



Russo & Nestler 2013

“Why have the present effects not been observed before? … One of the earliest studies of reward-probability coding in dopaminergic RPEs remarked on apparent diversity across dopamine neurons, but only in a footnote”

“The likelihood that individual neurons have distinct thresholds has critical implications for understanding the shape of the probability-response functions presented… Because these ranges are unknown, the only interpretation that should be given to the data at this time is that dopamine neuronal responses follow probability or uncertainty in a monotonic fashion.”

-Fiorillo et al 2003

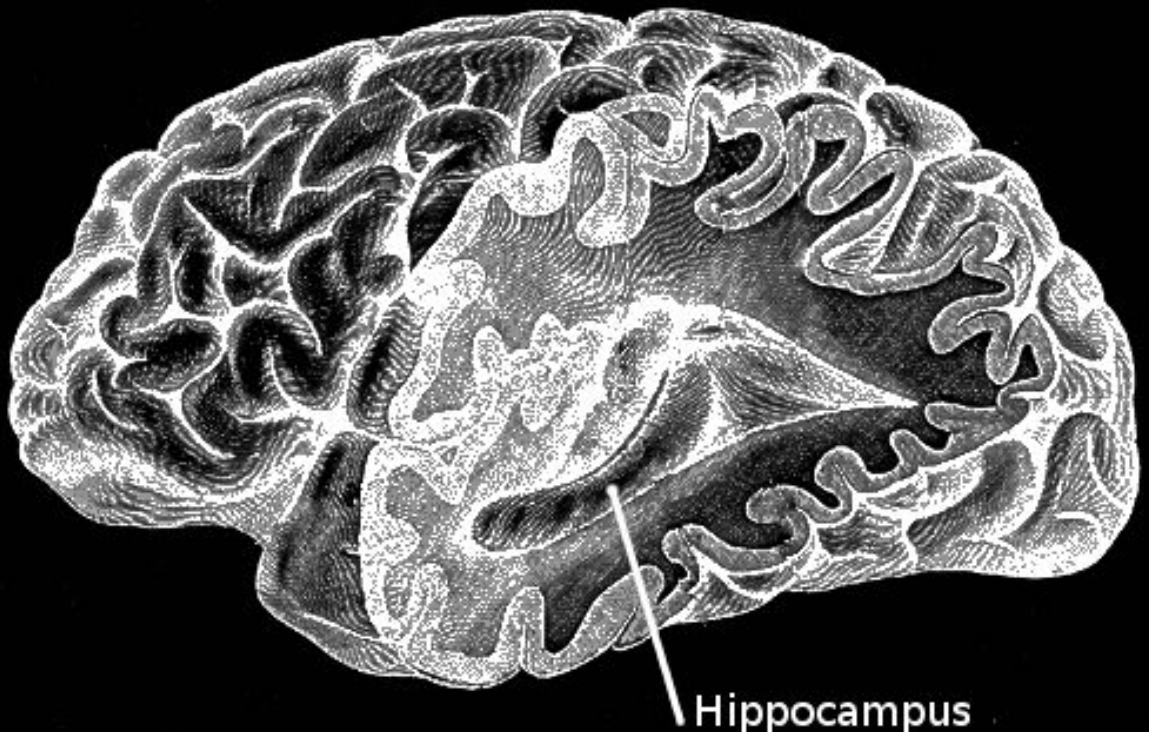
animals, or only between animals. In Extended Data Figure 6c we show that, indeed, even within individual animals the results seen in these past works are reproduced. Finally, observe in Extended Data Figure 6d that distributional TD, but not classical TD, predicts similar types of asymmetric diversity.

6 Predictions

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

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

References


Hippocampus

What would distributional coding mean for memory?

Might there be uncertainty about the rat's position?

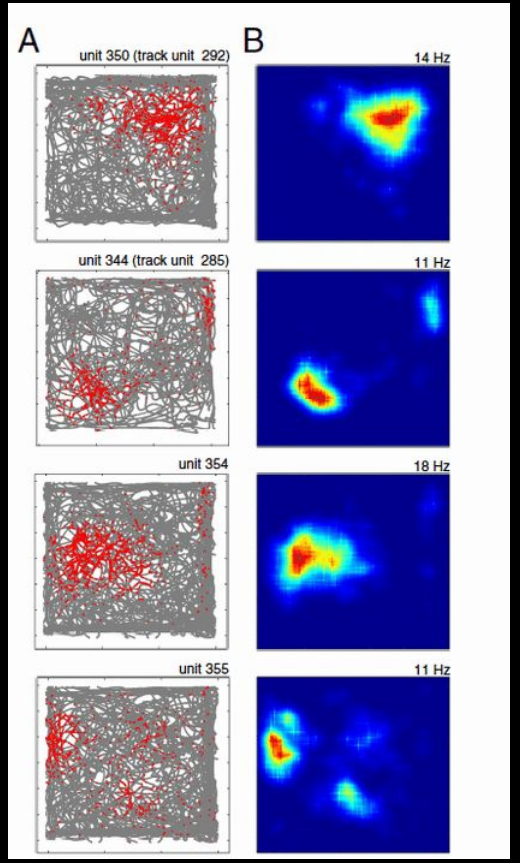
Partial remapping in an ambiguous environment
'Flickering' between place fields
Rate remapping?

Might there be uncertainty about the upcoming state?

Hippocampal preplay; imaginative sequences

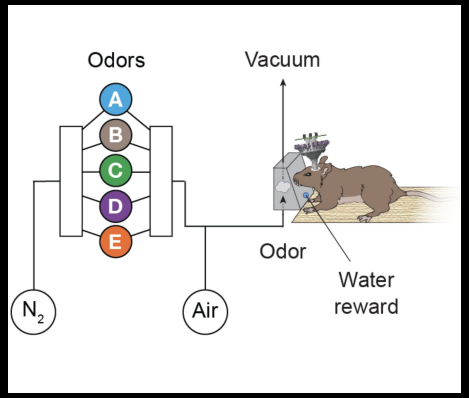
Human hippocampus encodes RPE (ERP study)

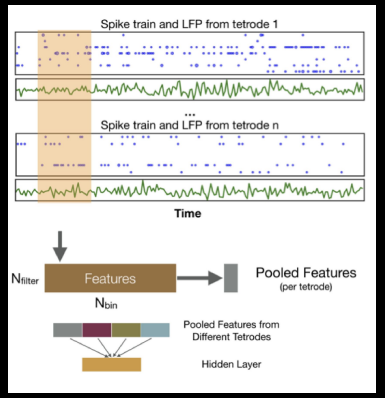
Fair amount of fMRI evidence

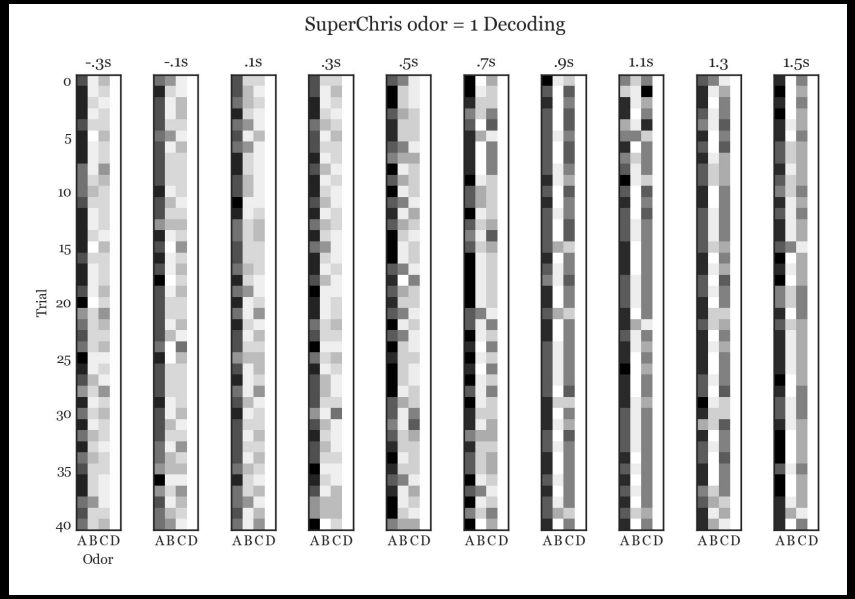


Zemel et al 1998; Resnik et al 2007; Pfeiffer et al 2017; Vanni-Mercier et al 2009

What would distributional coding mean for memory?



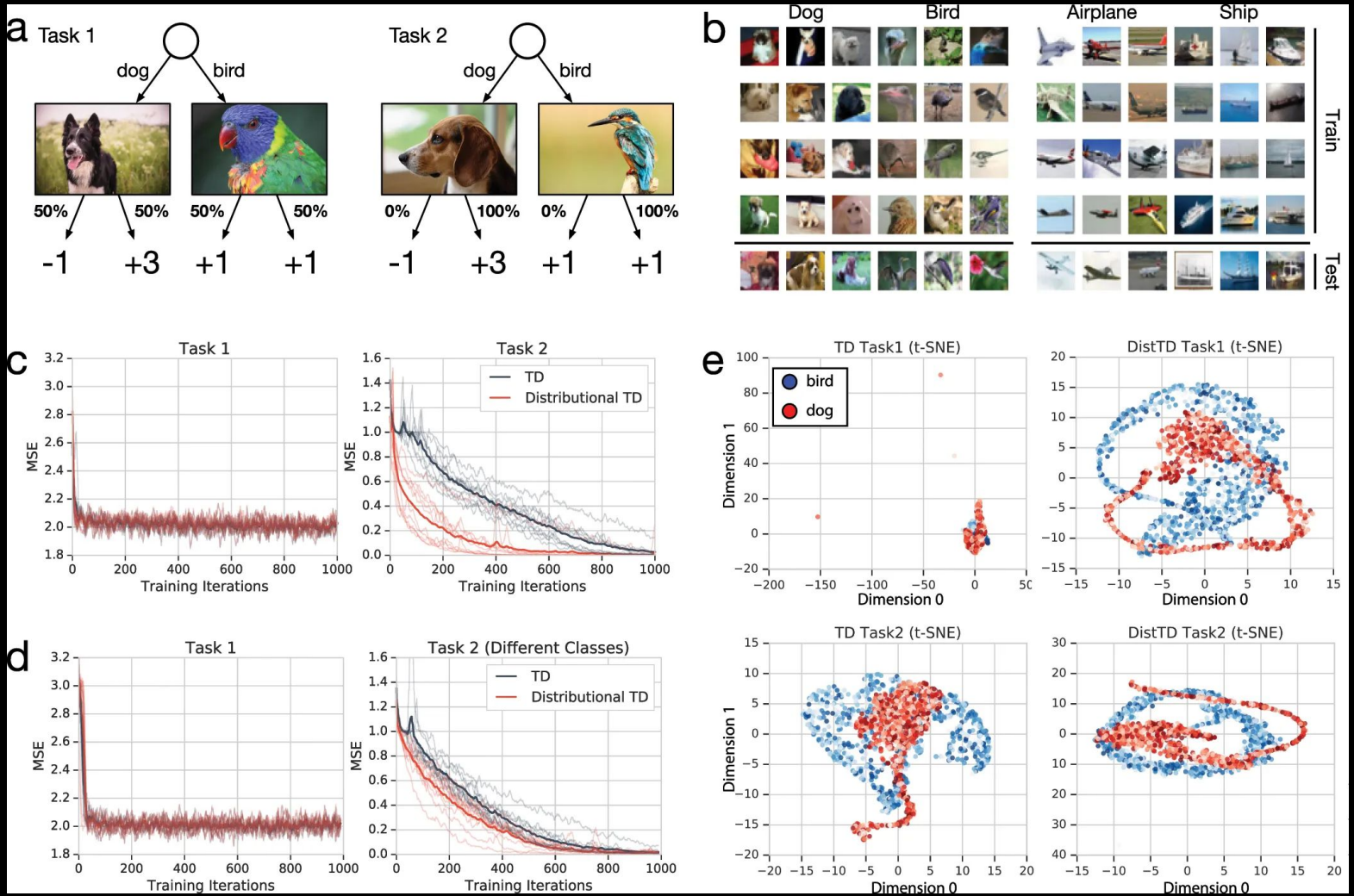




Sequence decoding via CNN reveals multiple representations of odors

Shahbaba et al 2019 (aRxiv); unpublished data

Thank you!



Representation learning