Distributional Reinforcement Learning in Prefrontal Cortex

Muller et al. (2024)

Discussed by
Harvey Huang

March 2024

Table of Contents

Table of Contents

Reinforcement learning, RPE and prefrontal cortex (PFC)

- Classic RL and the reword prediction error (RPE) hypothesis.
  - All neurons learn to predict the same expected reward (mean) => missing diversity (Wallis & Kennerley, 2010; Dabney et al., 2020).
- Distributional RL (Dabney et al., 2020).
  - Optimistic neurons encode values **above** the mean of the reward distribution.
  - Pessimistic neurons encode values **below** the mean of the reward distribution.
  - (Neural) risk preference (Morimura, Sugiyama, Kashima, Hachiya, & Tanaka, 2012)?
  - Link to reference-point and range adaptation at a neuron level (Rigoli, Friston, & Dolan, 2016; Bavard, Lebreton, Khamassi, Coricelli, & Palminteri, 2018)

Distributional RL and RPE

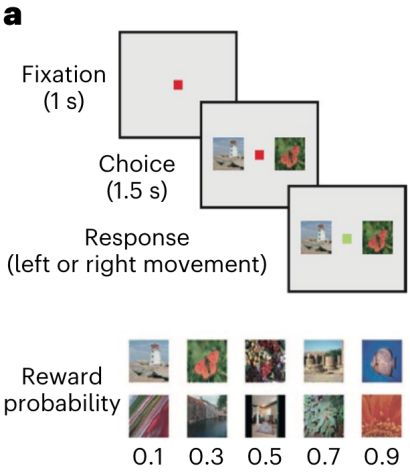
- RPE responses differed across individual neurons (Fiorillo, Tobler, & Schultz, 2003).
- Distributional RL provides a paradigm/framework/model to explain this(Dabney et al., 2020).
- The neurons switched from decreased to increased activity (compared to the baseline/mean activity) at different reward magnitudes.

Two experiments (datasets)

- First dataset (Kennerley, Behrens, & Wallis, 2011):
  - Prediction 1: different neurons carry different value predictions (optimistic/pessimistic).
  - Prediction 2: (the above neurons have) asymmetry in positive RPEs vs. negative RPEs.
  - Prediction 3: the above two forms of "diversity" correlate.
- Second dataset (Miranda, Malalasekera, Behrens, Dayan, & Kennerley, 2020):
  - Prediction: diverse asymmetries in the **rates** of learning from positive vs. negative RPEs <=> Optimistic cells learn faster from positive RPEs and slowly from negative RPEs, and pessimistic cells the opposite.

Table of Contents

Experiment 1: classic bandit choice-reward association task

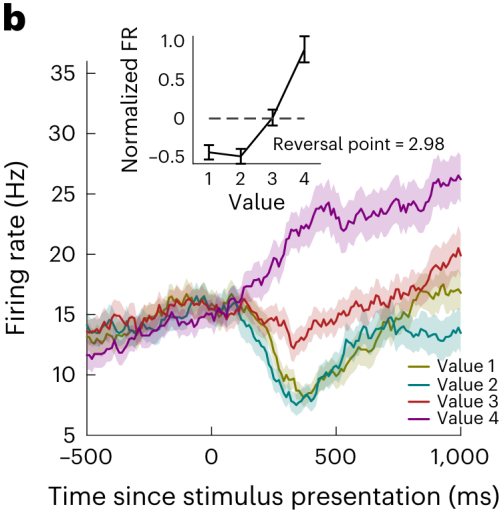


- Two "well-trained" non-human primates (NHPs, Macaca mulatta)
- Three PFC regions **recorded**: the lateral PFC (LPFC, n = 257), the orbitofrontal cortex (OFC, n = 140) and the anterior cingulate cortex (ACC, n = 213). See Kennerley et al. (2011) for the location of neurons.
- One stimuli (prob=0.1) was eventually dropped in the analysis due to exceptionally good performance.

Neuron inclusion criteria and assumptions (n) in this analysis

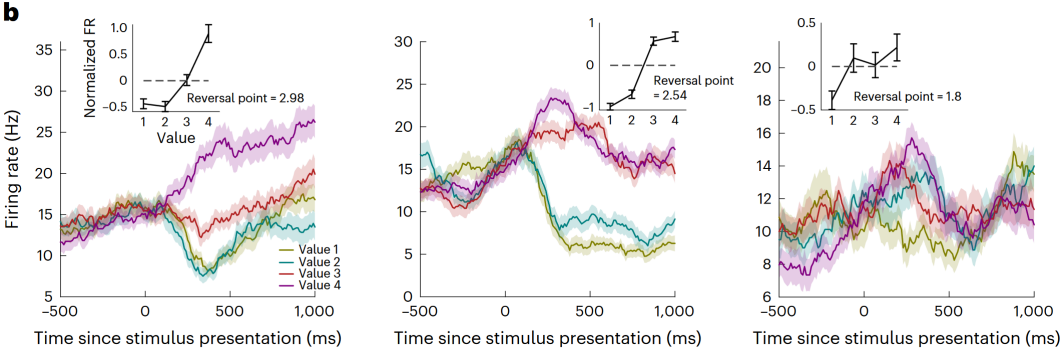
- Regression between (choice) probability level and mean firing rate on each trial in a 200- to 600-ms analysis window after cue onset.
  - pick neurons where $p < 0.05$.
  - See paper for the rationale of 200- to 600-ms window.
- Based on this criteria, 19% ACC neurons ($N = 41$), 6% OFC ($N = 7$) and 10% LPFC neurons ($N = 26$).
  - focus on ACC neurons only.
  - a region known to contain value-related learning signals (Kennerley et al., 2011), and important for risk-sensitive decision-making (Kolling, Wittmann, & Rushworth, 2014).

Measuring “optimism”: reversal point

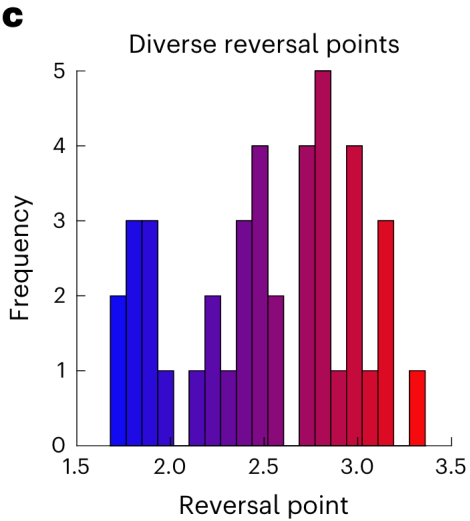


- Value 1 <=> reward $p = 0.3$, 2: 0.5, 3: 0.7, 4: 0.9
- **Reversal point:**
  - compare neuron’s response to the mean firing rate across trials in the analysis window after the cue (i.e., de-mean). Echo Dabney et al. (2020) with minor differences.
  - the point where the de-meaned value goes from -ve to +ve.

Three example neurons with different “reversal points”

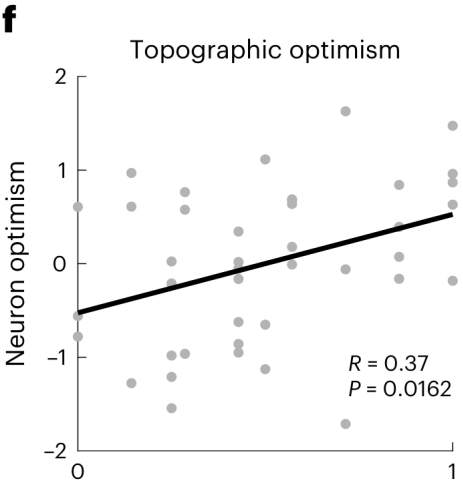


Prediction 1: different neurons carry different value predictions



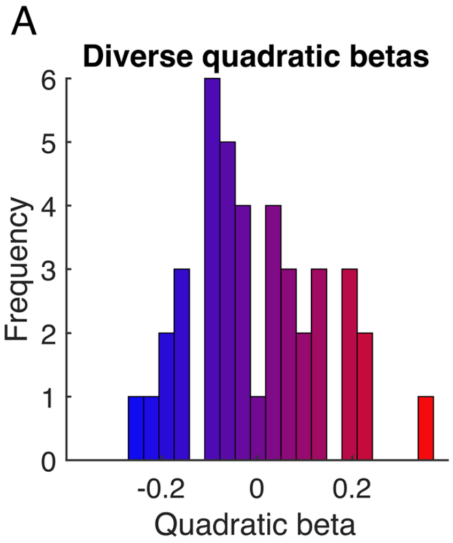
- ACC neurons only, $N = 213$.
- Blue <=> Pessimistic, Red <=> Optimistic
- In classical RL, all neurons have the same reversal points (mean).
- Expecting a uniform distribution? No, "learned reversal points are predicted to correspond to expectiles of the reward distribution" (Dabney et al., 2020).

How do optimistic neurons in one region communicate with those in another?



- Each data point denotes a neuron.
- X-axis: 0 <=> posterior insular, 1 <=> anterior insular (indicated by location).
- More anterior (located) neurons were more optimistic.

An alternative measure of "optimism": quadratic regression

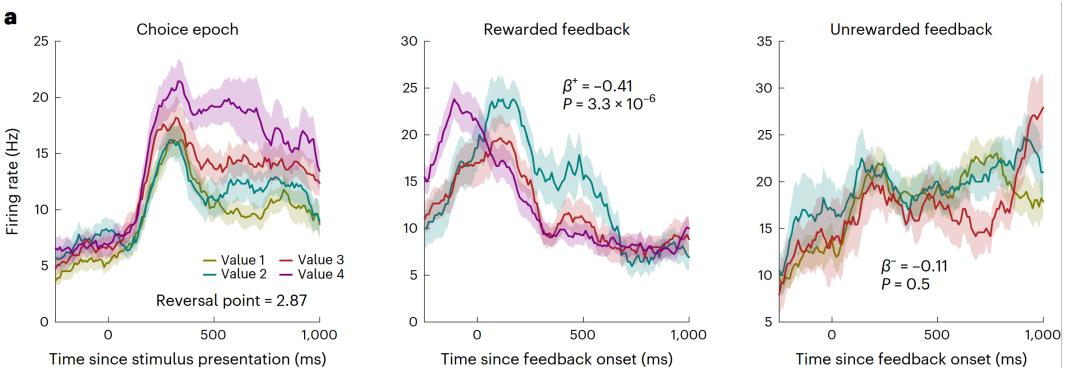


- $FR = \beta_0 + \beta_1 R + \beta_2 R^2$.
- where FR is the neuron firing rate, and R is the reward level.
- β_2 measures the convexity/concavity of the relationship.

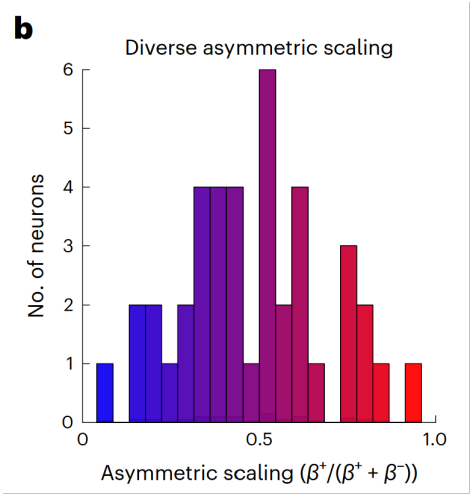
Prediction 2: (feedback) asymmetry in positive RPEs vs. negative RPEs.

- RPE = r (delivered reward) - V (reward probability)
- +ve RPE: (1− chosen reward prob) on rewarded trials (hence ≥ 0).
- −ve RPE: (0− chosen reward prob) on unrewarded trials (hence ≤ 0).
- "regress the chosen cue probability against the (mean) firing rate at feedback."
  - $FR \sim \alpha + \beta^{+/-} * V$
  - Coefficients: β^+ for +ve RPE and β^- for −ve RPE
  - Coefficients measure a neuron's sensitivity to reward probability when there is a surprise.
  - (Feedback) Asymmetric scaling measure = $\beta^+/(\beta^+ + \beta^-)$

Asymmetric scaling measure for an example neuron

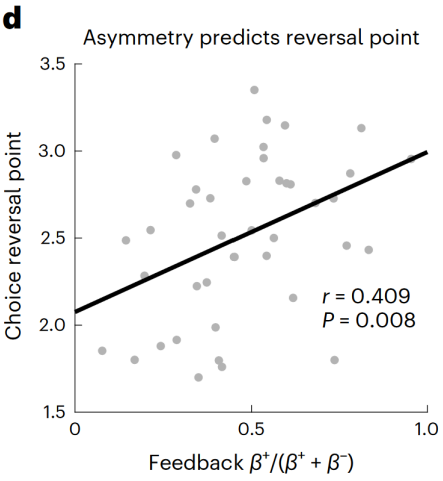


Prediction 2: (feedback) asymmetry in positive RPEs vs. negative RPEs.



- Blue <=> Pessimistic, Red <=> Optimistic

Prediction 3: reversal points correlate with asymmetry in positive and negative RPEs.



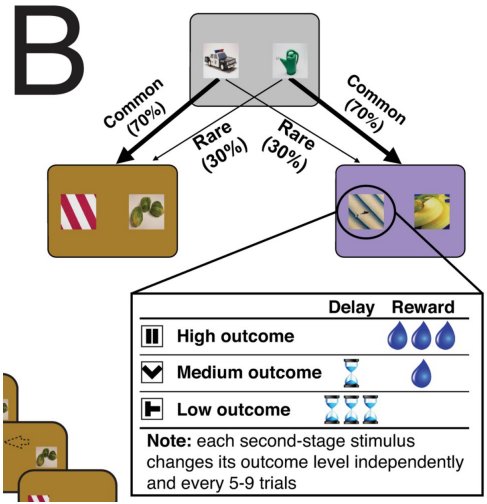
- In distributional RL, optimism arises from the asymmetry scaling of RPEs.
- If a neuron learns more from positive than negative RPEs, then the neuron naturally encodes more optimistic value prediction.

Table of Contents

Experiment 2

- In experiment 1, subjects are **well-trained** and there is no state transition: hence no learning.
- Hypothesis:
  - Diverse asymmetries in rates of learning from positive vs. negative RPEs.
  - Optimistic cells learn rapidly from positive RPEs and slowly from negative RPEs, and pessimistic cells the opposite.

Experiment 2: a two-stage bandit task



- Miranda et al. (2020)
- The outcome (juice as a reward) from the second stage is dynamic.
- Two rhesus monkeys.
- This paper focuses on the ACC neural activity at the feedback stage, i.e., when rewards are received.

Model for asymmetric learning

- One-step transition temporal difference (TD) learning model (Rescorla-Wagner) (Rescorla, 1972)

$$V \leftarrow V + \alpha\delta$$

where $\delta = r - V$ is the RPE, r is the reward on the current trial and V is the previous value estimate, and α is the learning rate.

- Distributional RL model (Dabney et al., 2020)

$$V \leftarrow V + \alpha^{+}\delta \quad \delta > 0$$

$$V \leftarrow V + \alpha^{-}\delta \quad \delta \leq 0$$

where α^{+} and α^{-} are separate learning rates for positive and negative RPEs.

(Proposed) asymmetric scaling model: from RPE to neuron firing rates.

- RW model

$$FR = \beta_0 + \beta_1 \delta$$

- Distributional RL model

$$FR = \beta_0 + \beta^+ \delta \quad \delta > 0$$

$$FR = \beta_0 + \beta^- \delta \quad \delta \leq 0$$

- Note asymmetric learning and asymmetric scaling measure are dissociated here (but connected with RPE δ).

Fitting

- A system with four parameters (α^+, α^-, β^+, β^-): very difficult to fit.
- Proposed the following equivalent:

$$FR = \beta_0 + \beta_1 \delta S, \qquad \delta > 0$$

$$FR = \beta_0 + \beta_1 \delta (1 - S), \qquad \delta \leq 0$$

  where $S \in [0, 1]$, acts as a single asymmetric scaling parameter.
- If S is near 1, positive RPEs are scaled greatly relative to negative RPEs.
- Note:

$$\frac{\beta^+}{\beta^+ + \beta^-} = \frac{S}{S + (1 - S)} = S$$

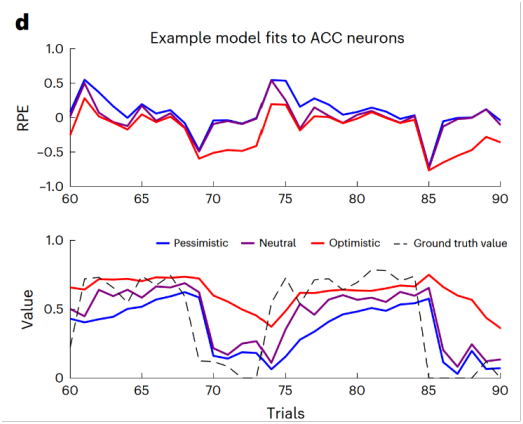
Model fitting

- Grid search for α^+, α^- and S, bounded between -1 and 1, with 0.025 size increments.
- Optimized for R^2.
- Four combinations of models in total.
- ALAS best fits while SLSS (full classic) model is the worst.

b

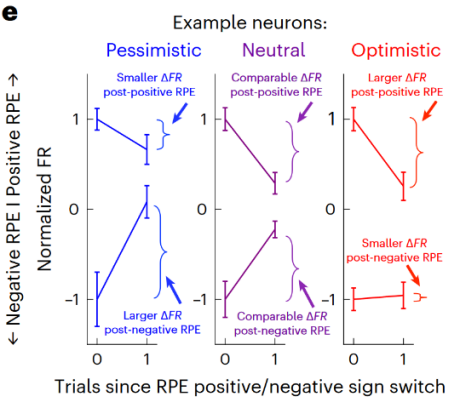
	Scaling	Learning
Symmetric	$FR = \beta_0 + \beta_1\delta$ (Equation (i)) SS	$V \leftarrow V + \alpha\delta$ (Equation (ii)) SL
Asymmetric	$FR = \beta_0 + \beta^+\delta,\ \delta > 0$ $FR = \beta_0 + \beta^-\delta,\ \delta \leq 0$ (Equation (iii)) AS	$V \leftarrow V + \alpha^+\delta,\ \delta > 0$ $V \leftarrow V + \alpha^-\delta,\ \delta \leq 0$ (Equation (iv)) AL

Example model fits to ACC neurons



- Change in state value in dashed black line.
- Asymmetric reactions from different neurons to state transitions.

Echo asymmetric firing rates



- Use the best-fitting model to define trials when the RPE switched from negative to positive, or vice versa.
- Plot the mean FR of the first trial (0) of the switch and the subsequent trial (1).
- Pessimistic neurons react more to negative RPEs while optimistic neurons react more to positive RPEs.

Thank you!

Bibliography I

Bavard, S., Lebreton, M., Khamassi, M., Coricelli, G., & Palminteri, S. (2018). Reference-point centering and range-adaptation enhance human reinforcement learning at the cost of irrational preferences. *Nature communications*, *9*(1), 4503.

Dabney, W., Kurth-Nelson, Z., Uchida, N., Starkweather, C. K., Hassabis, D., Munos, R., & Botvinick, M. (2020). A distributional code for value in dopamine-based reinforcement learning. *Nature*, *577*(7792), 671–675.

Farrugia-Roberts, M., Kruck, N., Premrudeepreechacharn, T., Santhanakrishnan, P., & Yang, S. (n.d.). Expectile-based distributional reinforcement learning and dopamine-associated mental disorders.

Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, *299*(5614), 1898–1902.

Kennerley, S. W., Behrens, T. E., & Wallis, J. D. (2011). Double dissociation of value computations in orbitofrontal and anterior cingulate neurons. *Nature neuroscience*, *14*(12), 1581–1589.

Kolling, N., Wittmann, M., & Rushworth, M. F. (2014). Multiple neural mechanisms of decision making and their competition under changing risk pressure. *Neuron*, *81*(5), 1190–1202.

Miranda, B., Malalasekera, W. N., Behrens, T. E., Dayan, P., & Kennerley, S. W. (2020). Combined model-free and model-sensitive reinforcement learning in non-human primates. *PLoS computational biology*, *16*(6), e1007944.

Bibliography II

Morimura, T., Sugiyama, M., Kashima, H., Hachiya, H., & Tanaka, T. (2012). Parametric return density estimation for reinforcement learning. *arXiv preprint arXiv:1203.3497*.

Muller, T. H., Butler, J. L., Veselic, S., Miranda, B., Wallis, J. D., Dayan, P., ... Kennerley, S. W. (2024). Distributional reinforcement learning in prefrontal cortex. *Nature Neuroscience*, 1–6.

Rescorla, R. A. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. *Classical conditioning, Current research and theory*, *2*, 64–69.

Rigoli, F., Friston, K. J., & Dolan, R. J. (2016). Neural processes mediating contextual influences on human choice behaviour. *Nature communications*, *7*(1), 12416.

Wallis, J. D., & Kennerley, S. W. (2010). Heterogeneous reward signals in prefrontal cortex. *Current opinion in neurobiology*, *20*(2), 191–198.