

Multi-timescale reinforcement learning in the brain

Corresponding Author: Professor Naoshige Uchida

Parts of this Peer Review File have been redacted as indicated to remove third-party material.

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Referee #1

(Remarks to the Author)

This paper presents analyses of mouse dopamine neuron activity and demonstrates that they encode temporal difference errors with different discount factors. The coincidence of observations from two experiments makes the argument convincing.

1) The introduction is quite verbose. Maybe the first paragraph, which doesn't talk about the core issue of temporal discounting, can be omitted and the paper is still as good as it is.

2) The first section presents a didactic example of how a collection of TD learners with different discount factors allows retrieval of reward timing, non-exponential discounting, and so forth, but given what were already demonstrated in the previous works, this section may be better put into the supplementary material and referenced as needed from the experimental sections.

On this topic, the works by Chris Reinke may better referenced:

Reinke C, Uchibe E, Doya K (2017). Average reward optimization with multiple discounting reinforcement learners. International Conference on Neural Information Processing (ICONIP 2017), 10634, 789-800, Springer Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-319-70087-8_81

Reinke C (2020). Time adaptive reinforcement learning. ICLR 2020 workshop: Beyond tabula rasa in RL (BeTR-RL), 10.48550/arXiv.2004.08600, arXiv:2004.08600. <https://doi.org/10.48550/arXiv.2004.08600>

One question about the merit of inverse Laplace approach is why the reward timing needs to be decoded this way, while there are other timing memory mechanisms in the brain.

Further question is that the second experiment suggests that the mouse learns a single value function using multiple TD errors. In this case the inverse Laplace approach doesn't work. Some previous works suggested spatial heterogeneity temporal discounting. In such a case learning multiple value functions may be feasible. In this work, however, no spatial organization was found, so that the dopamine terminals would be largely overlapped, which makes value learning distinct discount factors unlikely.

3) The first experiment is nicely designed to specifically measure the temporal discounting of each neuron. As not all the readers of this journal is RL theory, it would be better to show the TD error equation here, and explain that upon cue presentation, with $r(t) = V(s(t-1)) = 0$, $\delta(t) = \gamma V(s(t))$, so that it is proportional to the discounted value.

Or more precisely, because of the trial start cue, $V(s(t-1))$ should be the average of 4 cases. Figure 3c shows a response at the trial cue. It is interesting to also analyze whether the strength is the average of 4 cases, by further taking into account 1.25 sec to the odor cue.

It is written in the Methods that those neurons are from $n=8$ mice for the first experiment and $n=13$ for the second, but these numbers should be reported in the main text and data should be shown that the discount factors are varied within each mouse, rather than due to collection from different mice.

Is there any way to estimate each mouse's behavioral discount factor in this paradigm? Figure 3b shows the average licking, but for each trial, were there cases when the mouse was disengaged from the task and did not show predictive licking, like the errors in the lever release paradigm by Richmond group?

4) The reanalysis of the VR approach experiment is interesting, but did ramping happen also in the first experiment? If not, what is the reason for the difference? Internal timing versus continuously varying sensory cues?

Figure 4h uses cubic discounting, presumably for its resemblance to the fitted discount functions, but another example with hyperbolic discounting would be interesting.

If I understand correctly, the model assumes the same value function for all 13 mice, but that needs some justification. Is it possible to apply the method to neurons from a single mouse with many neurons recorded?

Again, is there any way of estimating the discount factor of each mouse from the behavior? Did the locomotion speed change with the time to reward?

5) The comparison of discount factors from two tasks is very interesting. Figure 5a shows a regression line with the discount factors in VR task as the regressor, but this would look different if those of the cued task is taken as the regressor. In either way, for checking the equivalence, it is better to show the 45 degree line.

Correlation coefficient may not be a good measure to check the correspondence; it can be high even when two values are in different scales.

What would be interesting is to check the ratio of two discount factors or discount rates, as they may be tuned for the relevant time scale of the task (~10 sec for the cued task and ~4 sec for the second task?)

6) Neural mechanisms: given this interesting observation that each dopamine neuron takes its own discount factor, it is interesting to consider possible circuit/synaptic mechanisms. As gamma is the ratio for the newer/older value functions, is this based on the difference in the balance of the excitatory/inhibitory inputs to the dopamine neurons?

How the recipient neurons/circuits adaptively utilize TD errors or value functions at different time scales is also an interesting question. In this regard, the possible role of serotonin in regulating the time scale of reward prediction may be relevant:

Doya K (2002). Metalearning and Neuromodulation. *Neural Networks*, 15, 495–506. [https://doi.org/10.1016/S0893-6080\(02\)00044-8](https://doi.org/10.1016/S0893-6080(02)00044-8)

Tanaka SC, Schweighofer N, Asahi S, Shishida K, Okamoto Y, Yamawaki S, Doya K (2007). Serotonin differentially regulates short- and long-term prediction of rewards in the ventral and dorsal striatum. *PLoS One*, 2, e1333. <https://doi.org/10.1371/journal.pone.0001333>

Referee #2

(Remarks to the Author)

This study combines computational approaches and analyses of neural data to characterize the temporal discount factors of dopamine neurons. First, the authors characterize the computational benefits of heterogeneous discount factors across dopamine neurons, including being able to separately decode reward magnitude and reward timing. They then characterize the discount factors of dopamine neurons recorded from mice performing a Pavlovian task in which cues predict rewards with different delays. They find that the cue responses reflect exponentially discounted future rewards, with individual neurons showing heterogeneity in the amount of discounting they exhibit. They show that this heterogeneity helps decode reward timing, and also explains complex dopamine ramping dynamics in a different VR task in which mice approach rewards on a linear track. Finally, in neurons that were recorded in both tasks, they show that discount factors are correlated across tasks, which they interpret as being cell-specific.

Overall, this is a technically sound paper that will be of interest to groups studying reinforcement learning, dopamine, and machine learning. The claims are generally well-supported by the data, and the results are a nice conceptual follow-up to their previous work on distributional RL. My main comments are about clarifying some of the results, especially in the computational part of the paper.

Major comments:

1. Regarding the policy learning step in their model, why are they using something as complicated as a DNN to decode the timing of rewards? Please clarify what is gained by using such a complex and uninterpretable model. Is the point that the DNN doesn't know the discount factors, but can still decode reward timing?

2. The authors argue that they can decode reward timing by inverting the discount functions matrix. Is this biologically plausible? Do they really think that the brain is performing this operation? If not, why is this being emphasized throughout the text? Please clarify.

3. I did not immediately realize that the major take-home message of Figure 2f,g is that the maximum of the blue and red curves are different. Making this more explicit in the figure would be helpful for the reader.

4. Figure 4j - Please clarify in the main text or legend the parametric form that was chosen for fitting the value function. The reader should not have to dig through the methods for this.

Minor comments:

Line 84: typo - should be timestep instead of time steps?

Line 337: exponential should be cubic

Line 380: predictions should be prediction

Line 962 - fitted should be fit?

Line 1163 - bootstrap should be bootstraps

Extended Data Figure 1 - the axis labels and tick labels are very tiny, the legends are so small that they are not readable.

Referee #3

(Remarks to the Author)

This is an interesting study whose results, if robust, are important as they suggest a qualitative change in our understanding of the information represented by the dopamine system. The study examines the activity of individual opto-tagged dopamine neurons in mice, during two different tasks; i) odor-reward conditioning using a set of cues which differ with respect to the timing of the predicted reward, and ii) approach to reward along a linear track in virtual reality (VR). The key claims are:

i) There is heterogeneity across the population of VTA dopamine neurons (DANs) in the discount factor used to calculate the value estimates that drive their reward prediction errors (RPEs). The discount factor determines how much rewards further in the future are discounted relative to those that will occur sooner, so this indicates that the relative weighting of expected rewards near and far in the future is different in the RPEs encoded by different dopamine neurons. This heterogeneity of discount factors removes the ambiguity inherent in standard discounted reward RL about whether a given value indicates a small reward soon or a larger reward later, causing the DAN population activity to carry information not just about how much discounted future reward is expected, but when in time it is expected to occur. This is conceptually related to recent proposals (Dabney et al. Nature 2020) that dopamine neurons encode not just average expected future reward but rather the distribution over expected reward magnitudes. The current work extends this it by proposing that the dopamine neurons carry information about the expected distribution of future rewards over time.

ii) Such distributional coding of when future reward is expected can improve the performance of reinforcement learning (RL) agents.

iii) Heterogeneity in the discount factor across dopamine neurons can explain heterogeneity in ramping of their activity during reward approach in VR.

iv) Discount factors of individual DANs estimated from the two tasks are correlated across the DAN population, indicating the discount factor of individual neurons is a cell specific property (or a property of the specific subnetwork it is embedded in), not a consequence of learning on a specific task.

The authors use a complementary set of analyses to support these claims, including directly assessing heterogeneity in discount rates from the response to cues predicting rewards at different delays, decoding analyses showing they can extract information about when reward is expected from the dopamine activity at the time of the cue, and computational modelling of how heterogeneity in discount rates may affect dopamine ramping, enabling estimation of discount rates in the VR task. This combination of strong theoretical motivation, technically demanding experiments, and sophisticated and creative analyses are real strengths of the study. However, I have some concerns that preclude full confidence in the validity of the claims at this point.

- The authors aggregate neurons across subjects without taking into account possible inter-subject variability – i.e. they assume fixed rather than random effects across subjects. This is never ideal (though remains common practice in animal neuroscience), as formally the statistical conclusions then only hold for these particular subjects not the population from which they were drawn. However, it is particularly problematic in this study because the key claim rests on there being heterogeneity across the population of neurons in their discount rates. If this heterogeneity is in fact across subjects (or sessions), not across neurons within a given subject at a given time, then the claim does not hold. To address this concern the authors could: 1) Provide versions of the plots showing heterogeneity in discount rates (figure 3E, 4K, Extended data figure 4) with the neurons color coded by subject. 2) Demonstrate that statistically significant heterogeneity in discount rates is present after controlling for possible systematic variation between subjects (and ideally sessions). Additionally, to help readers understand the strength of the evidence given inevitable cross subject variability, and how likely the results are to hold at the population level: 1). Specify the number of neurons recorded per subject and session. 2) Specify what the experimental unit is (e.g. neuron, trial, session, or subject) whenever reporting stats (including bootstrap tests) or confidence intervals. 3) Provide sufficient information for readers to assess how consistent the results are across subjects, e.g. by showing key results separately for individual subjects or providing cross-subject uncertainty measures (e.g. by bootstrap resampling subjects).

- If I have understood the author's arguments correctly there appears to be an incompatibility between the account of heterogeneity in cue evoked dopamine responses in the odor task and that of heterogeneity in ramping in the VR task. In the odor task, my understanding is that heterogeneity in the response to a given cue across DANs is proposed to be due to different DANs computing RPEs using cue value estimates learned using different discount factors – i.e. different DANs use different value estimates to compute RPE. However, the account of heterogeneity in the VR experiment presented in figure 4 assumes that RPEs are computed using the same value estimate for all DANs, but different discount rates across DANs in the RPE calculation. Though the authors do not directly allude to this apparent incompatibility in the text, they do present an alternative account of ramping heterogeneity in the VR experiment in extended data figure 8, in which different DANs use different value estimates to compute RPEs, each computed using different discount rates. However, this account makes the unusual assumption that the value estimates that drive dopamine responses are computed using an estimate of the expected timing of future reward. This is very different from standard RL accounts of basal ganglia in which the value estimates that drive dopamine responses are learned using temporal difference RL, and is also different from the author's previous account of dopamine ramping in Kim et al (Cell 2020) which assume uncertainty about the current state not the timing of future reward. It is also notable that the distribution of discount rates inferred using this alternative model (extended data figure 8F) is qualitatively different from that in the odor task (figure 3G). To address these concerns the authors could: i) Be explicit both in the main text and through diagrams in the figures about the conceptual models underlying their various accounts of experimental data, making clear where they imagine parallel channels carrying distributional codes for value, RPE etc and where they assume single common signals. ii) Ensure that their accounts of the VR and odor tasks are consistent with each other in their assumptions. iii) Ideally show that heterogeneity in discount rates can account for heterogeneity in ramping without making unusual and hard to justify assumptions about where the value signals that drive the RPEs come from, e.g. show that similar heterogeneity in dopamine ramping emerges when values are learned using TDRL with heterogeneity in discount factors across neurons.

- The arguments presented in figure 2F,G do not make a particularly convincing argument that heterogeneity in discount rates leads to better performance in RL agents. Figure 2F is a rather contrived situation in which a noisy reward function in most states combined with carefully placed high reward states causes value estimates learned using either high or low discount rates to work better at specific decision points. It is not obvious how relevant this is to biological learning as the assumption of noisy reward in most states does not seem typical of natural environments. Additionally to take advantage of the additional information learned using multiple discount factors the agent has to employ meta learning across many different task instances, and it is not obvious that this approach would generalise to natural tasks. The deep-Q network learning to play lunar lander (2G) suggests that different discount rates learn better value functions near and far from the goal, but as far as I could see does not demonstrate any actual performance difference between agents with and without a distribution of discount factors. The paper might benefit from focusing more on the core experimental findings and deemphasising these arguments.

Other points:

- Please quantify how consistent the heterogeneous ramping profiles in the VR task are across split halves of trials to demonstrate that this is genuine cross neuron heterogeneity rather than noisy estimation.

- Please quantify how well the ramping models with heterogeneous discount factors explain neural activity as compared to models with a single discount factor (e.g. using cross validated variance explained on split halves of the trials).

- It was not clear to me how the P values for the decoding accuracy of reward times (line 294) were calculated. Was this comparing the decoding accuracy with that obtained when the neuron identities are shuffled?

- In figure 3K would be useful to show the decoding obtained under the assumption that there is a single discount rate across neurons, as this seems like a better control than that obtained by shuffling the cue responses across neurons given the different baselines and gains of the different neurons. The stats for this comparison are presented in the main text but it would be useful to visualise the decoding accuracy under this important null model.

- Line 317: It is not clear what heterogeneity is being referred to in the sentence 'Here we show that some of this heterogeneity can be understood...'

- Line 947, why was this modified z score used rather than a standard z score?

Version 1:

Reviewer comments:

Referee #1

(Remarks to the Author)

I appreciate the authors' extensive revisions and explanations in response to my comments, especially the new analysis of licking behaviors. I think the paper can be published in the present form.

Referee #2

(Remarks to the Author)

The authors addressed my comments.

Referee #3

(Remarks to the Author)

The authors have addressed most of my concerns and I continue to think that this is an important and interesting work which will be of wide interest to the neuroscience community.

The one issue which I think is not well handled in the current manuscript is that of the different and incompatible assumptions in the account of the Pavlovian task presented in figure 2 and that of the VR task in figure 3. Specifically, the account of dopamine responses to the cue in figure 2 requires that separate value signals are learned using different discount factors, while the account of the heterogeneous ramping in figure 3 assumes a common value signal shared by all dopamine neurons. The authors have added a new section to the methods, 'Mixing in distributed RL models', discussing this issue, and I broadly buy their argument that partial mixing of value signals across channels with different discount rates could account for both datasets. However, I think that there remain some issues with how this is addressed in the manuscript:

- The manuscript is not straightforward in the main text about the different assumptions used in figure 2 and 3. The new methods section is motivated in the text as a way of reconciling two different models of the ramping activity (Figure 3 and ED fig 10), rather than reconciling the different models underlying figure 2 and 3 which is the more fundamental issue. Please make clear in the main results text that figure 2 assumes separate value signals for channels with different discount factors, while figure 3 assumes a single common value signal. Please also add the diagram from figure ED 12a to figure 2 to illustrate the parallel value channels needed for this account to work. The diagram presented in figure 3C also appears not to reflect the true structure of the common value signal model as it shows projections from each value unit to its corresponding prediction error unit, which do not exist in the common value signal model. Please update this diagram to correctly represent the structure of the model. It is OK to use models that make different simplifying assumptions in figure 2 and 3 for didactic reasons, and then reconcile them later through the partial mixing of value signals account, but it is important to be transparent with the reader about what you're doing to avoid confusion.
- I found the account of mixing of value signals between channels with different discount factors unclear in the current text. At line 1314 the motivation for some mixing of value between channels is motivated by anatomical considerations, and this seems reasonable. However, at line 1359 it is argued that in the absence of feedback about reward timing the channels are entirely decoupled, which does not appear compatible with mixing being a consequence of anatomical crosstalk. Rather than arguing for no value mixing in the Pavlovian task and value mixing in the VR task, a simpler way to reconcile the two datasets would seem to be to simulate a model with partial value mixing between channels on both tasks, and show that it reproduces the key qualitative features of the data. I'm not suggesting changing figures 2 and 3, or attempting to fit this partial mixing model to the data, but rather just to show via simulation that a single model can qualitatively reproduce both datasets.
- In equation 32 it is not clear why mixing between channels should occur for the value of the new state S_{t+1} but not for the old state S_t . Please explain in the manuscript the rationale for this assumption and whether it is important for the behaviour of the model. Also, my understanding is that the common value signal model (equation 25) does not share this assumption as it uses a single common value signal for both the new state and old state. If this is correct why the different assumptions in these two models?
- When describing figure 3 the text uses language suggesting that dopaminergic reward prediction errors contribute to learning a world model: Line 317 'we first assume that heterogeneous dopamine neurons contribute to learning a common model of the environment and therefore share a common value function.' Line 349: 'prediction errors are pooled to create a single value function and world model'. Dopaminergic RPEs are typically thought to be used for learning values or policies, rather than learning world models, and the conceptual model underlying figure 3 appears compatible with model-free reinforcement learning of state values. Discussing a world model here therefore seems an unnecessary complication, likely to confuse readers. If the authors feel that model-based RL is conceptually important in explaining these data then the reason for this should be made clear, otherwise I suggest removing these references to world models.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Our responses and comments are written in black while the original reviewer comments are in blue.

Summary of response to reviewers:

We thank the reviewers for their insightful comments on our manuscript. We greatly appreciate their recognition on the importance of our results and their opinion that our contribution would change our view of dopamine signaling in the brain. In the following, we will address the matters raised by the reviewers point by point. Here we provide a summary of the major changes:

- Following the reviewers' comments, we have reworked our presentation of the computational results and combined them into a single figure (Figure 1) in the main manuscript and clarified the presentation of the concepts in Extended Data Fig 1-3.
- An important point raised by several reviewers was whether some of our observations arises from variability across animals rather than across neurons. We have added a number of analyses to better present the variability within and across subjects and to show that the variability in discount factors across dopamine neurons is not an effect of polling across animals. In addition, we repeat our analyses for two animals in which we had sufficient numbers of neurons to perform population level analysis and show that our key results hold within single animals.
- Another related point raised by the reviewers is the relationship between the neural discounting measured in dopamine neurons and behavioral discounting. We added a number of analyses both in the cued delay task and the VR task to show that behavior (measured through licking) cannot account for the variability we observed at the neural level.
- We have also clarified the relationship between the models used for fitting the neural data in the two tasks and added derivations and simulations to show how they are part of a broader conceptual framework.

Overall, the constructive reviews have helped us strengthen our paper which, as the reviewers, we believe is an important contribution to the understanding of functional heterogeneity of dopamine signaling in the midbrain.

Referee #1 (Remarks to the Author):

This paper presents analyses of mouse dopamine neuron activity and demonstrates that they encode temporal difference errors with different discount factors. The coincidence of observations from two experiments makes the argument convincing.

1) The introduction is quite verbose. Maybe the first paragraph, which doesn't talk about the core issue of temporal discounting, can be omitted and the paper is still as good as it is.

We have shortened the introduction although it is not a drastic reduction. As this is for a general audience, we feel it would be difficult to jump too rapidly into a discussion of the role of the discount factor without introducing key concepts about reinforcement learning. Nonetheless, if the reviewer feels strongly, we would be happy to shorten further.

2) The first section presents a didactic example of how a collection of TD learners with different discount factors allows retrieval of reward timing, non-exponential discounting, and so forth, but given what were already demonstrated in the previous works, this section may be better put into the supplementary material and referenced as needed from the experimental sections.

On this topic, the works by Chris Reinke may better referenced:

Reinke C, Uchibe E, Doya K (2017). Average reward optimization with multiple discounting reinforcement learners. International Conference on Neural Information Processing (ICONIP 2017), 10634, 789-800, Springer Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-319-70087-8_81

Reinke C (2020). Time adaptive reinforcement learning. ICLR 2020 workshop: Beyond tabula rasa in RL (BeTR-RL), 10.48550/arXiv.2004.08600, arXiv:2004.08600. <https://doi.org/10.48550/arXiv.2004.08600>

One question about the merit of inverse Laplace approach is why the reward timing needs to be decoded this way, while there are other timing memory mechanisms in the brain.

We have reworked the first section, merged the first and second figures into a single figure, and moved several of the panels to the supplementary materials. We thank the reviewer for pointing us to this extremely relevant literature that we now cite.

The Laplace transform simply proves that the temporal information is present in the multi-gamma array. The DNN in Figure 2b (now Figure 1d) shows that decoding could be performed in a general, code-agnostic manner. But the experimental decoding shows that the code is compatible with the Laplace code (and possibly with other codes too). However, as suggested by our analysis and the work by Reinke, there are also other ways to exploit the multi-gamma information.

Further question is that the second experiment suggests that the mouse learns a single value function using multiple TD errors. In this case the inverse Laplace approach doesn't work. Some previous works suggested spatial heterogeneity temporal discounting. In such a case learning multiple value functions may be feasible. In this work, however, no spatial organization was found, so that the dopamine terminals would

be largely overlapped, which makes value learning distinct discount factors unlikely.

The reviewer points to an extremely interesting open question. Previous work has shown that there might be some gradient of discount factor across the scale of the entire striatum (Enomoto, Kazuki, et al. "Topographic distinction in long-term value signals between presumed dopamine neurons and presumed striatal projection neurons in behaving monkeys." *Scientific Reports* 10.1 (2020): 8912.

<https://doi.org/10.1038/s41598-020-65914-0> and Mohebi, Ali, et al. "Dopamine transients follow a striatal gradient of reward time horizons." *Nature Neuroscience* 27.4 (2024): 737-746. <https://doi.org/10.1038/s41593-023-01566-3>). Here, we focused our analysis at the single neuron level and in the VTA and did not find evidence for a local gradient within VTA. Our interpretation is that even though there might be some gradient at the large scale or at the level of the axonal projection target (i.e. striatum), there are representations at multiple timescales within each subregion. Previous work in distributional RL (Dabney, W., Kurth-Nelson, Z., Uchida, N. et al. A distributional code for value in dopamine-based reinforcement learning. *Nature* 577, 671–675 (2020). <https://doi.org/10.1038/s41586-019-1924-6>), has shown that an imprecise gradient can support a distributional representation. A similar effect would allow learning multiple timescales. Furthermore, although the dopamine release in striatum is not constrained to a synapse, it still has limited local diffusion. This observation, coupled with the dendritic structure of medium spiny neurons, can allow for learning multiple value loops Wörnberg, Emil, and Arvind Kumar. "Feasibility of dopamine as a vector-valued feedback signal in the basal ganglia." *Proceedings of the National Academy of Sciences* 120.32 (2023): e2221994120. <https://doi.org/10.1073/pnas.2221994120>).

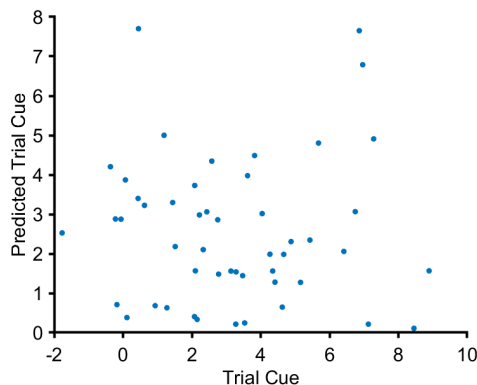
With respect to the difference between the single value function model in the second experiment, we point the reviewer to the new section in the Methods 'Mixing in distributed RL models' and Extended Data Fig. 12 which clarify the relationship between the models used for analysis in the first and second experiments.

3) The first experiment is nicely designed to specifically measure the temporal discounting of each neuron. As not all the readers of this journal is RL theory, it would be better to show the TD error equation here, and explain that upon cue presentation, with $r(t) = V(s(t-1)) = 0$, $\delta(t) = \gamma V(s(t))$, so that it is proportional to the discounted value.

Thank you for the suggestion. We updated the main text to include the TD error equation (Eq. 2).

Or more precisely, because of the trial start cue, $V(s(t-1))$ should be the average of 4 cases. Figure 3c shows a response at the trial cue. It is interesting to also analyze whether the strength is the average of 4 cases, by further taking into account 1.25 sec to the odor cue.

We performed the analysis and as shown below we do not observe a correlation ($r=-0.08$, $p=0.57$, Spearman rank correlation). This is not unexpected as the trial cue response carries a strong salience response (its timing is not predictable). This salience response is not governed by the value of the future reward (Schultz, Wolfram. "Dopamine reward prediction-error signaling: a two-component response." *Nature reviews neuroscience* 17.3 (2016): 183-195. <https://doi.org/10.1038/nrn.2015.26>). This is exactly why we introduced the trial cue so the reward delay predicting cue would be driven by the value response as much as possible. We have not included this analysis in the revised manuscript but would be happy to do so if the reviewer deems it necessary.



It is written in the Methods that those neurons are from $n=8$ mice for the first experiment and $n=13$ for the second, but these numbers should be reported in the main text and data should be shown that the discount factors are varied within each mouse, rather than due to collection from different mice.

We now report the numbers in the main text, and we have labeled with different colors the neurons from the different animals in Figure 2g, 3k and Extended Data Fig 4. We have also reported the distribution of neurons across tasks in Extended Data Table 1. Furthermore, we performed a number of analyses for the two animals (m3044 and m3054) for which we had a sufficient number of neurons. Together, these strongly suggest that the observed results are consistent with a distribution of discount factor across dopamine neurons rather than an across average effect.

Is there any way to estimate each mouse's behavioral discount factor in this paradigm?

We thank the reviewer for suggesting a more careful analysis of the relationship between behavior and neural activity.

Here we used a Pavlovian conditioning paradigm in which we did not explicitly measure the behavioral discounting of the animals. We nevertheless can use the average anticipatory lick rate after the odor cue to estimate the relative value of each cue and therefore the behavioral discounting. We can estimate the lick rate in the 600ms

following the odor cue and fit the reduction in lick rate to an exponential function, similarly to the firing rate (We baseline-subtract the lick rate as for the firing rate).

To avoid effects due to average differences in behavior at the across-animal level, we performed these analyses at the single animal level, for the two animals with sufficient amounts of neurons (m3044 and m3054). For plotting purposes, we normalize for each neuron the lick rate to the maximum lick rate across the 4 delays. In the heatmaps below (panels a,d), we sort the neurons by behavioral discounting (discount factor from fitting the lick rate) for both the lick rate (left panel) and the delay cue firing rate (right panels).

We do not observe a correlation between the behavioral discount factor and the neural discount factor (panels b,e). We also performed the decoding at the single animal level and show that the single animal discount matrices from panels c,f allows for decoding of subjective reward timing (panel g).

We added these panels in Extended Data Fig. 6a-g and a section ‘Analyzing behavioral discounting through the lick rate in the cued delay task.’ in the Methods.

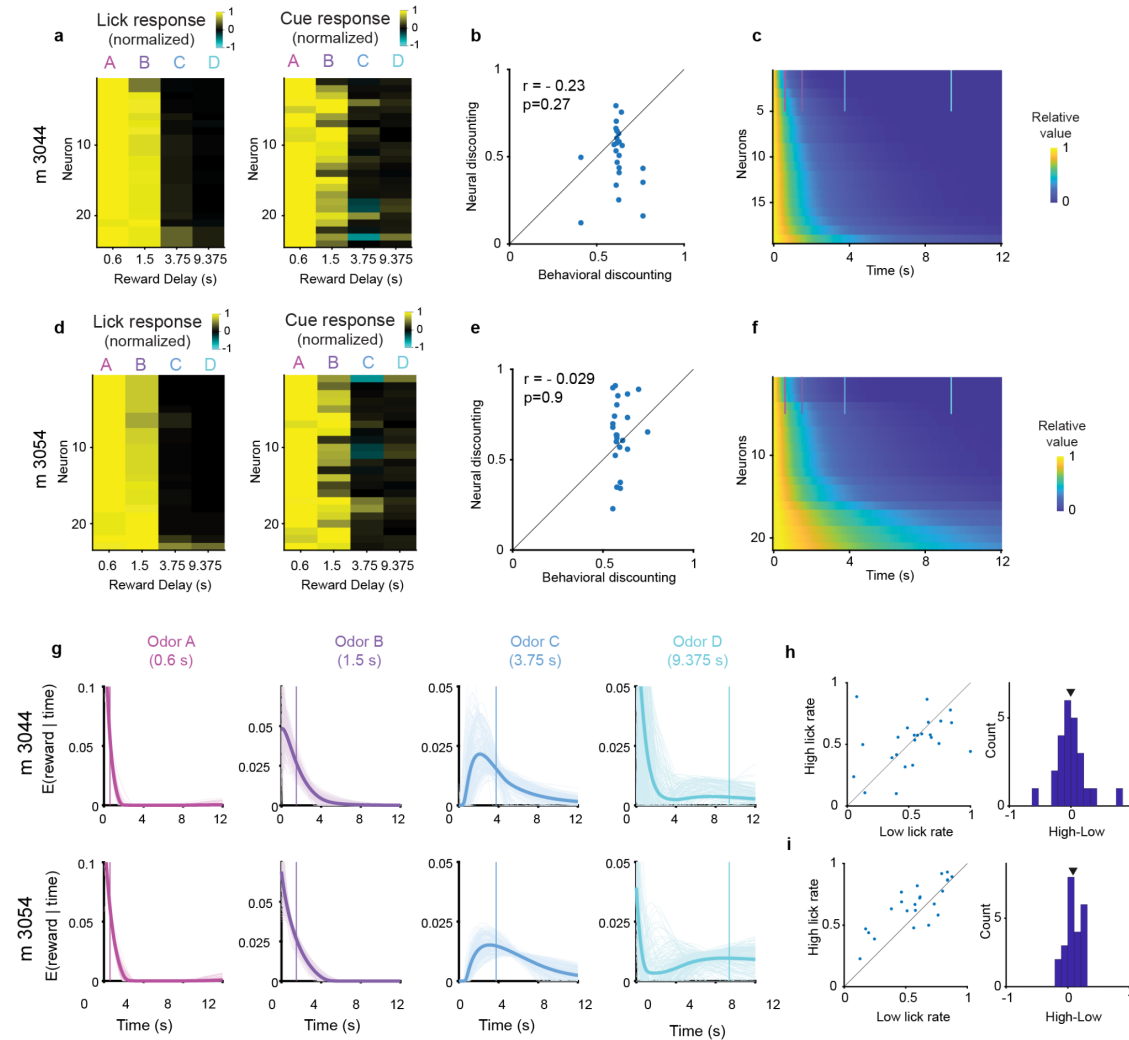
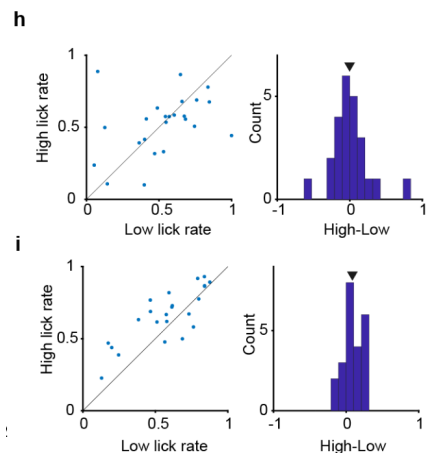


Figure 3b shows the average licking, but for each trial, were there cases when the mouse was disengaged from the task and did not show predictive licking, like the errors in the lever release paradigm by Richmond group?

The sessions were relatively short, which limited the number of trials but avoided disengagement on the task. To further test the contribution of anticipatory licking and inspired by the analysis in Sousa et al., 2023, bioRxiv (<https://doi.org/10.1101/2023.11.12.566727>). We tested the different inferred discount factor as a function of the anticipatory lick rate. If the animals were fully disengaged when lick rate was low, we would expect a difference in the RPE coding upon cue presentation.

For each trial type (reward delay), we computed the median anticipatory lick rate over the entire anticipation period and divided trials according to whether the lick rate was smaller than or superior or equal than the median lick rate for the trial delay. We then fitted the exponential discounting model to each neuron separately on the low and the high lick rate trials. To avoid across animal effects, we performed this analysis on the two animals with a sufficient number of neurons. We observed a significant effect for one animal (m3054, mean = 0.086, $p = 0.0062$, two-tailed t-test, panel i) and a non-significant effect for the other (m3044, mean = - 0.0024, $p = 0.96$, two-tailed t-test, panel h). We note that regardless of whether any effect would be significant with a larger dataset, the potential modulation by behavior of the discount factor within neuron is small in comparison to the heterogeneity of discount factors inferred across neurons (s.d. = 0.19 across neurons in m3054). This small modulation is consistent with the effect presented in Figure 3 of Sousa et al., 2023, bioRxiv (<https://doi.org/10.1101/2023.11.12.566727>). This effect could correspond to fluctuations in the value estimates, with higher value corresponding to higher lick rate and higher measured discount factors. Nevertheless, these modulations are small in comparison to the diversity across single neurons. It will be an interesting avenue to explore in future work where these fluctuations are manipulated explicitly in the behavioral task.



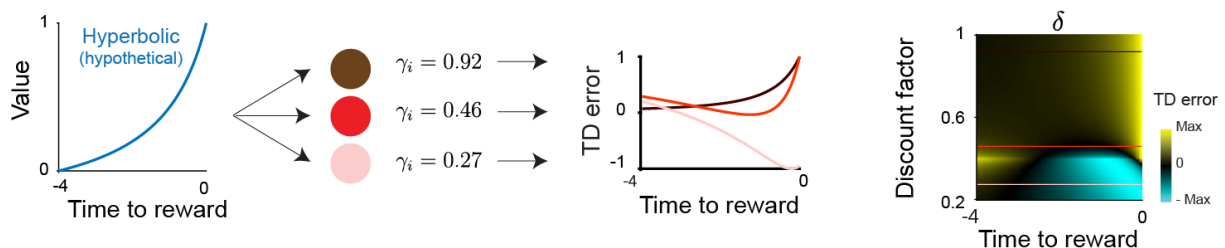
We have included these results in Extended Data Fig 6h,i. and a section titled 'Comparing the neural discount factor as a function of lick rate' in the Methods.

4) The reanalysis of the VR approach experiment is interesting, but did ramping happen also in the first experiment? If not, what is the reason for the difference? Internal timing versus continuously varying sensory cues?

Ramping does not occur in the cued delayed task (see Figure 2c). The reason for this difference is, as the reviewer suggests, the difference in the information related to reward timing that is available in both tasks. In the cued delay task, the mice have to rely on their internal estimate of time whose precision degrades with time elapsed (scalar timing). In contrast, in the VR task the uncertainty of reward timing reduces as the mice approach the reward time due to visual feedback in the VR maze. Previous work from the Uchida group has proposed a theory of when dopamine ramps occur or not based on this difference in state uncertainty and tested a non-trivial prediction of the theory and the result (the generation of dopamine “bump”) supported it (see Mikhael, John G., et al. "The role of state uncertainty in the dynamics of dopamine." *Current Biology* 32.5 (2022): 1077-1087. <https://doi.org/10.1016/j.cub.2022.01.025>).

Figure 4h uses cubic discounting, presumably for its resemblance to the fitted discount functions, but another example with hyperbolic discounting would be interesting.

We thank the reviewer for their interesting suggestion. We have added the following panel in Extended Figure 8e, a similar plot but for hyperbolic discounting. As the reviewer pointed out, the cubic value function is more similar to the fitted discount function. In the hyperbolic case, the slope near reward delivery increases faster than a similar exponential function, leading to different behavior closer to reward delivery. In the hyperbolic case, more neurons will ramp closer to reward delivery while the opposite occurs in the cub case. We therefore would prefer to keep the cubic value function in the main figure and add the following panels in Extended Data Figures.



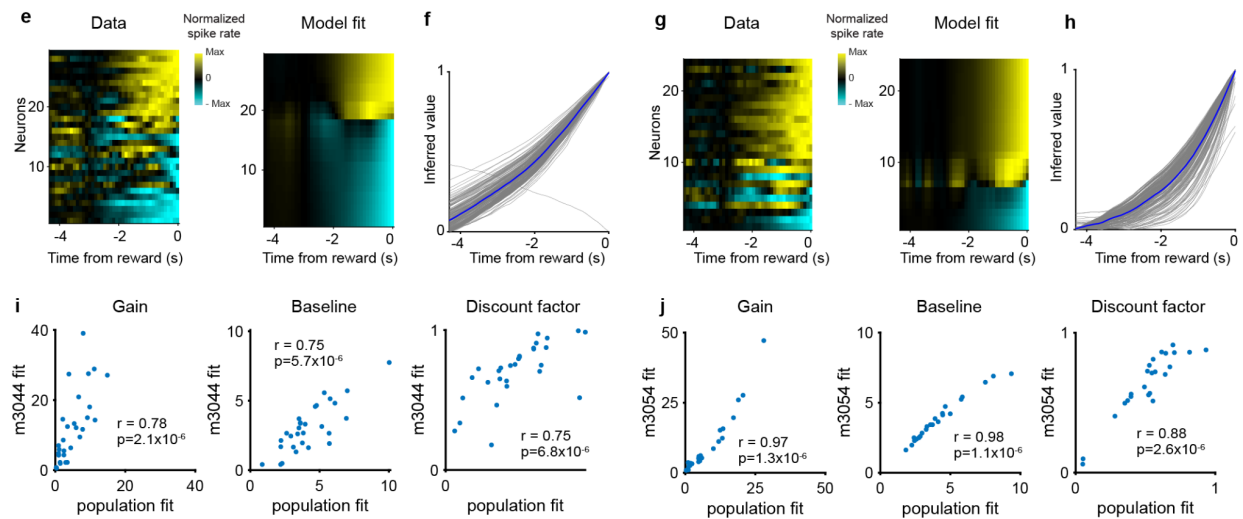
e, Intuition for diversity of ramping with a hyperbolic value function. Agents with a small discount factor exhibit a monotonic downward ramp (pink), while those with a large discount factor exhibit a monotonic upward ramp (brown). Agents with an intermediary discount factor tend to exhibit a downward then upward ramp. The hyperbolic value function gets increasingly convex as the reward approaches, so an increasing fraction of the agents have a positive prediction error as they approach the reward.

If I understand correctly, the model assumes the same value function for all 13 mice, but that needs some justification. Is it possible to apply the method to neurons from a single

mouse with many neurons recorded?

The reviewer is correct that in the analysis presented, we have grouped together all the neurons from the 13 animals and estimated a common value function across all neurons. Two of the animals had a sufficient number of neurons recorded, allowing us to perform the analysis at the single animal level.

The results of the single animal analyses (m3044, panels e,f,i and m3054, panels g,h,,j) are extremely similar to those in the grouped analysis. Importantly, there is a strong correlation across neurons between the parameter values estimated in the single animal analysis and the global analysis (panels i,j). These results suggest that the conclusions with a common value function hold when analyzed at the single animal level.



We have added these panels in Extended Data Fig. 9.

Again, is there any way of estimating the discount factor of each mouse from the behavior? Did the locomotion speed change with the time to reward?

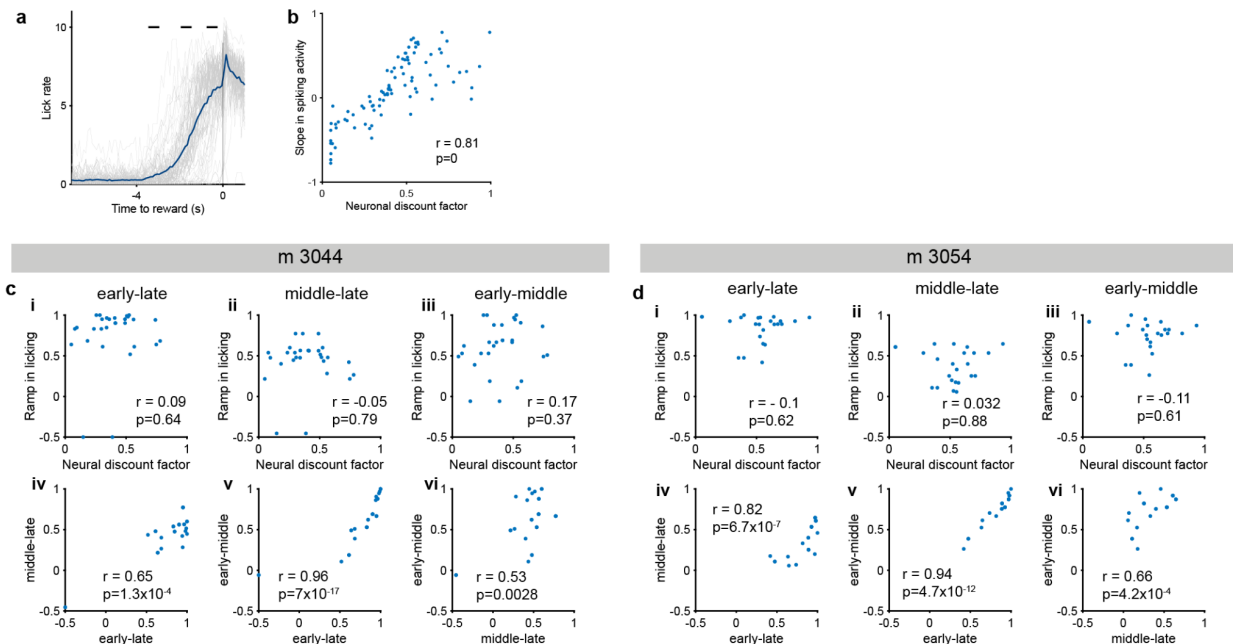
We thank the reviewer again for suggesting these analyses. We analyzed licking data from the two animals with a sufficient number of neurons to compare the properties of behavioral licking and neural activity.

The mice exhibit anticipatory licking prior to the reward delivery and there is some variability across sessions and animals (panel a). We therefore analyzed whether this variability could explain the differences in the inferred discount factor.

The dynamics of anticipatory licking are different from those of dopamine neuron firing. To compare their differences, we measure the slope of licking activity prior to reward delivery as a proxy for discounting (similarly to Figure 3b for neural activity). We note that in the case of dopamine firing, there is a strong correlation between the slope of

neural activity and the inferred discount factor ($r = 0.81$, $p = 0$, Spearman rank correlation, common value function model, panel b).

To ensure robustness of the analysis, we used three measures of the slope for anticipatory licking, using the period indicated by black marks in panel a. The three behavioral measures are significantly correlated to each other (subpanels iv-vi) but they are not correlated to the discount factor inferred from neural activity (subpanels i-iii). These results suggest that behavior does not drive the diversity of discount factors observed at the neural level.



We have included these results in Extended Data Figure 8a-d. and added a section titled 'Analyzing behavioral discounting through the lick rate in the VR task.' in the Methods.

We used licking as our measure of behavioral discounting as although the locomotion speed does change during the trial, the previous study from the Uchida group (Kim et al., Cell, 2020, <https://doi.org/10.1016/j.cell.2020.11.013>) had analyzed the locomotion speed change and showed that the modulation was not predictive of dopamine activity (we show below Figure 2 I-L from Kim et al., Cell, 2020). Panel L (left) shows that the speed of the visual scene rather than locomotion speed affects the ramp in dopamine signaling (the visual feedback about reward timing is what induces the ramps as discussed above).

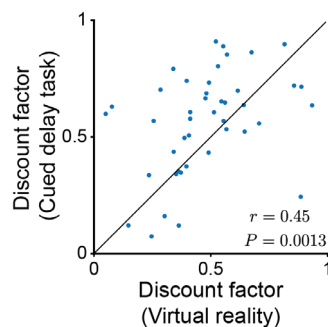
[FIGURE REDACTED]

Note that all the experiments presented here are in ‘open loop’ (the locomotion does not affect the speed of the visual scene), but similar results were obtained in a ‘closed loop’, operant setting (see Figure S7 in Kim et al., Cell, 2020).

Together, these results suggest that behavior is not the driver of the heterogeneity of discount factors measured across dopamine neurons.

5) The comparison of discount factors from two tasks is very interesting. Figure 5a shows a regression line with the discount factors in VR task as the regressor, but this would look different if those of the cued task is taken as the regressor. In either way, for checking the equivalence, it is better to show the 45 degree line.

We had added the regressor and uncertainty cone to help with visualization, but we can replace it with the 45-degree line instead if the reviewer thinks it is preferable (see the updated panel below now used in the Figure). We agree that the regression might not convey the quantitative relationship optimally, which is why we have used the Spearman rank correlation to quantify the relationship (see answer to next point).

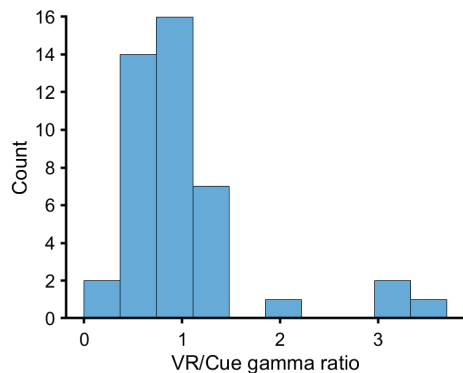


Correlation coefficient may not be a good measure to check the correspondence; it can be high even when two values are in different scales.

We thank the reviewer for their comment. We actually reported the rank correlation coefficient (Spearman) and clarified the wording in the text. We are happy to report either the Pearson correlation ($r=0.39$, $p=0.0047$) or Spearman correlation ($r=0.45$, $p=0.0013$). We can alternatively correlate the discount rate ($\lambda = -\log(\gamma)$), the rank correlation is unchanged (it is just a non-linear rescaling) and the Pearson correlation is significant ($r=0.32$, $p=0.018$).

What would be interesting is to check the ratio of two discount factors or discount rates, as they may be tuned for the relevant time scale of the task (~10 sec for the cued task and ~4 sec for the second task?)

We show below the ratio of the discount factors inferred in VR and Cued tasks. The average ratio across neurons is $d=1.02$. At the population level at least, we cannot identify an effect. We'll note that in the cued task, 3 of the cues are also within a ~4s timescale.



We also note that in the absence of explicit manipulation of the range of timescales it can be hard to interpret the absolute value of the discount factor. In the VR task, the relative shape of the ramping response is driven by the relative value of the discount factor, but the absolute value can depend on the maximum value of the value function and the gain. For example, although the scaling of discount factors in the common value function vs common reward uncertainty model fits is different, the rank correlation across neurons between the inferred discount factors is very high (Spearman's rank correlation, $\rho = 0.93$, $P < 1.0 \times 10^{-20}$, see Extended Data Fig. 10e). It will be interesting in future work to explicitly manipulate the timescales of tasks to test whether the timescale in single neurons also adapt. As observed in Sousa et al., 2023, BiorXiv (Fig. 5F-G, <https://doi.org/10.1101/2023.11.12.566727>), we expect this effect to be small relative to the diversity in discount factors across neurons.

6) Neural mechanisms: given this interesting observation that each dopamine neuron takes its own discount factor, it is interesting to consider possible circuit/synaptic mechanisms. As gamma is the ratio for the newer/older value functions, is this based on the difference in the balance of the excitatory/inhibitory inputs to the dopamine neurons?

This is a very interesting question. As pointed out by the reviewer, we note that in the TD error (whether written in its discrete form: $\gamma V(t+1) - V(t)$ or in the continuous form: $dV(t)/dt + \ln \gamma V(t)$), the contribution of the discount factor can be seen as a weight between two quantities, future and current value in the discrete case; and the derivative of value and value in the continuous case. As the reviewer suggests, it is easy to conceptualize a model in which these correspond to shared inputs onto dopamine neurons that each neuron weighs differentially, leading to a diversity of observed discount factors. However, clarifying the biological underpinning of discount factors requires completely different types of experiments to answer. While it is an important question, we believe that this is beyond the scope of the present study.

How the recipient neurons/circuits adaptively utilize TD errors or value functions at different time scales is also an interesting question. In this regard, the possible role of serotonin in regulating the time scale of reward prediction may be relevant:

Doya K (2002). Metalearning and Neuromodulation. *Neural Networks*, 15, 495–506.
[https://doi.org/10.1016/S0893-6080\(02\)00044-8](https://doi.org/10.1016/S0893-6080(02)00044-8)

Tanaka SC, Schweighofer N, Asahi S, Shishida K, Okamoto Y, Yamawaki S, Doya K (2007). Serotonin differentially regulates short- and long-term prediction of rewards in the ventral and dorsal striatum. *PLoS One*, 2, e1333.
<https://doi.org/10.1371/journal.pone.0001333>

We thank the reviewer for highlighting the literature on the potential role of serotonin in regulating the timescale of reward prediction. We added a sentence in the discussion to highlight these works.

“Our results provide evidence supporting the third model but the recruitment mechanisms of the neurons to adapt the global discounting function with task or context and the link between anatomical location and discounting (Enomoto et al, 2020) and the contribution of other neuromodulators, such as serotonin (Doya 2002, Tanaka et al. 2007), to this adaptation remain open questions.”

Referee #2 (Remarks to the Author):

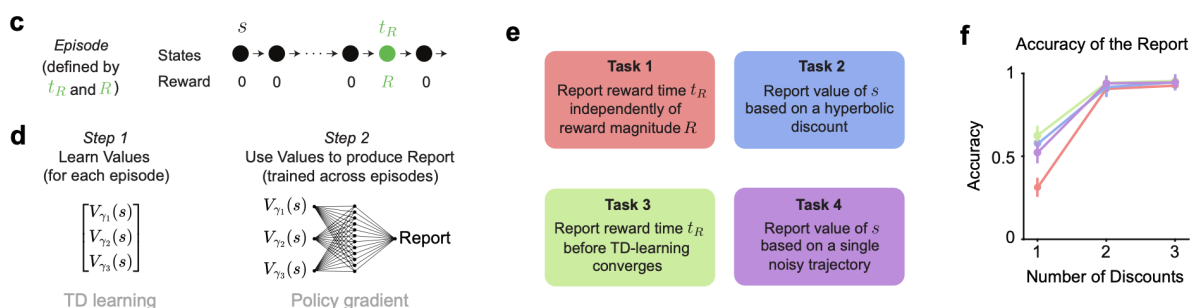
This study combines computational approaches and analyses of neural data to characterize the temporal discount factors of dopamine neurons. First, the authors characterize the computational benefits of heterogeneous discount factors across dopamine neurons, including being able to separably decode reward magnitude and reward timing. They then characterize the discount factors of dopamine neurons recorded from mice performing a Pavlovian task in which cues predict rewards with different delays. They find that the cue responses reflect exponentially discounted future rewards, with individual neurons showing heterogeneity in the amount of discounting they exhibit. They show that this heterogeneity helps decode reward timing, and also explains complex dopamine ramping dynamics in a different VR task in which mice approach rewards on a linear track. Finally, in neurons that were recorded in both tasks, they show that discount factors are correlated across tasks, which they interpret as being cell-specific.

Overall, this is a technically sound paper that will be of interest to groups studying reinforcement learning, dopamine, and machine learning. The claims are generally well-supported by the data, and the results are a nice conceptual follow-up to their previous work on distributional RL. My main comments are about clarifying some of the results, especially in the computational part of the paper.

Major comments:

1. Regarding the policy learning step in their model, why are they using something as complicated as a DNN to decode the timing of rewards? Please clarify what is gained by using such a complex and uninterpretable model. Is the point that the DNN doesn't know the discount factors, but can still decode reward timing?

We thank the reviewer for this insight. We have performed additional analysis and found that using a two-layer DNN was indeed unnecessarily complex for the decoding tasks (Tasks 1 to 4 in the new Fig. 1c-f, reproduced below). In general, the main motivation to use a general decoder instead of the L^{-1} decoder is indeed to show that the multi-timescale advantages can be exploited in a code-agnostic manner, that is, by a decoder that is not informed about the nature of the code, and it simply tries to use the multi-timescale information to minimize a cost function across tasks. In the current version we simplified the decoding model to a non-linear decoder with one hidden layer (Fig. 1d).



We find that this is the simpler decoder that can maintain high accuracy across the tasks, as a decoder without a hidden layer is not able to reliably decode the report in any of the tasks. We now highlight this at the end of the ‘Decoding Tasks’ section in Methods:

“Finally, note that we used a nonlinear decoder for the PG network instead of a linear one. As we showed in the previous section, in principle reward time can be decoded from value estimates with the $L^{(-1)}$ linear decoder. We use a nonlinear decoder for two reasons. First, optimal performance in Tasks 2 and 4 requires nonlinear operations over the learned values. Task 2 requires computing the hyperbolic value and Task 4 requires biasing towards myopic or farsighted estimates based on the estimated reward time. Second, even in tasks where the goal is only to report reward time (e.g. Tasks 1 and 3) the linear decoder $L^{(-1)}$ is only guaranteed to work well in optimal learning conditions (unnoisy value estimates with an infinite number of discounts). In incomplete learning conditions, the $L^{(-1)}$ decoder has been shown to be very sensible to noise (see ref[44]), which leads to poor performance in the tasks studied here. In general, our goal in these simulations was to illustrate the power of the multi-timescale representations over the single-timescale ones, as recoverable by a simple nonlinear decoder”

2. The authors argue that they can decode reward timing by inverting the discount functions matrix. Is this biologically plausible? Do they really think that the brain is performing this operation? If not, why is this being emphasized throughout the text? Please clarify.

The decoding through the inverse Laplace transform only requires a linear decoder which would be fixed if the discount factor of individual neurons is stable, as suggested by the results in Fig. 4. This could potentially be implemented through a learning rule to modulate the contribution of the different dopamine neurons to the learning signal in striatal neurons. Given that the decoding is linear, a simple biologically plausible learning rule like the delta rule would be sufficient.

However, we do not make a strong claim as to whether this decoding scheme is explicitly used in striatum or its downstream circuits to infer reward timing. Measuring the effect of individual dopamine neurons on individual striatal neurons is a technical challenge that would require an entirely different class of experimental methods poorly suited for experiments in-vivo in animals performing behavioral tasks.

3. I did not immediately realize that the major take-home message of Figure 2f,g is that the maximum of the blue and red curves are different. Making this more explicit in the figure would be helpful for the reader.

In the main text, Figure 2f,g is now replaced by Task 4 in Fig. 1, which conveys a clearer take-home message: multi-timescale learning produces more accurate value estimates in noisy and incomplete learning conditions.

Note that we moved the plots originally in Figure 2f,g to Extended Data Figure 3d and I (reproduced below). We have made the markers larger to indicate more clearly the maximum value of the accuracy as a function of the discount factor for the two conditions.

[FIGURE REDACTED]

We also extended our description and analysis of the myopic learning bias (previously Fig2 f,g, now Extended Data Figure 3) to highlight its generality, especially in situations that are more typical in natural environments. In general, as it is now explained in ‘The Myopic Learning Bias’ section in Methods (and also briefly in the main text), in incomplete learning scenarios, the accuracy of myopic vs. far-sighted value estimates depends on the uncertainty structure of the future. When the near future is more certain than the far future, myopic estimates are more accurate; and when the far future is more certain than the near future, far-sighted estimates are more accurate. We illustrate this phenomenon in the new Extended Data Figure 3a-b (reproduced below), where value estimates at the blue and red states are uncertain due to not having explored all possible paths. In both cases, after an infinite exploration of the environment, the upward branch would have a higher value than the downwards branch. After incomplete exploration of the environment (e.g. exploring the upward and downward branches once), we evaluate, for myopic vs. farsighted agents, whether the value of the upward branch is higher than the value of the downwards branch (panel b). In this case, when

the far future has more unexplored paths than the near future (blue), myopic estimates are more accurate (i.e. more representative of the true values after infinite exploration); and when the near future has more unexplored paths than the far future (red), far-sighted estimates are more accurate.

[FIGURE REDACTED]

A similar phenomenon occurs in the grid-world in Extended Data Figure 3f-i. As in the previous example, in the grid-world value estimates are uncertain due to having experienced only a limited number of trajectories. When learning from a limited number of trajectories, if the agent is close to the rewarded zone myopic estimates are more accurate; and if the agent is far from the rewarded zone far-sighted estimates are more accurate. In other scenarios, like in Task 4 in Fig. 1 and in Extended Data Fig. 3c-d, we also observe the myopic learning bias, but for a slightly different reason. In these cases, the uncertainty structure of the future is not determined by the incomplete exploration of the environment, but rather by the stochastic nature of rewards. The Lunar Lander environment, in which we also observe the myopic learning bias (Extended Data Fig. 3j-l) is a more realistic example, where potentially multiple factors combine to create uncertainty about the future (e.g. incomplete exploration and imperfect function approximation in the DQN).

4. Figure 4j - Please clarify in the main text or legend the parametric form that was chosen for fitting the value function. The reader should not have to dig through the methods for this.

The value function inferred (and showed in what is now Figure 3j) is non-parametric. It can take any value at any timestep. We have added in the main text the equation for the continuous form of the TD error to make it more explicit:

“To fit this model to the dopaminergic neurons, we used a bootstrapped constrained optimization procedure on a continuous formulation of the TD error^{71,73} ($\delta_i(t) = b_i + \alpha_i(\gamma_i^{dt} dV(t)/dt - \gamma_i^{dt} \ln(\gamma_i) V(t))$ and see Methods) by fitting a non-parametric common value function and neuron-specific gains, baselines and discount factors. “

Minor comments:

Line 84: typo - should be timestep instead of time steps?

Line 337: exponential should be cubic

Line 380: predictions should be prediction

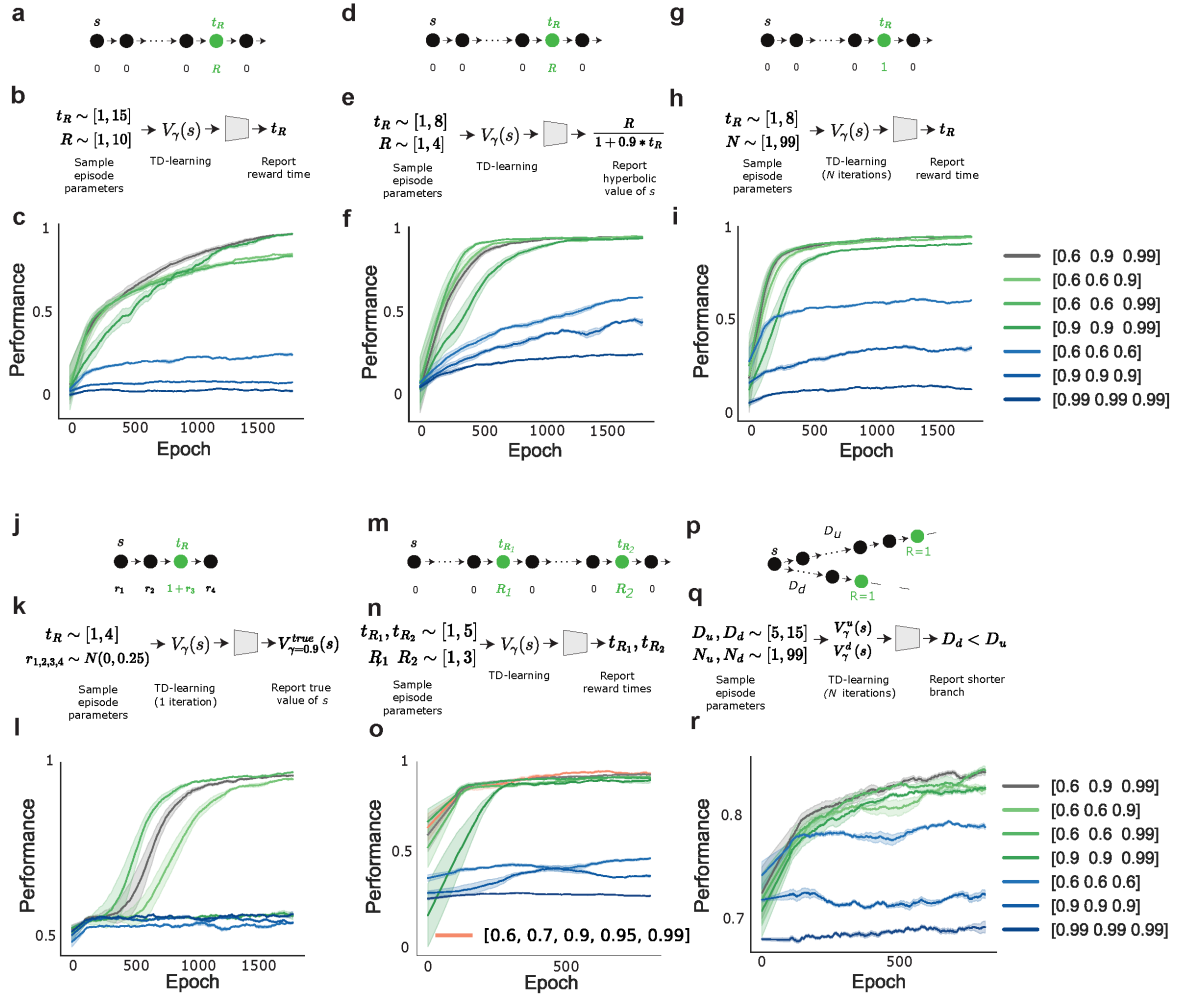
Line 962 - fitted should be fit?

Line 1163 - bootstrap should be bootstraps

We thank the reviewer for pointing out these typos, we have made the necessary corrections.

Extended Data Figure 1 - the axis labels and tick labels are very tiny, the legends are so small that they are not readable.

We have updated the Extended Data Figure 1 as shown below.



Referee #3 (Remarks to the Author):

This is an interesting study whose results, if robust, are important as they suggest a qualitative change in our understanding of the information represented by the dopamine system. The study examines the activity of individual opto-tagged dopamine neurons in mice, during two different tasks; i) odor-reward conditioning using a set of cues which differ with respect to the timing of the predicted reward, and ii) approach to reward along a linear track in virtual reality (VR). The key claims are:

- i) There is heterogeneity across the population of VTA dopamine neurons (DANs) in the discount factor used to calculate the value estimates that drive their reward prediction errors (RPEs). The discount factor determines how much rewards further in the future are discounted relative to those that will occur sooner, so this indicates that the relative weighting of expected rewards near and far in the future is different in the RPEs encoded by different dopamine neurons. This heterogeneity of discount factors removes the ambiguity inherent in standard discounted reward RL about whether a given value indicates a small reward soon or a larger reward later, causing the DAN population activity to carry information not just about how much discounted future reward is expected, but when in time it is expected to occur. This is conceptually related to recent proposals (Dabney et al. Nature 2020) that dopamine neurons encode not just average expected future reward but rather the distribution over expected reward magnitudes. The current work extends this it by proposing that the dopamine neurons carry information about the expected distribution of future rewards over time.
- ii) Such distributional coding of when future reward is expected can improve the performance of reinforcement learning (RL) agents.
- iii) Heterogeneity in the discount factor across dopamine neurons can explain heterogeneity in ramping of their activity during reward approach in VR.
- iv) Discount factors of individual DANs estimated from the two tasks are correlated across the DAN population, indicating the discount factor of individual neurons is a cell specific property (or a property of the specific subnetwork it is embedded in), not a consequence of learning on a specific task.

The authors use a complementary set of analyses to support these claims, including directly assessing heterogeneity in discount rates from the response to cues predicting rewards at different delays, decoding analyses showing they can extract information about when reward is expected from the dopamine activity at the time of the cue, and computational modelling of how heterogeneity in discount rates may affect dopamine ramping, enabling estimation of discount rates in the VR task. This combination of strong theoretical motivation, technically demanding experiments, and sophisticated and creative analyses are real strengths of the study. However, I have some concerns that preclude full confidence in the validity of the claims at this point.

- The authors aggregate neurons across subjects without taking into account possible

inter-subject variability – i.e. they assume fixed rather than random effects across subjects. This is never ideal (though remains common practice in animal neuroscience), as formally the statistical conclusions then only hold for these particular subjects not the population from which they were drawn. However, it is particularly problematic in this study because the key claim rests on there being heterogeneity across the population of neurons in their discount rates. If this heterogeneity is in fact across subjects (or sessions), not across neurons within a given subject at a given time, then the claim does not hold. To address this concern the authors could:

We thank the reviewer for this important point that was also raised by Reviewer 1. As we show below, we have clarified the distribution of the data across animals and have added single subject analyses to show that our results are not due to aggregation across animals or behavior but instead can be observed at the single animal level.

1) Provide versions of the plots showing heterogeneity in discount rates (figure 3E, 4K, Extended data figure 4) with the neurons color coded by subject.

We have updated the plots accordingly. Note that Figures 3 and 4 are now Figures 2 and 3 in the revised manuscript. We have added single subject analysis in Extended Data Fig 4 (panels i-j) as detailed below.

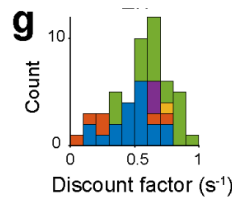


Figure 2g)

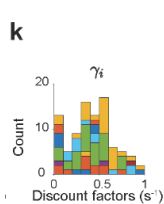
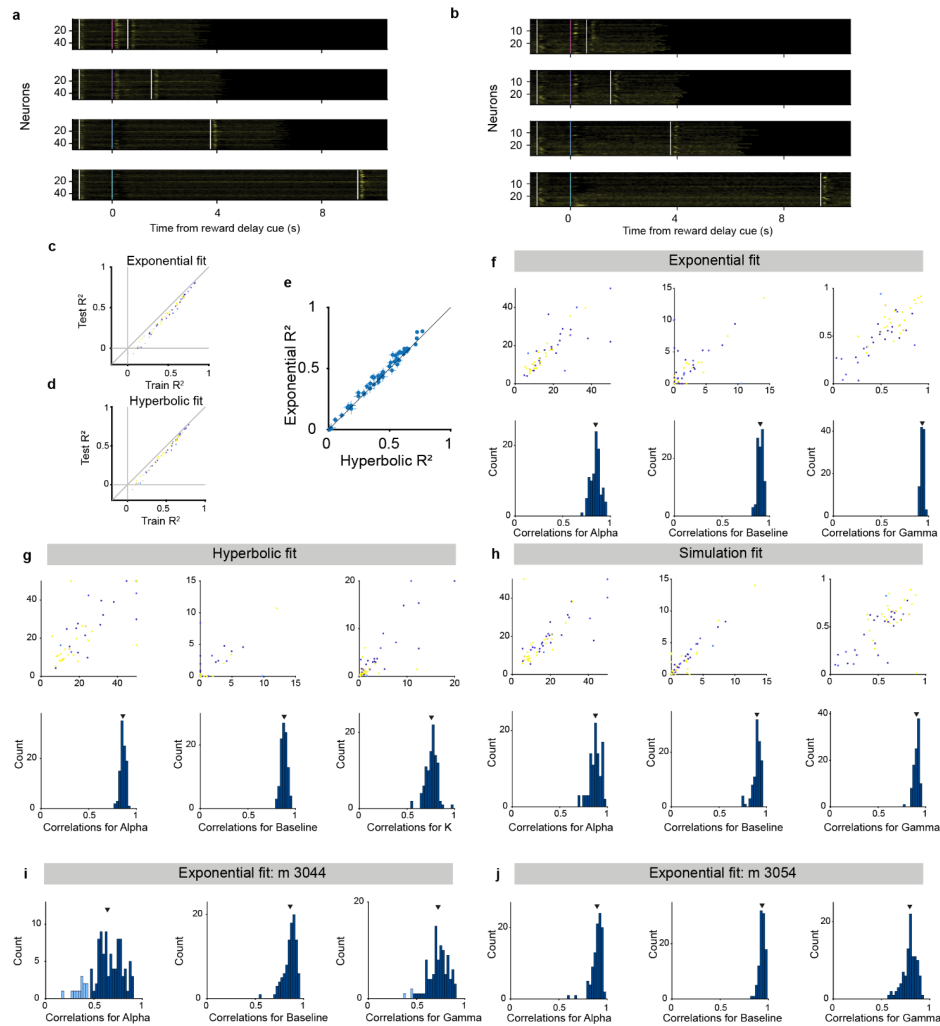


Figure 3k



2) Demonstrate that statistically significant heterogeneity in discount rates is present after controlling for possible systematic variation between subjects (and ideally sessions). Additionally, to help readers understand the strength of the evidence given inevitable cross subject variability, and how likely the results are to hold at the population level:

1). Specify the number of neurons recorded per subject and session.

We added the following table in the extended data (Extended Data Table 1) to more clearly show how the data is distributed across subjects.

“

Animal	Neurons recorded in cued delay task	Neurons recorded per session on cued delay task	Neurons recorded in VR task	Neurons recorded per session in VR task	Neurons recorded in both tasks	Neurons recorded per session in both tasks
m3014	0	0	2 (2)	1(1)	0	0
m3015	0	0	12 (11)	1.7±0.95(1.58±0.79)	0	0
m3017	0	0	1 (1)	1(1)	0	0
m3021	0	0	2 (2)	1(1)	0	0
m3044	36 (19)	2.4±0.83 (1.6±1)	37 (29)	1.9±0.91(1.71±0.77)	32 (15)	2.1±0.83(1.5±0.71)
m3045	7 (6)	1.2±0.41 (1.2±0.45)	13 (11)	1.2±0.4(1.1±0.32)	7 (5)	1.2±0.41(1.2±0.5)
m3047	1 (1)	1 (1)	1 (1)	1(1)	1 (1)	1(1)
m3048	3 (0)	1(0)	6 (6)	1(1)	3 (0)	1(0)
m3049	4 (0)	1.3±0.58 (0)	9 (0)	1.3±0.49(0)	4 (0)	1.3±0.58(0)
m3050	1 (0)	1 (0)	5 (0)	1(0)	1 (0)	1(0)
m3051	0	0	6 (0)	1(0)	0	0
m3053	3 (3)	1(1)	3 (3)	1(1)	2 (2)	1(1)
m3054	23 (21)	1.9±1.2 (1.75±0.97)	24 (24)	1.8±1.1(1.8±1.1)	22 (20)	1.8±1.0(1.7±0.78)

Table 1: Breakdown of the number of recorded neurons per animal and task. The numbers in parenthesis indicate the number of neurons included in the analysis. ± indicates standard deviation across sessions. The maximum number of neurons recorded in a single session was 4 in both the cued delay task and the VR task. “

2) Specify what the experimental unit is (e.g. neuron, trial, session, or subject) whenever reporting stats (including bootstrap tests) or confidence intervals.

We have clarified throughout the text the statistics used and the comparisons made (across neurons, across bootstraps, etc).

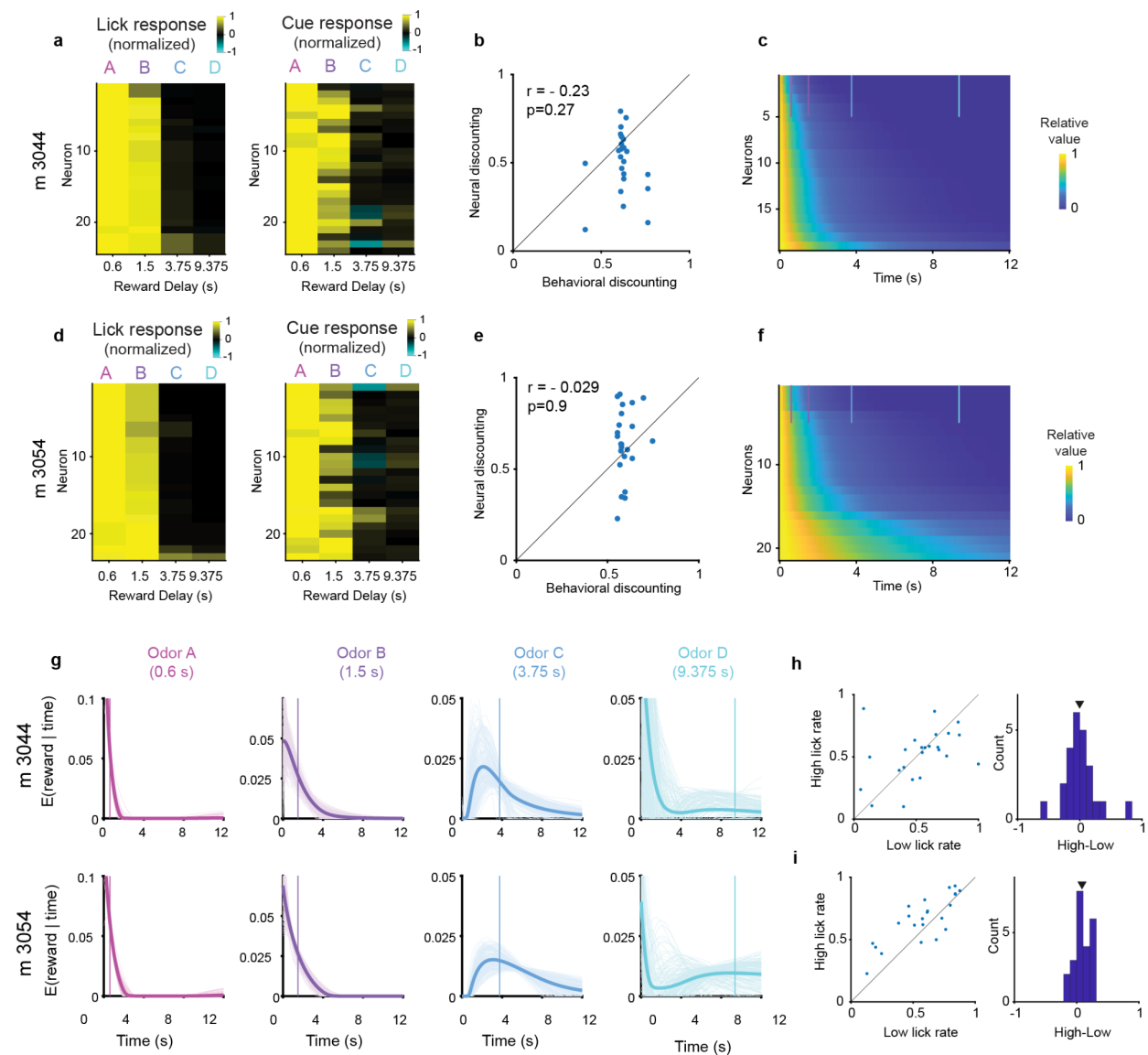
3) Provide sufficient information for readers to assess how consistent the results are across subjects, e.g. by showing key results separately for individual subjects or providing cross-subject uncertainty measures (e.g. by bootstrap resampling subjects).

. Due to the technical challenge of recording optogenetically identified dopaminergic neurons, we often do not have sufficient data in single behavioral sessions as highlighted in the above table. We did record a sufficient number of neurons in 2

animals (m3044 and m3054) to be able to do within-animal analysis for those. We also direct the reviewer to our response to reviewer 1 who raised similar concerns.

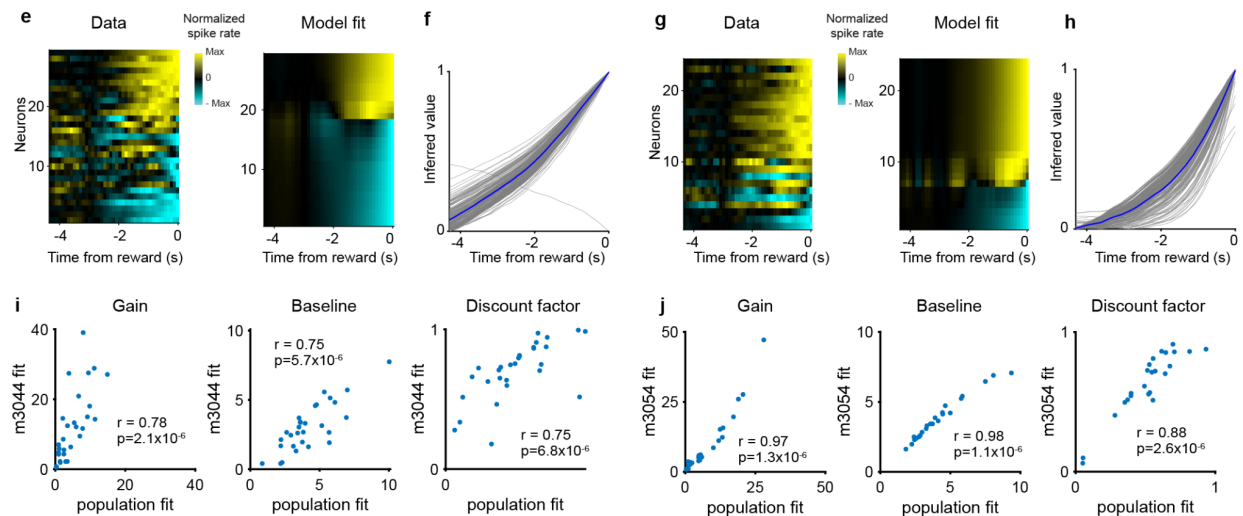
For these two animals, we repeated the main analyses performed throughout the manuscript to show that the results hold at the level of single animals.

In the cued delay task, we show two main results which are presented in Extended Data Figure 6 (shown below). In panels c,f, we show the inferred discount matrix for the two animals. In panels g, we show that we can successfully decode subjective reward timing from neurons recorded in a single animal. The additional panels in the Figure compare neural discounting and behavioral discounting to show that behavior weekly modulates at best the discount factor measured at the single neuron level (see response to reviewer 1).



For these two animals, we also have sufficient data in the VR task at the individual animal level to perform the analysis (m3044: panels e,f,i; m3054: panels g,h,j). We

again obtain very similar results than with the global analysis with a convex value function and a robust prediction of the ramping heterogeneity (panels e-h). Importantly, the inferred parameters at the single neuron level are strongly correlated between the individual animal level analysis and the global analysis (panels i,j).



These plots are part of Extended Figure 9 and we have added a description in the Methods.

Together these additional analyses show that our key argument is robust to single animal analysis and does not arise from pooling across animals and that modulation due to behavior cannot explain the scale of diversity in discount factors across neurons.

- If I have understood the author's arguments correctly there appears to be an incompatibility between the account of heterogeneity in cue evoked dopamine responses in the odor task and that of heterogeneity in ramping in the VR task. In the odor task, my understanding is that heterogeneity in the response to a given cue across DANs is proposed to be due to different DANs computing RPEs using cue value estimates learned using different discount factors – i.e. different DANs use different value estimates to compute RPE. However, the account of heterogeneity in the VR experiment presented in figure 4 assumes that RPEs are computed using the same value estimate for all DANs, but different discount rates across DANs in the RPE calculation. Though the authors do not directly allude to this apparent incompatibility in the text, they do present an alternative account of ramping heterogeneity in the VR experiment in extended data figure 8, in which different DANs use different value estimates to compute RPEs, each computed using different discount rates. However, this account makes the unusual assumption that the value estimates that drive dopamine responses are computed using an estimate of the expected timing of future reward. This is very different from standard RL accounts of basal ganglia in which the value estimates that drive dopamine responses are learned using temporal difference RL, and is also different from the author's previous account of dopamine ramping in Kim et al (Cell 2020) which assume uncertainty about the current state not the timing of

future reward. It is also notable that the distribution of discount rates inferred using this alternative model (extended data figure 8F) is qualitatively different from that in the odor task (figure 3G). To address these concerns the authors could: i) Be explicit both in the main text and through diagrams in the figures about the conceptual models underlying their various accounts of experimental data, making clear where they imagine parallel channels carrying distributional codes for value, RPE etc and where they assume single common signals. ii) Ensure that their accounts of the VR and odor tasks are consistent with each other in their assumptions. iii) Ideally show that heterogeneity in discount rates can account for heterogeneity in ramping without making unusual and hard to justify assumptions about where the value signals that drive the RPEs come from, e.g. show that similar heterogeneity in dopamine ramping emerges when values are learned using TDRL with heterogeneity in discount factors across neurons.

We thank the reviewer for these comments that point to an important aspect of our work. We answer these 3 points below:

i and ii.

We have clarified how the conceptual models used in the manuscript are related. We have added a diagram in Fig. 3c, added Extended Figure 12 (see below) and a section in the Methods ('Mixing in distributed RL models', see below) to show how the different conceptual models used in the manuscript are related. The key insight is to consider that in a distributed model-based RL model, there is likely to be a mismatch in the TD error computation between the model used for current value and the model used for future value. This can be conceptualized by a factor λ that control the amount of sharing in the TD error computation:

$$V_i(s_t) \leftarrow V_i(s_t) + \alpha \left(r_t + \gamma_i \tilde{V}_i(s_{t+1}) - V_i(s_t) \right)$$

$$\tilde{V}_i = \lambda \tilde{V} + (1 - \lambda) V_i$$

The different conceptual models in the manuscript correspond to different values of the mixing factor λ . This mismatch can occur due to imperfect anatomical segregation across the loops or due to additional information (such as sensory feedback) acting as a regularizer. (See Methods section below for details).

Importantly, the computational advantages highlighted in Fig.1 are still present in the partially shared model ('Using in the four tasks of Fig. 1e (see Decoding Tasks in Methods), the accuracy of the report of the agent with three discounts $\{0.6, 0.9, 0.99\}$ is $82 \pm 5\%$ in Task 1, $94 \pm 2\%$ in Task 2, $92 \pm 3\%$ in Task 3 and $93 \pm 3\%$ in Task 4')

"Mixing in distributed RL models"

When explaining how multi-timescale RL can explain the diversity of ramping activity, we proposed two possible interpretations: one with a common value function across all neurons and another with a common reward timing estimation; while in the cued delay task, each dopamine neuron contributes to learning an independent value function. Here we reconcile these

approaches as shown in Extended Data Fig. 12. We can understand these different models as a spectrum in which the value functions used for the computations for each discount factor is more or less shared across them.

In ‘classic’ multi-timescale TD learning⁵⁹, the values (V_i) and RPEs (δ_i) for the different discounts (γ_i) are updated independently, which guarantees its convergence. Now consider a situation in which the value-RPE circuits of the multiple discounts share a common value function, to a degree λ between 0 and 1. In this case, the next value estimate in the timescale-independent TD backup is corrected by:

$$V_i(s_t) \leftarrow V_i(s_t) + \alpha \left(r_t + \gamma_i \tilde{V}_i(s_{t+1}) - V_i(s_t) \right) \quad (32)$$

$$\tilde{V}_i = \lambda \tilde{V} + (1 - \lambda) V_i \quad (33)$$

Where $\tilde{V}$ is the mean value function across all discounts. The main motivation for this modification of the traditional TD backup is neuroanatomical, as it is plausible to consider a degree of common shared activity across nearby value units. Unlike the model with a fully shared common value function (i.e. $\lambda = 1$), multi-timescale learning with small values of the sharing parameter (e.g. $\lambda = 0.1$) preserve, to a large degree, all the computational advantages of multi-timescale learning, while being more biologically realistic than fully separated circuits (i.e. $\lambda = 0$). Using $\lambda = 0.1$ in the four tasks of Fig. 1e (while keeping the same simulation parameters as in Fig. 1, see Decoding Tasks in Methods), the accuracy of the report of the agent with three discounts $\{0.6, 0.9, 0.99\}$ is $82 \pm 5\%$ in Task 1, $94 \pm 2\%$ in Task 2, $92 \pm 3\%$ in Task 3 and $93 \pm 3\%$ in Task 4. Therefore, regularizing the independent value functions with a small degree of shared activity preserves all the multi-timescale advantages highlighted so far.

With this modification to the learning rule, V_i does not converge to a pure exponential form anymore (compare dashed lines with solid lines of Extended Data Fig. 12 a-c Value panels, middle row), even with a small sharing parameter ($\lambda = 0.1$ in Extended Data Fig. 12b). As a result, the RPE does not converge to 0 across the trajectory (Extended Data Fig. 12 b-c, δ , bottom row), so TD learning does not fully converge at the level of individual value estimators. However, we found empirically that learning stabilizes completely after 1000 TD backups (using $\alpha = 0.1$). Crucially, due to the non-exponential form of the learned value function, we observe that $\gamma < \tilde{\gamma}$ ($\gamma > \tilde{\gamma}$) have RPEs that mostly ramp down (up), so our characterization of cell-specific discounts based on ramping patterns is mostly independent of which of the ramping interpretations we adopt. These ramping patterns across timescales are robust when varying the magnitude of the sharing parameter λ ($\lambda \in [0, 1]$), the learning rate α and the number of TD backups (we use a learning rate α of 0.01 and 2000 TD backups for Extended Data Fig. 12). For the simulations in Extended Data Fig. 12, we use the discounts fitted experimentally in Figure 3k (90 units), and plot only three discounts in Extended Data Fig. 12 (renormalized to lie between -1 and 1) corresponding to the 20th, 70th and 90th percentiles, corresponding to discounts 0.25, 0.56 and 0.88.

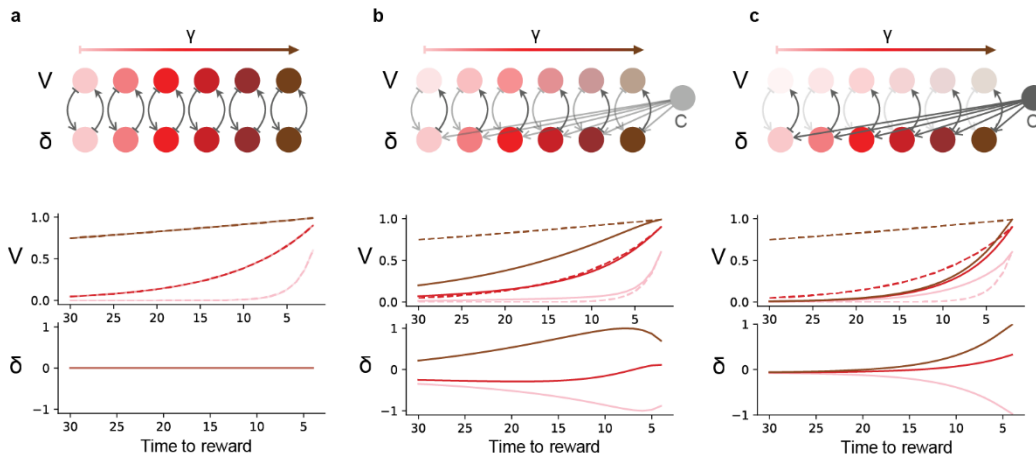
The common value function model we propose corresponds to a value of $\lambda = 1$, while the model in the cued delay task corresponds to $\lambda = 0$. The model in the VR task with a common reward timing estimation falls in between these two extremes. In this model, the TD error for neuron i can be written as follow: $\delta_i = \gamma_i V'_i - V_i$.

$V_i = \sum_{\tau=t}^T \gamma_i^{\tau-t} E(r|\tau, t, T) = \sum_{\tau=t}^T \gamma_i^{\tau-t} E_\tau$ is the value before the sensory feedback and $V'_i = \sum_{\tau=t+1}^T \gamma_i^{\tau-t-1} E'(r|\tau, t+1, T) = \sum_{\tau=t+1}^T \gamma_i^{\tau-t-1} E'_\tau$ the value of the next state including the sensory feedback and the reduction in uncertainty in reward timing. To highlight the contribution of the sensory feedback, we also introduce $V_i^{t+1} = \sum_{\tau=t+1}^T \gamma_i^{\tau-t-1} E'_\tau$, which would be the value at the next step in the absence of sensory feedback (and therefore no reduction in uncertainty about reward timing).

We can rewrite the TD error as follows:

$$\begin{aligned} \delta_i &= \gamma_i V'_i - V_i = \gamma_i V'_i - \gamma_i V_i^{t+1} + \gamma_i V_i^{t+1} - V_i \\ \delta_i &= \gamma_i V_i^{t+1} - V_i + \gamma_i \Delta V_i \end{aligned} \quad (34)$$

Here, the correction due to the sensory feedback appears as ΔV_i , which we can also write as $\sum_{\tau=t+1}^T \gamma_i^{\tau-t-1} (E'_\tau - E_\tau)$. Similarly than in ref [71] the sensory feedback acts as a correction term in the prediction error computation. Here the shared correction term is the reduction in uncertainty, so it take a slightly different from than in the general formulation with shared value $\tilde{V}$ and would correspond to a case where the sharing parameter depends on the discount factor. This source of ‘regularization’ could occur through different pathways. Here, it is the reduction in uncertainty due to the structure of the VR task that leads to a reduction in the uncertainty about reward timing as the mice approach the reward. This contribution of a shared signal or estimate from a parallel value estimation has also been used to explain non-canonical prediction errors in motor tasks⁹⁴. In the cued delay task, there is no feedback about reward timing and therefore we have $\Delta V_i = 0$ in eq. 34, and this corresponds to a situation where the loops are entirely decoupled ($\lambda=0$).



Extended Data Fig. 12 | Decoding reward timing in the cued delayed reward task using parameters inferred in the VR task. Inferred value functions and RPEs for the mixed RL model as a function of the common value function sharing-parameter λ , in a linear MDP of 30 steps (x-axis in the plots) with a deterministic reward equal to 1 in the last step and 0 everywhere else. Plots are shown after learning has empirically stabilized (after 3000 TD-learning iterations with a learning rate of 0.1). The dashed value function is the exponential value function without common value sharing ($\lambda=0$), which would lead to a

*flat RPE equal to 0 at every state. The actual value functions (solid lines) are not purely exponential, and thus lead to ramping RPEs. **a**, Circuit model in which each value estimation and their corresponding prediction error are part completely independent loops ($\lambda = 0$). At convergence, there is no more prediction error in the reward anticipation period (bottom row). **b**, Circuit model in which the prediction error for each dopamine neuron is influenced by both the independent value signal and the shared a common value signal ($C, \lambda = 0.1$). The dashed line indicate the value function corresponding to completely separate loops, and the solid function the actual value function due to the influence of the common value signal. The difference between them leads to ramping in the reward prediction error signals (bottom row). **c**, Circuit model with a strong influence of the common value signal ($\lambda = 1$) which also leads to ramping in the reward prediction error signals. See Methods for details.*

iii.

We have added simulations (Extended Data Fig. 12) to show that the heterogeneity in ramping occurs in an RL algorithm where the value loops are not entirely decoupled.

With regards to the difference between the tasks, the key difference in the reward anticipation period between the two tasks is the presence of sensory feedback affecting the timing estimate that is present in the VR tasks and absent in the cued delay task. In the VR task, the loops are effectively coupled through the sensory feedback that reduces uncertainty about reward timing. This coupling is either total ($\lambda = 1$) in the common value function model or partial in the common reward expectation model (see lines 1338-1361 in the Methods section ‘Mixing in distributed RL models’). Note that any sources of coupling (through the mixing factor λ) will lead to heterogenous ramping. In the cued delay task, the absence of feedback about reward timing implies that there is no correction term ($\Delta V_i = 0$), and it corresponds to a situation where the loops are entirely decoupled ($\lambda = 0$, see text of methods section above)

In the VR task, the reduction in the uncertainty in reward timing due to the visual scene has been extensively studied (see Mikhael, John G., et al. "The role of state uncertainty in the dynamics of dopamine." *Current Biology* 32.5 (2022): 1077-1087. <https://doi.org/10.1016/j.cub.2022.01.025>). In this paper the ramps turn negative if the VR linear track is darkened (indicative of greater uncertainty in the future) as mice approach the reward (Fig 5D-E from the above paper below).

[FIGURE REDACTED]

This leads to an additional term in the TD error computation that corresponds to the reduction in uncertainty due to the visual feedback. We therefore believe that the shared input due to sensory feedback is not just an assumption but a key feature of the VR task. The exact form of this term can take several forms (the two models we propose for

the VR task) but in all cases the diversity of discount factors is key to obtain a heterogeneity of ramping activity in dopamine neurons.

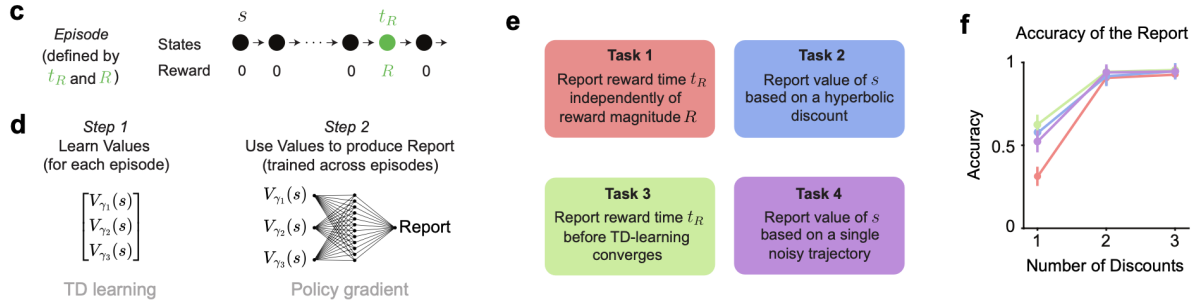
As the reviewer points out, our formulation is in terms of uncertainty of reward timing rather than uncertainty in the state. However, these formulations are essentially identical depending on the point of view we take. In the case of state uncertainty, the current state is unknown, but the reward always occurs in the same state. In the case of timing uncertainty, it is the state in which reward is delivered which is uncertain. The formulation as state uncertainty allowed for the formal derivations in Kim et al, 2020 and Mikhael et al, 2022 but the core effect of the sensory feedback as reducing uncertainty is the same in both interpretations.

- The arguments presented in figure 2F,G do not make a particularly convincing argument that heterogeneity in discount rates leads to better performance in RL agents. Figure 2F is a rather contrived situation in which a noisy reward function in most states combined with carefully placed high reward states causes value estimates learned using either high or low discount rates to work better at specific decision points. It is not obvious how relevant this is to biological learning as the assumption of noisy reward in most states does not seem typical of natural environments. Additionally to take advantage of the additional information learned using multiple discount factors the agent has to employ meta learning across many different task instances, and it is not obvious that this approach would generalise to natural tasks. The deep-Q network learning to play lunar lander (2G) suggests that different discount rates learn better value functions near and far from the goal, but as far as I could see does not demonstrate any actual performance difference between agents with and without a distribution of discount factors. The paper might benefit from focusing more on the core experimental findings and deemphasising these arguments.

We thank the reviewer for highlighting some limitations of our previous theoretical analysis, as it helped us to significantly improve the clarity of the theoretical exposition in the paper. We adapted our descriptions, analyses and simulations in several directions to address these issues.

1. Deemphasize theoretical findings. We simplified the previous Fig. 2 and merged it with Fig.1, such that the 4 multi-timescale advantages (Tasks 1 to 4) use the same example MDP and are given similar importance in the main text (Fig.1c-f, reproduced below). We simplified the explanations for Task 1,2 and 3 in the main text. The myopic learning bias, previously shown in Fig. 2f,g is now replaced by Task 4 in Fig.1d, which shows that multi-timescale learning produces more accurate value estimates when learning from incomplete and noisy information. We believe the paper is easier to read now, as it deemphasizes the theoretical advantages with respect to the core experimental findings, while still displaying the essence for these advantages in the simpler MDP of Fig. 1. We have moved our results in more complex tasks (e.g. Lunar Lander) to the methods (Extended Data Fig. 3). However, we are willing to further

condense the theoretical descriptions in the main text or move part of it to the Methods if the reviewer considers it appropriate.



2. Generality of the representational advantages in naturalistic environments. We extended our description and analysis of the myopic learning bias (previously Fig2 f,g, now Extended Data Figure 3, reproduced below) to highlight its generality, especially in situations that are more typical in natural environments, going beyond the assumption of noisy rewards. As we now explain in ‘The Myopic Learning Bias’ section in Methods (and also briefly in the main text), in incomplete learning scenarios, the accuracy of myopic vs. far-sighted value estimates depends fundamentally on the uncertainty structure of the future. When the near future is more certain than the far future, myopic estimates are more accurate; and when the far future is more certain than the near future, far-sighted estimates are more accurate.

The myopic learning bias holds independently of the source of uncertainty about the future. It holds in cases where uncertainty arises due to noisy rewards (e.g. Task 4 and Extended Data Figure 3c-e), and also when uncertainty arises due to incomplete exploration of the full state space (e.g. Extended Data Figure 3a-b and 3f-i). Consider the scenario shown in Extended Data Figure 3a-b (newly added as part of the revision). After experiencing the trajectories an infinite number of times, in both states the higher-value option is the upwards branch. But consider the case where value estimates at the blue and red states are uncertain due to not having explored all possible paths. (e.g. experiencing the upwards and downwards branch only once). In this case, when the far future has more unexplored paths than the near future (blue state), myopic estimates are more accurate, that is, more representative of the true value after infinite exposure. In contrast, when the near future has more unexplored paths than the far future (red state), far-sighted estimates are more accurate. A similar phenomenon is illustrated more generally in the grid-world in Extended Data Figure 3f-i. In this case, value estimates are uncertain due to having experienced only a limited number of trajectories in the grid-world. When learning from a limited number of trajectories, if the agent is close to the rewarded zone myopic estimates are more accurate; and if the agent is far from the rewarded zone far-sighted estimates are more accurate. The Lunar Lander environment, in which we also observe the myopic learning bias (Extended Data Fig. 3j-l) is a more realistic example, where potentially multiple factors combine to create uncertainty about the future (e.g. incomplete exploration and imperfect function approximation in the DQN).

[FIGURE REDACTED]

3. Necessity of additional learning mechanisms to exploit the multi-timescale representation. The general result is that the accuracy of myopic vs. far-sighted value estimates is determined by the uncertainty structure of the future: when the near (resp., far) future is more certain, myopic (resp., far-sighted) estimates are more accurate. This result holds independently of whether the agent *knows* the uncertainty structure of the future. Naturally, however, actually improving the accuracy of value estimates requires knowing the uncertainty structure, in order to choose between myopic and far-sighted estimates. In other words, exploiting the myopic learning bias requires two components: (1) having available, at every state, myopic and far-sighted value estimates and (2) knowing, at every state, whether the near future is more certain or uncertain than the far future.

In some scenarios, like in Task 4 in Fig. 1 and the MDP in Extended Data Fig. 3c, the second component can be inferred by looking at the multi-timescale values (i.e. by determining t_R in these tasks), which indeed improve the performance of multi-timescale agents compared to single-timescale ones (Fig. 1f and Extended Data Fig. 3e). In other scenarios, the uncertainty structure of the future depends on more complex factors like the distance to the landing zone in Lunar Lander or to the rewarded zone in the grid world. In these cases, the multi-timescale values could potentially still serve as useful proxies to estimate these distances. But in general, the second component might require the agent to rely on separate uncertainty estimates, like counting the fraction of paths left unexplored at different moments along the trajectories, or estimating reward uncertainty at different moments in the future using the full Laplace code presented in our previous work (Tano et al, NeurIPS, 2020).

Therefore, as suggested by the reviewer, taking advantage of the representational advantage provided by the myopic learning bias could indeed require separate task meta-learning in some scenarios (in this case, a meta-learning method to infer the uncertainty structure of the future). The key idea, however, is that only multi-timescale systems can exploit this advantage in the first place, as it requires having myopic and far-sighted value estimates at every state.

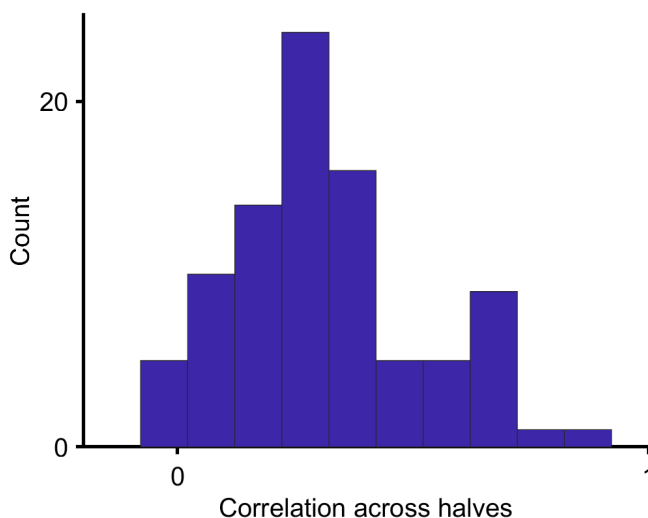
This result aligns with our general theoretical goal in the paper. Previous work ([Fedus et al 2019, Reinke et al 2017, Reinke 2020], now cited in the main text) has already demonstrated increased performance by multi-timescale RL across a wide variety of problems. Our goal in the paper is to explain the fundamental representational advantages that multi-timescale systems have over their single-timescale counterparts, rather than focusing on how these advantages are exploited to produce higher performance by specific architectures.

For example, we did not show performance improvement in the Lunar Lander as classic deep RL algorithms can reach state of the art performance. Tasks on which performance requires computational power outside of the reach of academic research groups like ours and performance would be dependent on a too large range of factors to provide any insight into the representational advantages of multi-timescale RL.

Other points:

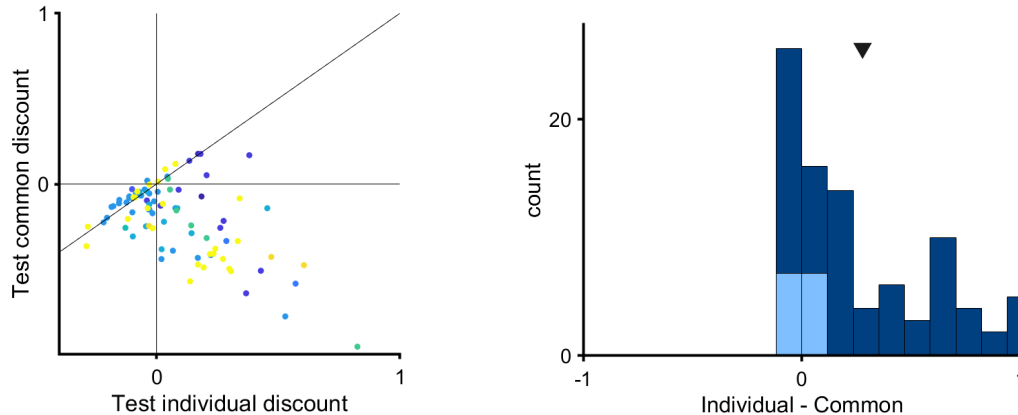
- Please quantify how consistent the heterogeneous ramping profiles in the VR task are across split halves of trials to demonstrate that this is genuine cross neuron heterogeneity rather than noisy estimation.

We computed the correlation between PSTHs across the two halves of the trials. The plot below shows the distribution of average (across bootstraps) correlations across neurons (at the population level across neurons, mean = 0.31, $P = 2.6 \times 10^{-25}$, t -test. At the single neuron level across bootstraps, for 85/90 neurons $r > 0$ and $P < 0.05$, for 3/90 neurons $r < 0$ and $P < 0.05$ and for 2/90 neurons $P > 0.05$). This shows that the observed ramping is reliable across neurons. Due to space constraints, we have not included these results in the revised manuscript but would be happy to do so if the reviewer deems it necessary.



- Please quantify how well the ramping models with heterogeneous discount factors explain neural activity as compared to models with a single discount factor (e.g. using cross validated variance explained on split halves of the trials).

We have added a plot similar to the one in the cued delay task, showing the cross-validated (on test data from the test half of the data) variance explained between recorded PSTH and predicted PSTH for the model with individual discount factors for each neuron and a model with a single discount factor across all neurons (taken to be the average of the inferred discount factors across neurons). The analysis shows that the model with individual discount factors is a much better fit to the data. (At the population level across neurons, mean ΔR^2 : 0.28, $P < 2.3 \times 10^{-10}$, two-tailed t -test; and at the individual neuron level across bootstraps positive difference and $P < 0.05$ for 57 neurons and negative difference and $P < 0.05$ for 19 neurons). Due to space constraints, we have not included these results in the revised manuscript but would be happy to if the reviewer deems it necessary.



- It was not clear to me how the P values for the decoding accuracy of reward times (line 294) were calculated. Was this comparing the decoding accuracy with that obtained when the neuron identities are shuffled?

We thank the reviewer for pointing out the confusing wording. Yes, the p-values on (what was) line 294 correspond to a comparison with the decoding with shuffled neuron identities. We clarified it in the revised text:

“(compared to shuffle control: $P = 1.2 \times 10^{-4}$ for 0.6 s reward delay, $P < 1.0 \times 10^{-20}$ for the other delays, one-tailed Wilcoxon signed rank test; Extended Data Fig. 5e, see Methods). “

- In figure 3K would be useful to show the decoding obtained under the assumption that there is a single discount rate across neurons, as this seems like a better control than that obtained by shuffling the cue responses across neurons given the different baselines and gains of the different neurons. The stats for this comparison are presented in the main text but it would be useful to visualise the decoding accuracy under this important null model.

It is true that the single discount comparison is closer to a comparison with classic RL models. We have changed the controls presented in Fig 3k (Now Figure 2k). We now show the discounting for a common discount factor in Fig 2k. and the random shuffle control in Extended data Fig 5f.

- Line 317: It is not clear what heterogeneity is being referred to in the sentence ‘Here we show that some of this heterogeneity can be understood...’

We clarified the wording of the sentence. We are referring to the heterogeneity/diversity of ramping activity.

“Here we show that some of the diversity in ramping activity across neurons can be understood as evidence for multi-timescale RL across dopamine neurons.”

- Line 947, why was this modified z score used rather than a standard z score?

The ramping starts from a baseline activity earlier in the trial. The goal is to show how the activity deviates from this baseline. If we used standard z-score, by computing the mean activity across all of the trials, for neurons with downward ramp, the start of trial baseline would be above the mean, while for upward ramping neurons, the start of trial baseline would be below the mean. Using the modified z-score, allows us to show more clearly across the neurons the differences in ramping behavior as the mice approach reward delivery.

We thank the reviewers for their support towards publication of this manuscript and the constructive review process. Reviewer comments are in blue and our response in black.

Referees' comments:

Referee #1 (Remarks to the Author):

I appreciate the authors' extensive revisions and explanations in response to my comments, especially the new analysis of licking behaviors. I think the paper can be published in the present form.

We thank the reviewer for the constructive review process and their support to publish our manuscript.

Referee #2 (Remarks to the Author):

The authors addressed my comments.

We thank the reviewer for the constructive review process and their support to publish our manuscript.

Referee #3 (Remarks to the Author):

The authors have addressed most of my concerns and I continue to think that this is an important and interesting work which will be of wide interest to the neuroscience community.

We thank the reviewer for the constructive review process and their support to publish our manuscript. We answer their points below:

The one issue which I think is not well handled in the current manuscript is that of the different and incompatible assumptions in the account of the Pavlovian task presented in figure 2 and that of the VR task in figure 3. Specifically, the account of dopamine responses to the cue in figure 2 requires that separate value signals are learned using different discount factors, while the account of the heterogeneous ramping in figure 3 assumes a common value signal shared by all dopamine neurons. The authors have added a new section to the methods, 'Mixing in distributed RL models', discussing this issue, and I broadly buy their argument that partial mixing of value signals across channels with different discount rates could account for both datasets. However, I think

that there remain some issues with how this is addressed in the manuscript:

- The manuscript is not straightforward in the main text about the different assumptions used in figure 2 and 3. The new methods section is motivated in the text as a way of reconciling two different models of the ramping activity (Figure 3 and ED fig 10), rather than reconciling the different models underlying figure 2 and 3 which is the more fundamental issue. Please make clear in the main results text that figure 2 assumes separate value signals for channels with different discount factors, while figure 3 assumes a single common value signal. Please also add the diagram from figure ED 12a to figure 2 to illustrate the parallel value channels needed for this account to work. The diagram presented in figure 3C also appears not to reflect the true structure of the common value signal model as it shows projections from each value unit to its corresponding prediction error unit, which do not exist in the common value signal model. Please update this diagram to correctly represent the structure of the model. It is OK to use models that make different simplifying assumptions in figure 2 and 3 for didactic reasons, and then reconcile them later through the partial mixing of value signals account, but it is important to be transparent with the reader about what you're doing to avoid confusion.

We thank the reviewer for helping us clarify our models used in the manuscript. We have added the model in Figure 2l and updated the model for Figure 3 (now panel m) to better reflect the underlying assumptions.

- I found the account of mixing of value signals between channels with different discount factors unclear in the current text. At line 1314 the motivation for some mixing of value between channels is motivated by anatomical considerations, and this seems reasonable. However, at line 1359 it is argued that in the absence of feedback about reward timing the channels are entirely decoupled, which does not appear compatible with mixing being a consequence of anatomical crosstalk. Rather than arguing for no value mixing in the Pavlovian task and value mixing in the VR task, a simpler way to reconcile the two datasets would seem to be to simulate a model with partial value mixing between channels on both tasks, and show that it reproduces the key qualitative features of the data. I'm not suggesting changing figures 2 and 3, or attempting to fit this partial mixing model to the data, but rather just to show via simulation that a single model can qualitatively reproduce both datasets.

We have added simulations (Extended Data 7i-k) to show that if there is partial mixing, the information allows the timing information to be decoded in an environment similar to the cued delayed reward task, while observing ramping heterogeneity in an environment similar to the VR task (Extended Data Fig 7j).

- In equation 32 it is not clear why mixing between channels should occur for the value of the new state S_{t+1} but not for the old state S_t . Please explain in the manuscript the rationale for this assumption and whether it is important for the behaviour of the model. Also, my understanding is that the common value signal model (equation 25) does not share this assumption as it uses a single common value signal for both the new state and old state. If this is correct why the different assumptions in these two models?

We have clarified in the methods the distinction between the models. The full common value model of eq. 25 is used to quantify the RPE but it would learn an exponential function that is the average of the discount factors (if we ignore taking into account the added information from the reduction in uncertainty due to the visual feedback). In order to compute the RPE, dopamine units need to know the value of the next state, and in this feedback signal that gammas are mixed. Mixing at s_t would imply a communication among value units that is independent of the RPE computation. We have updated the methods to include the following paragraph (line 1187) :

The model in the cued delay task corresponds to $\lambda = 0$. The common value function model we propose would be similar to a value of $\lambda = 1$ but note that the common value model used for fitting in eq. 25 is slightly different than in eq. 32, as in eq. 32 the shared value is only used for estimating future value. The model from eq. 32 corresponds more closely to the common reward timing estimation model (Extended Data Fig. 9) where the reduction in uncertainty with visual feedback affects the estimate of future value as outlined below. In this model, the TD error for neuron i can be written as follow:

- When describing figure 3 the text uses language suggesting that dopaminergic reward prediction errors contribute to learning a world model: Line 317 'we first assume that heterogeneous dopamine neurons contribute to learning a common model of the environment and therefore share a common value function.' Line 349: 'predictions errors are pooled to create a single value function and world model'. Dopaminergic RPEs are typically though to be used for learning values or policies, rather than learning world models, and the conceptual model underlying figure 3 appears compatible with model-free reinforcement learning of state values. Discussing a world model here therefore seems an unnecessary complication, likely to confuse readers. If the authors feel that model-based RL is conceptually important in explaining these data then the reason for this should be made clear, otherwise I suggest removing these references to world models.

We have rephrased the text on line 317 to say

‘we first assume that heterogeneous dopamine neurons contribute to learning a common model of the value of the states in the environment and therefore share a common value function’

And removed the mention of ‘world model’ in line 349