We thank the reviewers for their time and thorough comments, as well as their valuation of our work including its relevance for NeurIPS. We will fix the typos, expand the references, further separate results from hypotheses, and fill in the details suggested by the reviewers (e.g. switching cost is subtracted from the reward after switch; training for each set of restart values is performed separately). For the larger discussion items, please find the detailed comments below.

Alternative models, experiments and analyses. We agree that several claims about the R-model would be strengthened by including formal comparison across models. To this end, we performed additional manipulations of animal task parameters (inter-trial delays; reward depletion rates; switch costs) and tested non-deep models (port/action-value V/Q-learning). These manipulations support the R-learning model and we briefly mention them in the paper (lines 270-275), but we did not include the corresponding data/details to meet the space constrains and keep the text clarity. Additionally, the reviewers highlighted the importance of quantitative fits. We agree and have performed these analyses that will be included in the revised paper. Briefly, we used gradient descent to minimize the negative log likelihood w.r.t the parameters of the MVT / leaky MVT – the decision rules, as we prove, optimal for the V/R models. We found that, according to the leaky MVT, mice averaged the past reward with the time constant $\kappa = 0.86 \pm 0.04$, perceived the switching cost as $\langle c_{sw} \rangle = 2.3 \pm 1.3 \mu$l, and exhibited the perceptual noise of $\alpha = 0.33 \pm 0.24$. In the revised paper, we will use these parameters in our deep R-model for quantitative similarity between the data and the model. We further used AIC to confirm that the leaky MVT decision rule explains the data better than the MVT ($\Delta AIC = -101 \pm 62$).

Neuronal substrates. We agree that differentiating between the V- and R-models based on TD errors is challenging, nonetheless we argue that consistency between VTA DA and R-learning TD error is an important result in light of the advantages of R-learning. We currently attempt to differentiate between these models using additional manipulations. Note that in R-learning, both expected and unexpected rewards are discounted by the *same* value of an exponentially averaged reward, hence TD error in a converged R-model matches that of the V-model, and is independent from the average reward (lines 234-240). At the same time, theory, model and data all reveal that the exponentially averaged reward varies substantially ($1 - \kappa \approx 0.14$ of reward variation) in contrast with the intuition offered by the reviewer #4.

Optimality of R-learning and Bayesian inference. We argue in the discussion that R-learning behavior can be optimal in the real-world settings. Unlike our tasks with repetitive reward patterns, natural environments are dynamic, i.e. the reward richness may change over time. To behave optimally in such settings, the animals have to keep track of the environmental variables such as the average reward rate. This is exactly what happens in the leaky MVT with its short-term averaging of the reward. The optimal timescales for exponential averaging w.r.t the dynamics of an environment have been previously studied in hidden Markov models using Bayesian inference approach (Ref [29]). It has been suggested that animals, evolutionally exposed to dynamic environments, may have finetuned their decision rules to the average-case environmental dynamics, and use these in fixed experimental settings by extension (Ref [30]). This logic is in line with viewing deep RL networks as meta-learning circuits optimizing the generalized decision rules in their weights (Ref [8]). Although the reviewers are correct to point out that we did not perform Bayesian inference modeling, we connect it to RL via literature on optimality of exponential filtering (cited above / lines 302-320).

Theoretical implications. R-learning may be advantageous for computation. In dynamic environments, R-learning – converging to the leaky MVT decision rule – enables agents to adjust to changes in reward contiguity. Such adjusment does not necessitate costly updates of high-dimensional state values. The R-model allows a simple implementation: exponential averaging may be performed with a recurrent neuron. Such threshold update mechanism may not only be computationally efficient ($\mathcal{O}(n)$), but also has potential of outpefroming more complex methods in time-varying stochastic environments. This is akin to (Sutton, 1992) showing that an $\mathcal{O}(n)$ gain tuning beats the $\mathcal{O}(n^2)$ Kalman filter in dymanic environments, because the latter is only optimal if the model of the world is precise. R-learning also broadens the class of models for DA activity, potentially impacting interpretation of a range of experimental findings.

Relation to foraging and neuroeconomics literature. Our work builds upon results in the field including Ref [2] mentioned by the reviewer #1; we will further clarify this in the discussion. Two key differences are the comparison of a broader class of models and the the choice of species, allowing us to monitor and manipulate neuronal activity. There are critical differences in the reward schedule: Ref [2] manipulates low/high initial reward values in blocks, whereas we draw random initial rewards intermittently. Our design aims to: 1) reduce autocorrelation in reward sequences to enable analyses as logistic regression, and 2) disentangle the impact of the initial and average reward. Each initial reward is preceeded with the same variety of reward sequences – yet the average leaving thresholds are unique for each initial value. This observation enabled us to pursue the hypothesis of the leaky estimate of average reward. Our approach, summarized in the leaky MVT rule, explains overharvesting (Ref [2]) mentioned by the reviewer #1. Specifically, the exponentially averaged reward drops in parallel with the current reward – after which they meet at a lower threshold than predicted by the MVT. The models used in our work are similar to these in (Ref [2]) with the exception of the definition of state = history of the rewards, which, we argue, allowed our models to generalize predictions over various reward schedules. In a revision we will clarify these issues and also discuss the relation to the previous studies including Ref [2], and the prospect theory / reference point literature (e.g. Mobbs et al, 2018; Constantinople et al, 2019).